
Caracterización de las infracciones de tránsito en Bogotá en el año 2017 utilizando un modelo Cox Log – Gaussiano de patrones puntuales

Characterization of transit infractions in Bogota through a Cox Log – Gaussian model of point patterns

Karen Osorio Guerrero.^a
karen.osorio@usantotomas.edu.co

Andrés Ortíz Rico.^b
andresortiz@usantotomas.edu.co

Resumen

El análisis espacial de patrones puntuales se usa para conocer el comportamiento de un conjunto de datos a lo largo de una ventana de observación espacial, para luego así proponer modelos que permitan hacer predicciones acerca del número de ocurrencias de un fenómeno en el lugar de estudio. En este trabajo, se hace una extensión de dichos análisis, donde no solo se estudia el comportamiento espacial, sino la dependencia de esta a través del tiempo. Para esto, se usa una base de datos de infracciones de tránsito ocurridas en la ciudad de Bogotá durante el año 2017, pero se seleccionó la infracción con mayor recurrencia durante ese año, en este caso, la infracción C02, que corresponde a parquear en sitios prohibidos. Para el análisis se consideran varias opciones: primero, observar el comportamiento puramente espacial sin incluir el tiempo como una variable de estudio, trabajar el tiempo como una marca cualitativa, una marca cuantitativa y finalmente, tratarlo como una dimensión más. Se concluyó que esta última es la opción óptima para el análisis, encontrando que tanto a nivel espacial como a nivel espacio temporal, se tiene un patrón agregado. Finalmente, se hayan los riesgos relativos del modelo, haciendo pronósticos en algunos tiempos elegidos al azar.

Palabras clave: espacio temporal, MCMC, Lgcp, proceso puntual inhomogeneo, estadística espacial.

Abstract

The spatial analysis of point patterns is used to know the behavior of a set of data throughout a spatial observation window, and then to propose models that allow making predictions about the number of occurrences of a phenomenon in the place of study. In this work, an extension of these analyzes is made, where not only the spatial behavior is studied, but its dependence through time. For this, a database of traffic violations that occurred in the city of Bogota during the year 2017 is used, but the violation with the greatest occurrence during that year was selected, in this case, the traffic violation C02, which corresponds to parking in prohibited places. For the analysis, several options are considered: first, observe purely spatial behavior without including time as an analysis variable, working time as a qualitative mark, a quantitative mark and

^aEstudiante Maestría en Estadística Aplicada.

^bDecano Facultad de Estadística Universidad Santo Tomás.

finally, as one more dimension. It was concluded that the latter is the most optimal option for the analysis, finding that both at a spatial level and at a space-time level, there is an added pattern. Finally, we have the relative risks of the model, making forecasts at some times chosen at random.

Keywords: Spatio temporal, MCMC, Lgcp, inhomogeneous point process, spatial statistics.

1. Introducción

Según Medicina Legal (2019), a lo largo del año 2019 en Colombia se registraron 6.892 muertes por accidentes de tránsito y 33.539 lesionados por choques, además, en promedio cada día mueren 15 personas en las vías del país. Un informe del Observatorio Nacional de Seguridad Vial manifiesta que 4 de cada 10 accidentes en el país deja muertos o heridos, y son causados en su mayoría por desobedecer las normas de tránsito, en especial, exceder los límites de velocidad. Los motociclistas son los que aportan la mayor cantidad del número de fallecidos, siendo el 51.2% del total.

Lo anterior implica qué, en promedio cada 94 minutos muere un colombiano en las vías, en donde los siniestros viales se han convertido en la segunda causa de muerte en el país; en la capital, en ese mismo año murieron 497 personas, en mayoría entre los 20 y los 34 años, por lo qué, aproximadamente cada 5.6 minutos ocurre un siniestro vial.

Según el SIMIT¹, en Bogotá, las infracciones que más se registraron durante el año 2019 fueron: no realizar la revisión técnico mecánica, conducir sin llevar licencia de conducción, manejar un vehículo a una velocidad no permitida, y conducir con la licencia vencida. Con respecto a las vías donde se presenta mayor accidentalidad se encuentra en primer lugar la avenida Boyacá, la avenida Ciudad de Cali y la avenida Caracas.

En los últimos años se han desarrollado varios trabajos que hacen uso de estadística espacial en problemas de movilidad y accidentes viales: en el año 2014, el SIMIT publicó un documento llamado TRANSITEMOS, en donde hace un análisis espacial de las infracciones de tránsito cometidas en el país desde el año 2004 al 2014, en este análisis se presenta un conteo de las infracciones por regiones, días y meses en el periodo de análisis, pero no hacen un modelamiento con los datos. Steenberghen et al. (2010) proponen una metodología para calcular concentraciones espaciales en eventos basados en puntos sobre una red. En el 2013, Escobar (2013) realizó un análisis espacial aplicando la metodología de patrones puntuales para describir el comportamiento de los accidentes de tránsito ocurridos en la capital en el año 2010. Luego, en el 2014, Jiang et al (2014), estudian los accidentes en vías de la ciudad de Florida (USA), haciendo uso de varios modelos bayesianos: Poisson Gamma, Poisson log-normal, Poisson con efectos aleatorios temporales y Poisson con efectos aleatorios espacio temporales con el fin de identificar cual de los modelos es mejor para identificar lugares de riesgo de accidentalidad en la carretera. Ya en el año 2016, Arias (2016) realizó un análisis espacio temporal de la accidentalidad por tramos en la avenida Boyacá en Bogotá, que se centró en la estimación kernel para identificar puntos de accidentalidad y su cambio durante cinco años. Y, finalmente, para el año 2018, Moncada (2018) estudia los accidentes de tránsito en Bogotá en el año 2015 utilizando la metodología de patrones puntuales sobre la red lineal. La variable tiempo no la consideran dentro del modelo, sino que, realizan la estimación de la intensidad en cada mes.

En este trabajo, se presenta un análisis espacio temporal haciendo uso de las infracciones de tránsito reportadas y/o capturadas durante el año 2017, cuyo objetivo principal es evaluar el comportamiento de la infracción C02² abordando el problema de 4 maneras: haciendo uso de patrones puntuales a nivel espacial, espacial con marcas cuantitativas y cualitativas y finalmente, haciendo uno de un modelo Cox Log – Gaussiano.

¹Sistema Integrado de información sobre multas y sanciones por infracciones de tránsito.

²Parquear en sitios prohibidos

2. Marco teórico

2.1. Análisis espacial de patrones puntuales

Inicialmente, en este trabajo se tendrán en cuenta dos definiciones importantes:

Proceso estacionario: Sea $A \subseteq \mathbb{R}^2$ la región de interés. Un proceso es estacionario si la distribución de probabilidad del proceso en cualquier región $B \subseteq A$ es invariante bajo cualquier traslación arbitraria de B en A .

Proceso isotrópico: Sea $A \subseteq \mathbb{R}^2$ la región de interés. Un proceso es isotrópico si la distribución de probabilidad del proceso en cualquier subregión $B \subseteq A$ es invariante bajo cualquier rotación arbitraria de B en A .

Sea D un conjunto aleatorio, los patrones puntuales son realizaciones de experimentos aleatorios y se distinguen entre patrones completamente aleatorios, agrupados y regulares.



Figura 1: Patrón completamente aleatorio (izquierda), agregado (medio) regular (derecha).

Un patrón puntual es aleatorio (CSR) si se cumplen los siguientes requisitos:

- El promedio de eventos por unidad de área, la intensidad $\lambda(s)$ es homogénea a lo largo de D .
- El número de eventos en dos regiones que no se solapan son independientes.
- El número de eventos en cualquier subregión sigue una distribución Poisson.

Para medir el grado de aleatoriedad de los puntos, existen varias funciones que estiman lo anterior:

- Función G : Distancia al evento más cercano. Esta función mide la distribución de las distancias de un evento arbitrario al evento más cercano. Si estas distancias están definidas como $d_i = \min_j d_{ij}, \forall j, i = 1, \dots, n$, entonces la función G puede ser estimada como:

$$\hat{G}(r) = \frac{\#\{d_i : d_i \leq r, \forall i\}}{n} \quad (1)$$

Donde, j indica el evento más cercano, i , un evento arbitrario, el numerador es el número de elementos en el conjunto de distancias que son menores o iguales a d y n es el número total de puntos. Bajo un proceso completamente aleatorizado la función G es:

$$G(r) = 1 - \exp(-\lambda\pi r^2) \quad (2)$$

Donde λ representa el número promedio de eventos por unidad de área (intensidad), para saber si es compatible con un patrón completamente aleatorio se traza la función empírica $\hat{G}(r)$ versus la función teórica, esta se puede obtener simulando la intensidad en la región de estudio y, una vez se tenga la función empírica debe estar contenida en ella.

El valor esperado de la función G bajo CSR con intensidad λ es $G(r) = 1 - e^{-\lambda\pi r^2}$. Cuando $G(r) > 1 - e^{-\lambda\pi r^2}$ se sugiere un patrón agregado, mientras que si $G(r) < 1 - e^{-\lambda\pi r^2}$ se sugiere un patrón regular.

- Función F : Distancia de un punto al evento más cercano. Esta función mide la distribución de todas las distancias de un punto arbitrario del plano a su evento más cercano. Esta función es usualmente llamada el espacio vacío, porque es una medida del espacio promedio dejado entre dos eventos. Cuando el patrón es completamente aleatorio, la función es:

$$F(r) = 1 - \exp(-\lambda\pi r^2) \quad (3)$$

La función F bajo CSR con intensidad λ es $F(r) = 1 - e^{-\lambda\pi r^2}$, $F(r) < 1 - e^{-\lambda\pi r^2}$ sugiere un patrón agregado y $F(r) > 1 - e^{-\lambda\pi r^2}$ sugiere un patrón regular.

- Función K para marca cualitativas

Para cualquier par de eventos i, j , la función K , $K_{ij}(r)$ es el número esperado de puntos de tipo j dentro de una distancia r de un típico punto de tipo i , estandarizado al dividir por la intensidad de puntos de tipo j :

$$K_{ij}(r) = \frac{1}{\lambda_j} \mathbb{E} \left[t(u, r, \mathbf{X}^{(j)}) \mid u \in \mathbf{X}^{(i)} \right] \quad (4)$$

para $r \geq 0$, donde $X^{(j)}$ es el subproceso de puntos de tipo j , con intensidad λ_j y $t(y, r, x)$ es el conteo de r -vecinos. Dividiendo por λ_j nos permite comprar patrones puntuales con diferentes intensidades. Si dos tipos i y j son iguales, entonces $K_{ij}(r) = K_{ii}(r)$ es la función K usual del proceso X^i de puntos de tipo i . Esta tiene la misma interpretación que la función K usual: por ejemplo, el valor del punto de referencia $K_{ij} = \pi r^2$ (en dos dimensiones) es consistente con un proceso Poisson para los puntos de tipo i .

Sin embargo, si $i \neq j$, la interpretación de $K_{ij}(r)$ es bastante diferente: ahora mide la dependencia o asociación entre los tipos i y j . Suponga que los puntos de tipos i y j son independientes, entonces el número esperado de puntos de tipo j caen dentro de una distancia r de un punto de tipo i es simplemente el número esperado de puntos de tipo j cayendo en cualquier disco de radio r , llamado $\lambda_j \pi r^2$, así que $K_{ij} = \pi r^2$. Esto es, el valor del punto de referencia $K_{ij} = \pi r^2$ es consistente con independencia entre los puntos de tipos i y j . Incluso es consistente con la propiedad más débil de que los puntos de los tipos i y j no están correlacionados.

- Estimadores para análisis espaciales con marcas cuantitativas.

En un patrón puntual marcado de la forma $y = \{(x_1, t_1), \dots, (x_m, t_m)\}$ donde las marcas t_i son números reales. La intensidad con marcas reales, usa el suavizado de Nadaraya-Watson:

$$\tilde{t}(u) = \frac{\sum_i t_i \kappa(u - x_i) b(x_i)}{\sum_i \kappa(u - x_i) b(x_i)} \quad (5)$$

El suavizador de Nadaraya-watson está relacionado con la estimación de la intensidad. Para un proceso puntual marcado en dos dimensiones, con marcas reales, la intensidad $\lambda(u, t)$ si existe, es una función de tres coordenadas (dos espaciales y la marca variable). Integrando sobre la marca variable produce la intensidad del proceso puntual de las ubicaciones,

$$\lambda_{\bullet}(u) = \int_{-\infty}^{\infty} \lambda(u, t) dt$$

En un punto dado u , la razón

$$p(t | u) = \frac{\lambda(u, t)}{\lambda_{\bullet}(u)}$$

es la densidad de probabilidad condicional de la marca t en esta ubicación, y el valor esperado de la marca en esta ubicación es

$$e(u) = \int_{-\infty}^{\infty} t p(t | u) dt = \frac{\int_{-\infty}^{\infty} t \lambda(u, t) dt}{\int_{-\infty}^{\infty} \lambda(u, t) dt}$$

La función marca correlación $k_f(r)$ de un proceso puntual marcado estacionario Y es una medida de dependencia entre las marcas de dos puntos del proceso a una distancia r . Se define la función marca correlación por

$$k_f(r) = \frac{\mathbb{E}[f(t(u), t(v)) | u, v \in \mathbf{X}]}{\mathbb{E}[f(M, M')]}$$

donde M, M' son independientes, idénticamente distribuidas marcas aleatorias la cual tienen la misma distribución como la marca de un punto elegido al azar. $k_f(r)$ es no correlación en el usual sentido estadístico. Como la función par correlación puede tomar cualquier valor real no negativo y el valor 1 sugiere falta de correlación, la interpretación de valores grandes o pequeños que 1 depende la elección de la función f . En este caso, para marcas con valores reales no negativos, $f(m, m') = mm'$.

para cada ubicación espacial u , donde k es el kernel suavizado en un espacio dos dimensional, y $b(x_i)$ es un factor de corrección.

$$K_{ij}(r) = \frac{1}{\lambda_j} \mathbb{E} \left[t(u, r, \mathbf{X}^{(j)}) | u \in \mathbf{X}^{(i)} \right] \quad (6)$$

para $r \geq 0$, donde $X^{(j)}$ es el subproceso de puntos de tipo j , con intensidad λ_j y $t(y, r, x)$ es el conteo de r -vecinos. Dividiendo por λ_j nos permite comprar patrones puntuales con diferentes intensidades. Si dos tipos i y j son iguales, entonces $K_{ij}(r) = K_{ii}(r)$ es la función K usual del proceso X^i de puntos de tipo i . Esta tiene la misma interpretación que la función K usual: por ejemplo, el valor del punto de referencia $K_{ij} = \pi r^2$ (en dos dimensiones) es consistente con un proceso Poisson para los puntos de tipo i .

Sin embargo, si $i \neq j$, la interpretación de $K_{ij}(r)$ es bastante diferente: ahora mide la dependencia o asociación entre los tipos i y j . Suponga que los puntos de tipos i y j son independientes, entonces

el número esperado de puntos de tipo j caen dentro de una distancia r de un punto de tipo i es simplemente el número esperado de puntos de tipo j cayendo en cualquier disco de radio r , llamado $\lambda_j \pi r^2$, así que $K_{ij} = \pi r^2$. Esto es, el valor del punto de referencia $K_{ij} = \pi r^2$ es consistente con independencia entre los puntos de tipos i y j . Incluso es consistente con la propiedad más débil de que los puntos de los tipos i y j no están correlacionados.

2.2. Modelos para procesos puntuales espaciales

Proceso de Poisson homogéneo: Un proceso Poisson homogéneo está caracterizado por representar un proceso en donde todos los eventos son independientes y están uniformemente distribuidos en una región A donde el proceso puntual ocurre. Esto quiere decir que la ubicación de un punto no afecta la probabilidad de que otros puntos aparezcan cerca y que no hay regiones en donde los eventos tienen más probabilidad de suceder. Se define como sigue:

1. El número de eventos en A con área $|A|$ sigue una distribución Poisson con media $\lambda |A|$, donde λ es la constante de intensidad del proceso puntual.
2. Dado n eventos observados en la región A , estos están uniformemente distribuidos en A .

El proceso Poisson homogéneo también es estacionario e isotrópico, es estacionario porque la intensidad es constante y la intensidad de segundo orden dependen solo de la posición relativa de dos puntos; es isotrópico porque la intensidad de segundo orden es invariante a la rotación, por lo tanto, el proceso puntual tiene intensidad constante y su intensidad de segundo orden depende solo de la distancia entre los dos puntos sin importar la posición relativa de estos.

Todo lo anterior indica que la intensidad del proceso puntual es constante, esto, es $\lambda(s) = \lambda > 0, \forall s \in A$, y los eventos son independiente de los otros. El proceso de Poisson homogéneo es la definición formal de un proceso completamente aleatorio.

Proceso de Poisson no homogéneo

En la práctica, el proceso de Poisson homogéneo no es usual que se cumpla, puesto que la distribución espacial de los puntos se puede ver afectada por distintos factores, el proceso de Poisson no homogéneo es una generalización del proceso homogéneo en donde la intensidad no es constante sino variable, el supuesto de independencia entre los eventos se mantiene pero la variación espacial es más diversa, y hay eventos que tienen mayor probabilidad de ocurrencia en unos lugares que en otros, por lo anterior, la intensidad $\lambda(s)$ será una función que varía espacialmente.

- Función de intensidad: La intensidad se define como el número de eventos ocurridos por unidad de área. Si las áreas son muy pequeñas, se tiene,

$$\lambda(s) = \lim_{|s| \rightarrow 0} \frac{E(N(s))}{|s|} \quad (7)$$

Donde, $N(s)$ es el número de eventos en la región s , $|s|$ es el área de la región s

Y si es un proceso de Poisson homogéneo,

$$\lambda(s) = \frac{n}{|s|} \quad (8)$$

La función de intensidad también se puede hallar a través de la estimación kernel, el cual es un método no paramétrico de estimación, en el cual, en la formación de función de densidad no se imponen supuestos tan rígidos a los datos como sucede en el planteamiento paramétrico. La función Kernel $\kappa(u)$ es una función que satisface las siguientes condiciones:

1. $\int k(u) = 1$
2. $\int uk(u) = 0$

Cuando se estima una función de intensidad haciendo uso del kernel en el caso bi - dimensional, la función kernel necesita ser reemplazada por una función que se acomode a las coordenadas, esto puede ser obtenido multiplicando dos funciones kernel univariadas, la elección de dichas funciones puede hacerse a conveniencia, lo que implica que hay ausencia de interacción entre ambas coordenadas. Ya se había visto que en un patrón completamente aleatorio:

$$\lambda(s) = \frac{n}{|A|} \quad (9)$$

Si se tiene x_i y y_i las coordenadas de localización del sitio s_i , entonces el producto kernel lleva a una estimación de la intensidad dada por,

$$\hat{\lambda}(s) = \frac{1}{|A| h_x h_y} \sum_{i=1}^n k\left(\frac{x - x_i}{h_x}\right) k\left(\frac{y - y_i}{h_y}\right) \quad (10)$$

$$\hat{\lambda}(s) = \frac{1}{|A|} \sum_{i=1}^n \frac{1}{h} k\left(\frac{s - s_i}{n}\right) \quad (11)$$

Donde h_x y h_y son los anchos de banda en las direcciones respectivas del sistema de coordenadas. Por otro lado, cuando hay problemas con el efecto del borde $\delta(s)$, hay un estimador que lo considera:

$$\hat{\delta}(s) = \frac{1}{h^2} \sum_{i=1}^n \frac{1}{h} k\left(\frac{s - s_i}{n}\right) \quad (12)$$

Con $\delta(s) = \int_A \frac{1}{h^2} k\left(\frac{s-u}{h}\right) du$, $k(u)$ es una función simétrica y bivariada y $\delta(S)$ es una corrección del borde para compensar las observaciones perdidas que ocurren cuando x está cerca al borde de la región A .

Se debe resaltar que cuando se estima la intensidad con suavizado kernel, el punto importante está en escoger el ancho de banda, pues kernel diferentes producen anchos de banda parecidos, pero el mismo kernel con diferente ancho de banda produce resultados diferentes.

- Verosimilitud: Un método semiparamétrico para estimar la intensidad es por la verosimilitud. La log - verosimilitud de una realización de n eventos independientes de un proceso de Poisson no homogéneo con intensidad $\lambda(s)$ es: (Diggle, 2003, p. 104)

$$L(\lambda) = \sum_{i=1}^n \log \lambda(s_i) - \int_A \lambda(s) ds \quad (13)$$

Donde $\int_A \lambda(s) ds$ es el número esperado de casos de un proceso Poisson no homogéneo con intensidad $\lambda(s)$ en la región A . Cuando la intensidad es estimada paramétricamente la verosimilitud puede ser maximizada obteniendo los estimadores de los parámetros del modelo.

La función de intensidad tiene unas propiedades de segundo orden las cuales miden la fuerza y el tipo de interacción entre eventos de procesos puntuales. La intensidad de segundo orden de dos puntos s_1 y s_2 , refleja la probabilidad de que cualquier par de eventos ocurran en las vecindades de s_1 y s_2 , respectivamente. Una alternativa para medir las propiedades espaciales de segundo orden cuando el proceso es Poisson homogéneo es por medio de la función K. Esta función mide el número de eventos encontrados a una distancia dada de cualquier evento en particular y se define como:

$$K(s) = \lambda^{(-1)} E[N_0(s)] \quad (14)$$

Donde $E[\cdot]$ denota la esperanza y $N(s)$ representa el número de eventos a una distancia d de un evento arbitrario.

Una extensión del proceso de Poisson no homogéneo es capaz de modelar la interacción estocástica entre eventos es el proceso de Cox. Esta extensión está basada en construir una variable aleatoria $\Lambda(x)$ para la función de intensidad. De este modo, diferentes simulaciones del patrón puntual espacial

2.3. Los modelos espacio - temporales de patrones puntuales.

Gabriel, Rowlingson, Diggle (2013) hacen una introducción al paquete stpp en R y se basan en lo siguiente:

Los eventos de un proceso espacio temporal de patrones puntuales forman un conjunto contable de puntos $P = \{(s_i, t_i) : i = 1, 2, \dots\}$, en el cual $s_i \in \mathbb{R}^2$ es la localización y $t_i \in T \subset \mathbb{R}^+$ es el tiempo de ocurrencia del i -ésimo evento. En la práctica, los datos disponibles para el análisis son los puntos $(x_i, t_i) : i = 1, \dots, n$ que forman la realización parcial del proceso restringido a un dominio espacio â temporal finito de observaciones, $S \times T$ donde usualmente S es un polígono y T es un intervalo cerrado.

■ Propiedades de primer y segundo orden.

Las propiedades de primer orden están descritas por la intensidad del proceso:

$$\lambda(s, t) = \lim_{|ds \rightarrow 0| |dt \rightarrow 0|} \frac{E[Y(ds, dt)]}{|ds| |dt|} \quad (15)$$

Donde ds define una pequeña región espacial alrededor de la ubicación s , $|ds|$ es su área, dt es un pequeño intervalo contenido en el tiempo d , $|dt|$ es el ancho del intervalo y $N(ds, dt)$ se refieren a el número de eventos en $ds \times dt$. $\lambda(s, t)$ es el número promedio de eventos por unidad de tiempo en la ubicación (s, t) . Un proceso en donde $\lambda(s, t) = \lambda$ para todo (s, t) es llamado homogéneo.

Las propiedades de segundo orden describen la relación entre el número de eventos en pares de subregiones dentro de $S \times T$. La intensidad de segundo orden está definida como:

$$\lambda_2((s_i, t_i), (s_j, t_j)) = \lim_{|D_i \rightarrow 0| |D_j \rightarrow 0|} \frac{E[Y(D_i, D_j)]}{|D_i| |D_j|} \quad (16)$$

Donde $D_i = ds_i \times dt_i$ y $D_j = ds_j \times dt_j$ son pequeños cilindros que contienen los puntos (s_i, t_i) y (s_j, t_j) respectivamente. La covarianza se define como,

$$\gamma((s_i, t_i), (s_j, t_j)) = \lambda_2((s_i, t_i), (s_j, t_j)) - \lambda(s_i, t_i)\lambda(s_j, t_j) \quad (17)$$

Y la función de correlación es,

$$g((s_i, t_i), (s_j, t_j)) = \frac{\lambda_2((s_i, t_i), (s_j, t_j))}{\lambda(s_i, t_i)\lambda(s_j, t_j)} \quad (18)$$

- **Estacionariedad** Un proceso espacio temporal puntual (s, t) , $s \in S, t \in T$ es estacionario de primer y segundo orden:

1. En espacio, si $\lambda(s, t) = \lambda(t)$ y $\lambda_2((s, t), (s', t)) = \lambda_2(s - s', t)$
2. En tiempo, si $\lambda(s, t) = \lambda(s)$ y $\lambda_2((s, t), (s', t)) = \lambda_2(s, t - t')$
3. En ambos, espacio y tiempo, si $\lambda(s, t) = \lambda$ y $\lambda_2((s, t), (s', t')) = \lambda_2(s - s', t - t')$

Un proceso espacio temporal puntual estacionario también es isotrópico si

$$\lambda_2((s, t), (s', t')) = \lambda_2(u, v)$$

donde (u, v) es vector de diferencia espacio temporal $u = \|s - s'\|$ y $v = \|t - t'\|$.

- **Separabilidad** Un modelo espacio temporal puntual es separable de primer orden si su intensidad $\lambda(s, t)$ puede ser factorizada como:

$$\lambda(s, t) = m(s)\mu(t), \forall (s, t) \in S \times T \quad (19)$$

Donde, $m(s)$ y $\mu(t)$ son las intensidades espacial y temporal, respectivamente.

Un proceso espacio temporal puntual estacionario es separable de segundo orden si la covarianza $\gamma(u, v) = \lambda_2(u, v) - \lambda^2$ se factoriza como,

$$\gamma(u, v) = \gamma_s(u)\gamma_t(v) \quad (20)$$

- **Función K en espacio tiempo no homogéneo** Para la intensidad de segundo orden en un proceso espacio temporal puntual isotrópico la función K se define como:

$$K = 2\pi \int_0^v \int_0^u g(u', v') u' du' dv' \quad (21)$$

Donde $g(u, v) = \frac{\lambda_2(u, v)}{\lambda(s, t)\lambda(s', t')}$ $u = \|s - s'\|$ y $v = \|t - t'\|$

- **Función de correlación en espacio tiempo no homogéneo**

Un estimador de la correlación está definido como:

$$\hat{g}(u, v) = \frac{1}{|S \times T| 4\pi u} \sum_{i=1}^n \sum_{j \neq i} \frac{1}{w_{ij} v_{ij}} \frac{k_s(u - \|s_i - s_j\|) k_t(v - \|t_i - t_j\|)}{\lambda(s_i, t_i) \lambda(s_j, t_j)} \quad (22)$$

Donde w_{ij} y v_{ij} son los factores de corrección espacio temporales y $k_s(\cdot)$, $k_t(\cdot)$ son funciones kernel con anchos de banda h_s y h_t .

- **Proceso de Poisson homogéneo** Este proceso es el mecanismo estocástico más simple para la generación de patrones puntuales espacio temporales. Un proceso de Poisson homogéneo está definido como sigue:

1. Para algún $\lambda > 0$, el número $Y(S \times T)$ de eventos dentro de la región $S \times T$ sigue una distribución Poisson con media $\lambda |S| |T|$, donde $|\cdot|$ denota un área (dos dimensiones) una longitud de acuerdo al contexto.
2. Dado $Y(S \times T) = n$, los n eventos en $S \times T$ forman una muestra aleatoria independiente de una distribución uniforme en $S \times T$. La intensidad de primer y segundo orden de un proceso Poisson homogéneo reduce a las constantes, $\lambda(s, t) = \lambda$ y $\lambda_2((s_i, t_i), (s_j, t_j)) = \lambda_2$.

■ **Proceso de Poisson no homogéneo** Es el más simple proceso puntual no estacionario. Se obtiene reemplazando la intensidad constante λ por una función de intensidad espacial y/o temporal función de intensidad $\lambda(s, t)$. Se define de la siguiente manera:

1. El número $Y(S \times T)$ de eventos dentro de la región $S \times T$ sigue una distribución Poisson con media $\int_S \int_T \lambda(s, t) dt ds$.
2. Dado $Y(S \times T) = n$, los n eventos en $S \times T$ forma una muestra aleatoria independiente de una distribución en $S \times T$ con función de densidad de probabilidad $f(s, t) = \lambda(s, t) \int_S \int_T \lambda(\tilde{s}, \tilde{t}) d\tilde{t} d\tilde{s}$.

■ **Proceso de Cox Log – Gaussiano** Un proceso de Cox es un proceso puntual doblemente estocástico formado por un proceso Poisson no homogéneo con una intensidad estocástica. La definición en espacio tiempo es:

1. $\Lambda(s, t) : s \in S, t \in T$ es un valor no negativo de un proceso estocástico.
2. Condicional en $\Lambda(s, t) : s \in S, t \in T$, los eventos forman un proceso no homogéneo de Poisson con intensidad $\lambda(s, t)$.

Suponga que $Z = Z(s, t); s \in S, t \in T$ es un valor real del proceso gaussiano con media

$$\mu(s, t) = E[Z(s, t)]$$

y función de covarianza

$$c((s_i, t_i), (s_j, t_j)) = cov(Z(s_i, t_i), Z(s_j, t_j))$$

Si la función de intensidad está definida por $\Lambda(s, t) = \exp[Z(s, t)]$, entonces el correspondiente proceso Y es un proceso de Cox Log – gaussiano.

Según (Taylor et al. 2013), una función de intensidad puramente espacial provee el número esperado de puntos por unidad de área a través de una región espacial dada W cuando $x \in W$. Esto se convierte en una intensidad espacio temporal cuando las observaciones son tiempo dependientes, tal que, ahora se busca estimar el número promedio de puntos por unidad de área cuando están situados en una ubicación $x \in W$ y tiempo $t \in T$.

Sea $W \subset \mathbb{R}^2$ una ventana de observación en espacio y $T \subset \mathbb{R}_{\geq 0}$ un intervalo de tiempo de interés. Los casos ocurren en ubicaciones espacio temporales $(x, t) \in W \times T$ de acuerdo a un proceso de Cox espacio temporal inhomogeneo, es decir, un proceso de Poisson con una intensidad estocastica $R(x, t)$.

El número de casos $X_{S, [t_1, t_2]}$, surgen en algún $S \subseteq W$ durante el intervalo $[t_1, t_2] \subseteq T$, se distribuyen Poisson condicionado en R ,

$$X_{S, [t_1, t_2]} \sim \text{Poisson} \left\{ \int_S \int_{t_1}^{t_2} R(s, t) ds dt \right\} \quad (23)$$

De acuerdo con (Diggle et al. 2005), la intensidad se descompone multiplicativamente como

$$R(s, t) = \lambda(s) \mu(t) \exp\{\mathcal{Y}(s, t)\} \quad (24)$$

En la ecuación 24, la componente espacial fija, $\lambda : \mathbb{R}^2 \mapsto \mathbb{R}_{\geq 0}$ es una función proporcionada por el usuario, pro proporcional a la población en riesgo en cada punto en el espacio y escalada de forma que,

$$\int_W \lambda(s) ds = 1 \quad (25)$$

mientras que, la componente temporal fija, $\mu : \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$, es una función proporcionada por el usuario de forma que,

$$\mu(t) = \lim_{|\delta t| \rightarrow 0} \left\{ \frac{\mathbb{E}[X_{W, \delta t}]}{|\delta t|} \right\} \quad (26)$$

La función $\mathcal{Y}$ es un proceso Gaussiano, continuo en espacio y tiempo.

El proceso Gaussiano $\mathcal{Y}$, es de segundo orden estacionario con función de covarianza minimamente parametrizada,

$$\text{cov}[\mathcal{Y}(s_1, t_1), \mathcal{Y}(s_2, t_2)] = \sigma^2 r(\|s_2 - s_1\|; \phi) \exp\{-\theta(t_2 - t_1)\} \quad (27)$$

Donde $\|\cdot\|$ es la norma Euclidiana en $\mathbb{R}^2$, y $\sigma, \phi, \theta > 0$ son parámetros conocidos.

■ Estimación de parámetros por método de mínimo contraste

El método del mínimo contraste implica encontrar los valores del parámetro que minimizan el cuadrado de la discrepancia entre la forma paramétrica asumida de la característica de interés de segundo orden y una correspondiente estimación no paramétrica del mismo. Las funciones paramétricas típicamente usadas para mínimo contraste espacial son la función Par-correlación (PCF) y la función K. Estas son convenientes porque son teóricamente tratables dentro del marco de referencia de un proceso de Cox Log - Gaussiano, y también son accesibles a estimadores no paramétricos. Las formas funcionales de ambas, g y K dependen de la elección de la función de correlación espacial. $J_{\sigma^2, \phi}$ representan las funciones g o K , y $\hat{J}$ representa la estimación no paramétrica en la data observada. La forma general del criterio de mínimo contraste con respecto al rezago espacial u es:

$$\mathcal{M}_J(\sigma^2, \phi) = \int_{u_0}^{u_{\max}} w(u) \left[v\{\hat{J}(u)\} - v\{J_{\sigma^2, \phi}(u)\} \right]^2 du \approx u_{\text{diff}}^{-1} \sum_{u \in I} w(u) \left[v\{\hat{J}(u)\} - v\{J_{\sigma^2, \phi}(u)\} \right]^2 \quad (28)$$

donde u_0 es el rezago espacial más pequeño a ser considerado, $u_{\max}$ es el límite superior en la distancia considerada, $w(u)$ representa un conjunto opcional de peso rezago - dependientes y $v\{\cdot\}$ es una transformación opcional para aplicar a las cantidades de interés.

En la práctica, la integral es aproximada sumando una secuencia de valores uniformemente espaciada $U = u_0, u_1, \dots, u_{\max}$, de tal manera que u_{diff} es la diferencia entre cualquier de dos terminos consecutivos en U .

3. Metodología

Es una investigación en estadística aplicada al contexto colombiano, específicamente a la situación de tránsito y transporte que se vive en la capital del país. Según el índice TomTom³, Bogotá ocupó la segunda posición de los territorios con mayor congestión vehicular de los países estudiados, con un nivel de congestión de 63 %, es decir, que los trayectos se tardan un 63 % más de lo normal.

3.1. Fuentes de información y variables

La fuente de información es una base de datos que contiene las infracciones de tránsito detectadas por lápiz, cámara⁴, dispositivos deap, centros comerciales, centros de gestión y convenios, durante los años 2016 y 2017, pero para este trabajo se decidió trabajar con el año 2017. Las variables que incluye la base de datos se pueden ver en la Tabla 1.

Tabla 1: Descripción de variables

Variable	Descripción
OBJECTID	Id de la infracción dentro de la base de datos.
OBJECTID 1	ID de la infracción dentro de la base de datos del SIMIT.
PLACA	Placa del vehículo en el que se cometió la infracción.
NUM COMPARENDO	Identificación de la infracción dentro del sistema de movilidad.
FECHA HORA	Fecha y hora en la que se detectó la infracción en formato dd-mm-aaaa-hh-mm.
AÑO	Año en el que se detectó la infracción, la base de datos contiene la información de los años 2016 y 2017.
HORA OCURRENCIA	Hora en la que se detectó la infracción.
MES	Mes en el que se detectó la infracción.
MEDIO DETENCIÓN	Medio que usó la autoridad para la captura de la infracción.
CLASE VEHÍCULO:	Tipo de vehículo que cometió la infracción (automóvil, bicicleta, bus, buseta, camión, camioneta, campero, etc.)
INFRACCIÓN	Código de la infracción cometida, dentro de la clasificación del Código Nacional de Tránsito.
DES INFRACCIÓN:	Descripción de la infracción de tránsito.
LOCALIDAD	Localidad en donde se hizo la detención.
DIRECCIÓN	Dirección donde se hizo la detención de la infracción.
LATITUD	Coordenada geográfica correspondiente a la localización de un lugar, en dirección Norte o Sur desde el Ecuador.
LONGITUD	Coordenada geográfica correspondiente a la localización de un lugar, en dirección Este u Oeste desde el meridiano de referencia 0°, o meridiano de Greenwich.

3.2. Análisis exploratorio de datos

En este análisis exploratorio de datos, se identifican datos faltantes en las variables latitud y longitud. La imputación de estos datos se hace a través del proceso de codificación geográfica. La codificación geográfica es el proceso de convertir direcciones en coordenadas geográficas (latitud y longitud). Una de las mejores herramientas para geocodificar es la API de Google maps, por medio de la plataforma de Google Maps

³Este índice toma como referencia 56 países y 403 ciudades midiendo cuanto más se pueden tardar los trayectos en la ciudad por la congestión vehicular.

⁴Centro automático de despacho

haciendo una conexión mediante una API Key y un proceso de facturación, se carga un archivo con extensión .csv y regresa un archivo con la misma extensión y las correspondientes coordenadas.

3.3. Análisis descriptivo

Haciendo uso del software RStudio, se estudia el comportamiento de todas las infracciones durante el año 2017, como todas las variables de la base son categoricas (excepto latitud y longitud), se utilizan gráficos que describan el comportamiento de las mismas. Inicialmente se observa el número de infracciones por mes, por medio de un gráfico de líneas para ver si hay algún tipo de tendencia o, identificar posibles picos de infracciones durante el año. También se utilizan diagramas de barras y columnas para identificar las infracciones que más ocurrieron durante el año de estudio, así como los tipos de vehículos que más cometen infracciones en la ciudad.

Se hace uso de técnicas multivariadas con el objetivo de encontrar posibles relaciones entre las variables. Para este caso, se eligen las variables tipo de infracción, tipo de vehículo, día de la semana y hora del día, esto con el fin de ver si existe alguna relación al cometer la infracción entre la hora del día y el tipo de vehículo, o el día de la semana y la infracción que se comete.

3.4. Análisis espacial y espacio temporal

Para esta parte, se selecciona unicamente la infracción C02, puesto que es la que tiene mayor ocurrencia en el año 2017 y se aborda de 4 distintas maneras:

- Análisis espacial: este análisis se aborda unicamente desde la perspectiva espacial de patrones puntuales, considerando el comportamiento de los puntos sobre el espacio. Inicialmente se hace uso del software Qgis, para identificar el comportamiento de los puntos en cada mes del año, y posibles zonas de calor a lo largo de la ciudad. Adicional, se evalúa identificar el tipo de proceso a través de las funciones K y G.
- Análisis espacial con marca cualitativa o categórica: Se incorpora el tiempo como una variable cualitativa en el análisis, busca evaluar el comportamiento del patrón agregando una marca, en este caso se trabaja el tiempo como el mes en el que se cometió la infracción. Se hace por medio del paquete spatstat de RStudio, donde se tiene un dataframe de dos dimensiones con una marca categórica.
- Análisis espacial con marca cuantitativa: al igual que en el análisis con marca categórica se trabaja con un dataframe con de dos dimensiones pero una marca cuantitativa, en este caso se trabaja el tiempo como el día del año en el que se cometió la infracción.
- Análisis espacio temporal: se considera el tiempo como una tercera dimensión, se utiliza el paquete Stpp de RStudio, este paquete provee las herramientas para hacer estimaciones espacio temporales análogas a las que se hace en un análisis espacial. Para poder calcular la función K inhomogénea y la función par correlación se usan las funciones STIKHAT y PCFHAT. Finalmente, para hacer la estimación de un modelo espacio temporal, se usa el paquete Lgcp en RStudio, este paquete primero hace una estimación de la componente espacial, luego la componente temporal y finalmente la componente espacio temporal. Para la componente temporal, se usa una regresión local Lowess y la estimación de la componente espacio temporal se hace a través de simulaciones de método de Montecarlo basado en cadenas de Markov (MCMC), usando el algoritmo de Metropolis Langevin ajustado (MALA), se eligen los tiempos $T = 50, 80$ y 120 como tiempos al azar para realizar la predicción.

4. Resultados

4.1. Análisis descriptivo

A continuación se presenta el análisis descriptivo de la base de datos de infracciones en el año 2017. En la Fig.2 se observan los meses en donde hay mayor cantidad de infracciones en toda la ciudad y de cualquier tipo, los meses vacacionales como enero, diciembre, junio y julio son los que tienen menor cantidad de infracciones registradas, a diferencia de los meses pico como son febrero y septiembre.

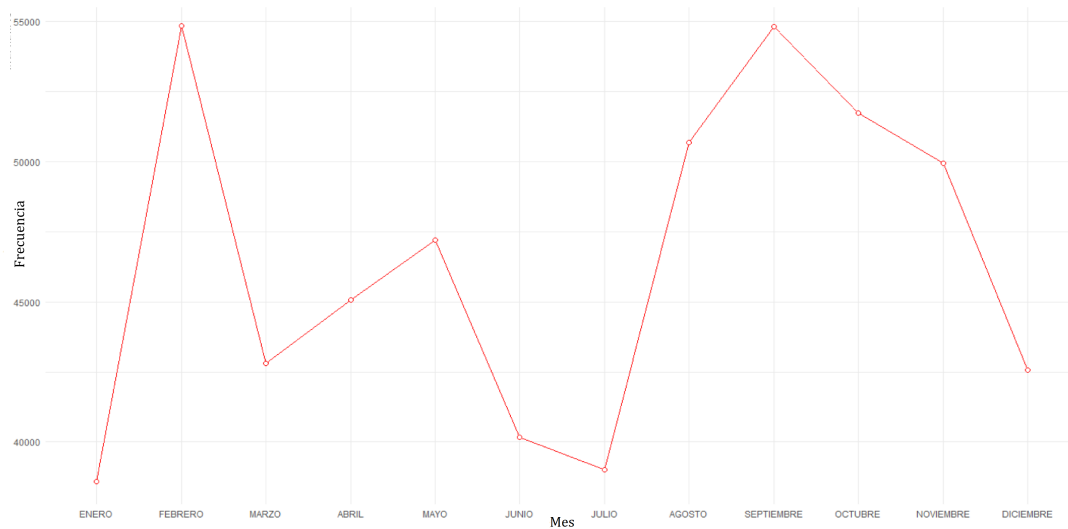
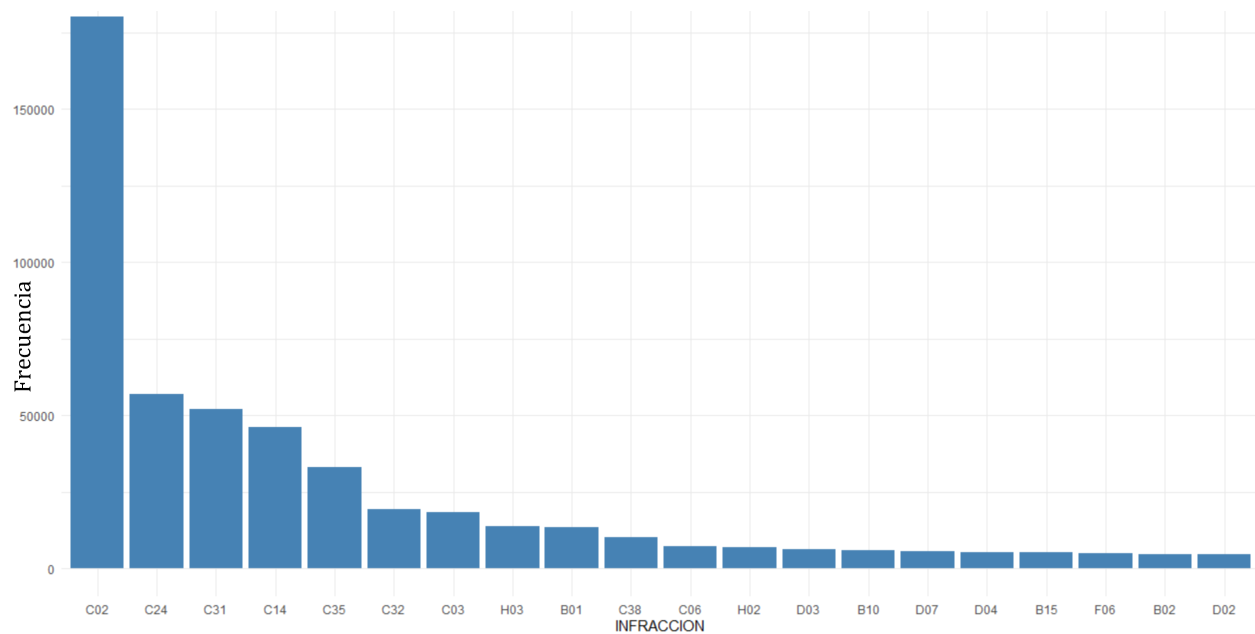


Figura 2: Número de infracciones por mes.

Respecto a los tipos de infracciones, en la Fig. 3 se puede ver cual fue la mayor cantidad de infracciones durante el año de estudio, donde la infracción con mayor incidencia fue la C02, parquear en sitios prohibidos, seguida de las infracciones C24 y C31, que corresponden a conducir motocicleta sin observar las normas establecidas en el código nacional de tránsito terrestre y no acatar las señales de tránsito, respectivamente. La infracción que menos se cometió fue la B02, que prohíbe transitar por zonas prohibidas.

Sobre los tipos de vehículos que más cometen infracciones (Ver Fig. 4), en primer lugar se encuentran los automóviles, seguido de las motocicletas y luego los camiones, esta distribución concuerda con la distribución actual del parque automotor de la ciudad, en donde actualmente, en Bogotá, circulan cerca de 2.400.000 vehículos. De estos, el 50 % son automóviles, el 20 % motocicletas y el 14 %, camionetas. Solo un 5 % corresponde a transporte de servicio público y un 2 %, a taxis amarillos.



C02 - Parquear en sitios prohibidos. C24- Transitar por zonas prohibidas. C31- No acatar las señales o requerimientos impartidos por los agentes de tránsito.
 C14 - Transitar por sitios restringidos o en horas prohibidas por la autoridad competentes. C35 - No realizar la revisión técnica mecánica y de emisiones contaminantes.
 C32 - No respetar el paso de peatones que cruzan en una vía en sentido permitido para ellos. C03 - Bloquear una calzada o intersección con vehículo, salvo cuando el bloqueo obedezca a un accidente.
 H03 - El conductor, pasajero o peatón, que obstaculice, peridique o ponga en riesgo a las demás personas o no cumpla normas o señales de tránsito.
 B01 - Conducir un vehículo sin llevar consigo la licencia de conducción. C38 - Usar sistema móviles de comunicación o teléfonos instalados en los vehículos al momento de conducir.
 C06 - No utilizar el cinturón de seguridad por parte de los ocupantes del vehículo. H02 - El conductor que no porte la licencia de tránsito.
 D03 - Tránsito en sentido contrario al estipulado para la vía, calzada o carril.
 B10 - Conducir un vehículo con vidrios polarizados, entintados y oscurecidos, sin portar el permiso respectivo, de acuerdo a la reglamentación vigente sobre la materia.
 D07 - Conducir realizando maniobras altamente peligrosas, siempre y cuando la maniobra viole las normas de tránsito o ponga en peligro a las personas.
 D04 - No detenerse ante una luz roja o amarilla en el semáforo. B15 - Conducir un vehículo de servicio público que no lleve el aviso de tarifas oficiales.
 B02 - Conducir un vehículo con la licencia de conducción vencida. D02 - Conducir sin portar el seguro obligatorio de accidentes de tránsito ordenado por la ley.

Figura 3: Número de infracciones por tipo.

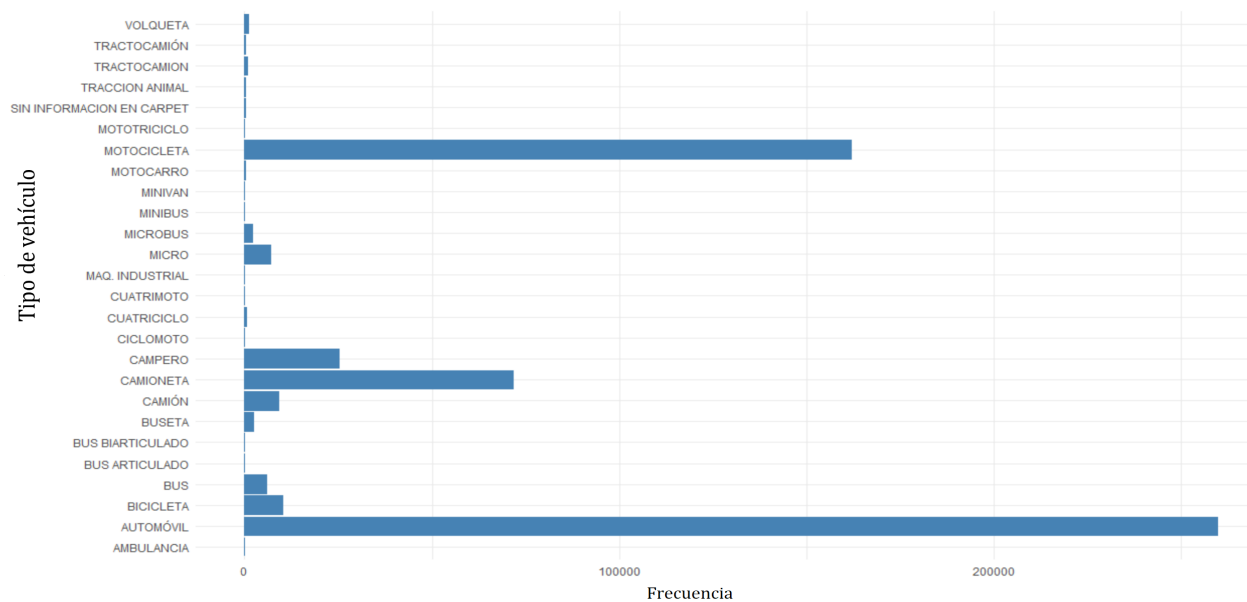


Figura 4: Número de infracciones por tipo de vehículo.

Finalmente, para ver el comportamiento de las infracciones por localidad, en la Fig. 5, es posible ver como Chapinero es la localidad en donde se presenta el mayor número, esto se puede dar por la afluencia de personas en esta localidad, bares y establecimientos de comercio. Una localidad como Fontibón puede presentar una alta cantidad de infracciones debido a la afluencia de carros por ser una salida principal de la ciudad.

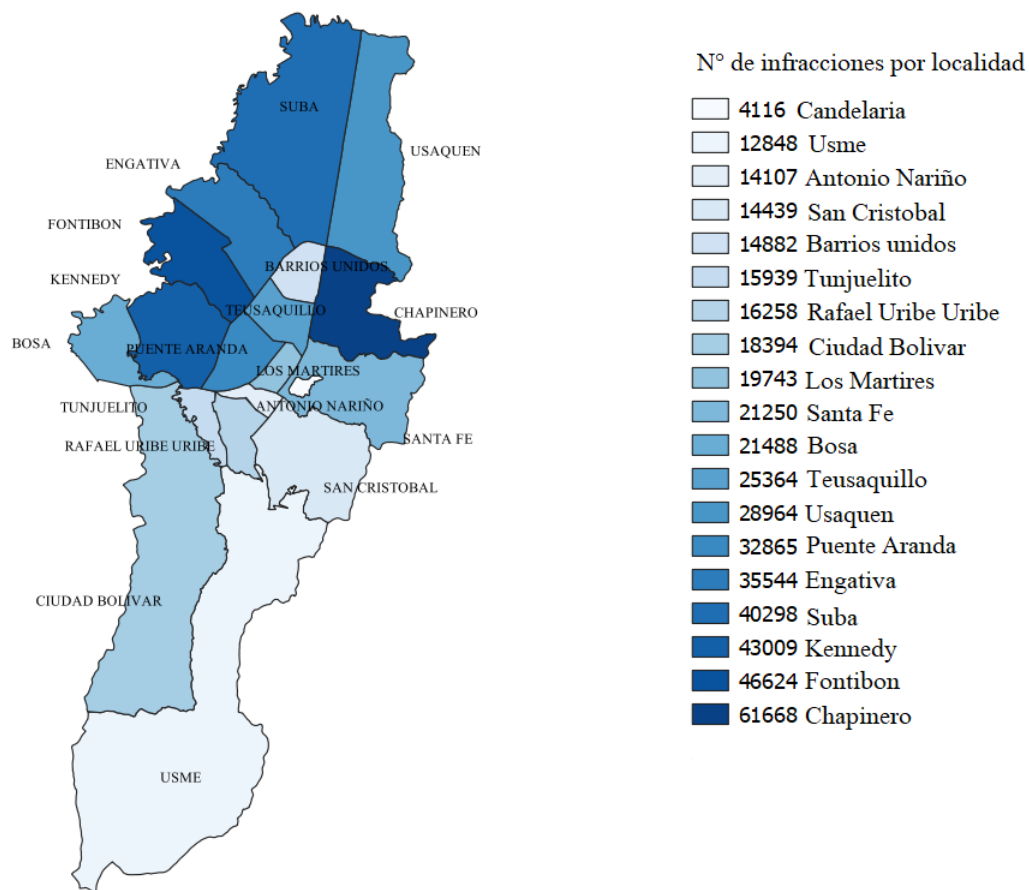


Figura 5: Número de infracciones por localidad

4.2. Análisis de correspondencias múltiples

Para este análisis se utilizaron las variables: TIPO_VEHICULO⁵, HORA.CLA⁶, DIA⁷, INFRACCIÓN⁸ y MES.

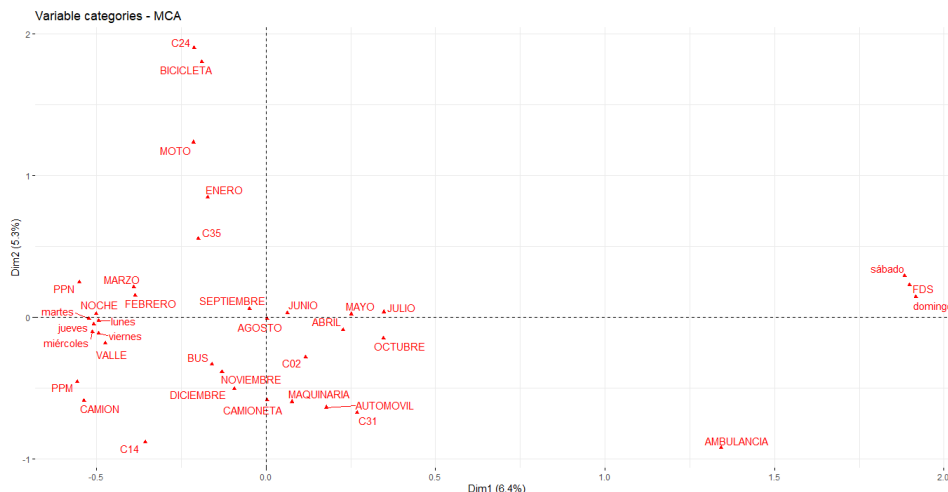


Figura 6: Análisis de correspondencias múltiples.

El análisis de correspondencias múltiples se usa para detectar estructuras y posibles relaciones en el conjunto de datos. En este análisis la inercia indica la cantidad de varianza explicada por las variables. Si la inercia es considerablemente alta, quiere decir que las variables están explicando en buena medida las componentes nuevas.

Las dos primeras dimensiones explican únicamente el 11.6 % de la inercia total, es hasta la dimensión 14 que se recoge más del 50 % de la variabilidad. Esto puede ser una consecuencia de la frecuencia de cada una de las variables, por ejemplo, en el caso de la ambulancia, tienen una frecuencia muy baja y esto puede afectar considerablemente el aporte de algunas variables a la inercia total.

Se observan algunas relaciones en la parte superior derecha del gráfico, la infracción C35 junto a MOTO y el mes de enero, la infracción C35 es impuesta a los vehículos que no tienen vigente o no portan el certificado de revisión técnico mecánica y son las motocicletas quienes presentan mayor afinidad con esta infracción. (Ver Fig. 6)

Con respecto al aporte de cada variable en las tres primeras dimensiones (ver Tabla 2) se tienen los valores de eta 2 que miden el grado de relación de cada variable con cada dimensión. Se puede ver que la dimensión (Dim.1) está muy relacionada con el día y la hora en que se cometió la infracción. La infracción está especialmente relacionada con la infracción y el tipo de vehículo, y la tercera con el mes y la infracción.

⁵Las 29 categorías originales de esta variable fueron agrupadas en 7 categorías generales: ambulancia, automóvil, bicicleta, bus, camión, camioneta, maquinaria, moto de acuerdo a la actividad principal.

⁶Se elaboraron las siguientes categorías dependiendo del comportamiento del tráfico en la ciudad: PPM: horario correspondiente al pico y placa en horas de la mañana (6:00 a.m a 8:30 am), PPN: horario correspondiente al pico y placa en horas de la noche (3:00 pm a 7:30 pm), NOCHE: 7:30 p.m. a 5:59 a.m., VALLE: 8:31 a.m a 2:59 p.m., FDS: sábados y domingos.

⁷Corresponde al día de la semana (lunes, martes miércoles, jueves, viernes, sábado y domingo)

⁸Se seleccionaron las 5 infracciones más frecuentes en el año de análisis: C02: parquear en sitios prohibidos, C14: transitar por sitios prohibidos, C31: no acatar las señales de tránsito, C24: Conducir motocicleta sin observar las normas de tránsito y C35: No realizar la revisión tecnomecánica

Tabla 2: Valores η^2

η^2			
	Dim.1	Dim.2	Dim.3
Mes	0.060	0.090	0.427
Infracción	0.043	0.785	0.622
Tipo vehículo	0.035	0.758	0.100
Día	0.961	0.016	0.007
Hora	0.962	0.050	0.217

4.3. Análisis espacial

Inicialmente, se analiza el comportamiento espacial de la infracción más común durante el año en estudio: C02.

La Fig. 7, muestra el comportamiento de la infracción a lo largo del año, en donde los meses de febrero, abril y marzo tienen la mayor ocurrencia.

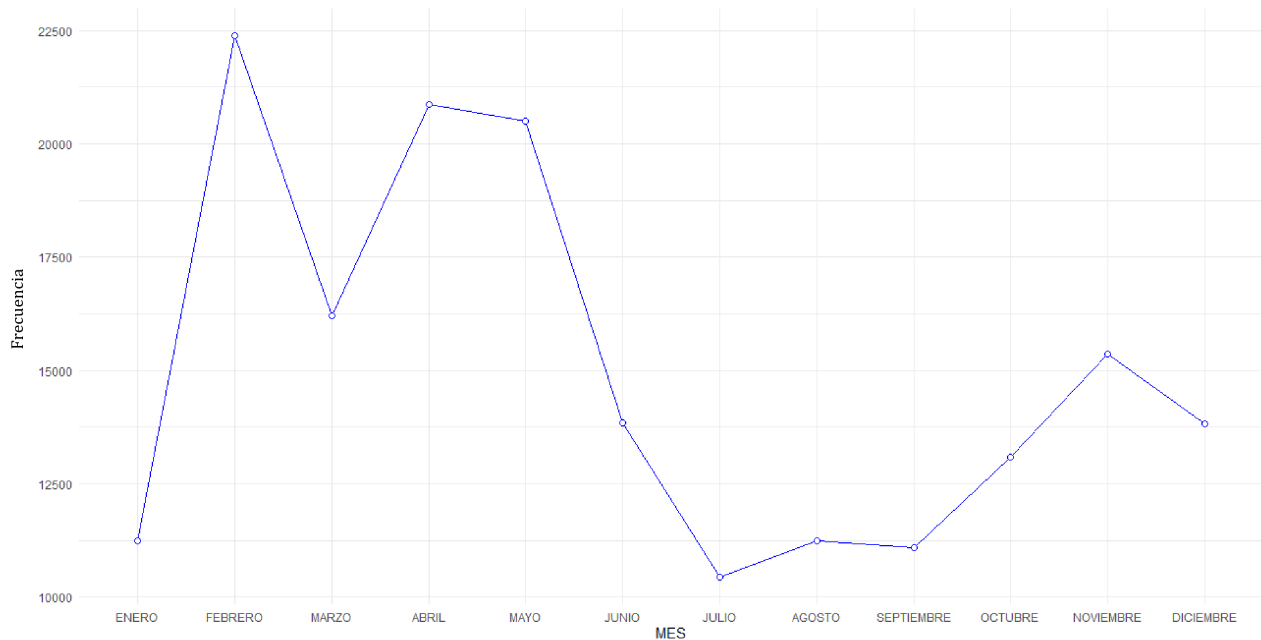


Figura 7: Comportamiento de la infracción C02 en el año 2017

Ahora, por medio del software Qgis, utilizando la opción Mapa de calor se gráficán los puntos de calor mes a mes⁹. En la Fig. 8, es posible ver como hay una variación tanto espacial como temporal. En los meses

⁹Esta interpolación difiere del total de infracciones presentado en la Fig. 7, ya que para esta se tienen en cuenta únicamente los puntos con coordenadas latitud y longitud completas. Luego del proceso de georreferenciación con Google Maps mencionado

de alta ocurrencia de la infracción, como en febrero marzo y mayo, hay puntos de calor a lo largo de casi toda la ciudad, pero se concentra especialmente en las localidades de Chapinero y Usaquén. Ya en el mes de junio, en donde la cantidad de infracciones cometidas disminuye, hay puntos específicos de la ciudad en donde está la concentración espacial, como lo son la zona de outlets cerca al centro comercial Plaza Central, el Aeropuerto Internacional El Dorado, en la localidad de Los Martires cerca a la estación de La Sabana, y en la localidad de Chapinero por las carreras 15 y 7 desde la calle 72 hasta la calle 85. Algunos de estos puntos son constantes en los meses del año, pero en particular empieza a disminuir la concentración a lo largo de toda la ciudad.

El comportamiento de la Fig. 8, muestra como es necesario incorporar el tiempo como tercera variable en el análisis y reafirma la necesidad de un análisis espacial y temporal.

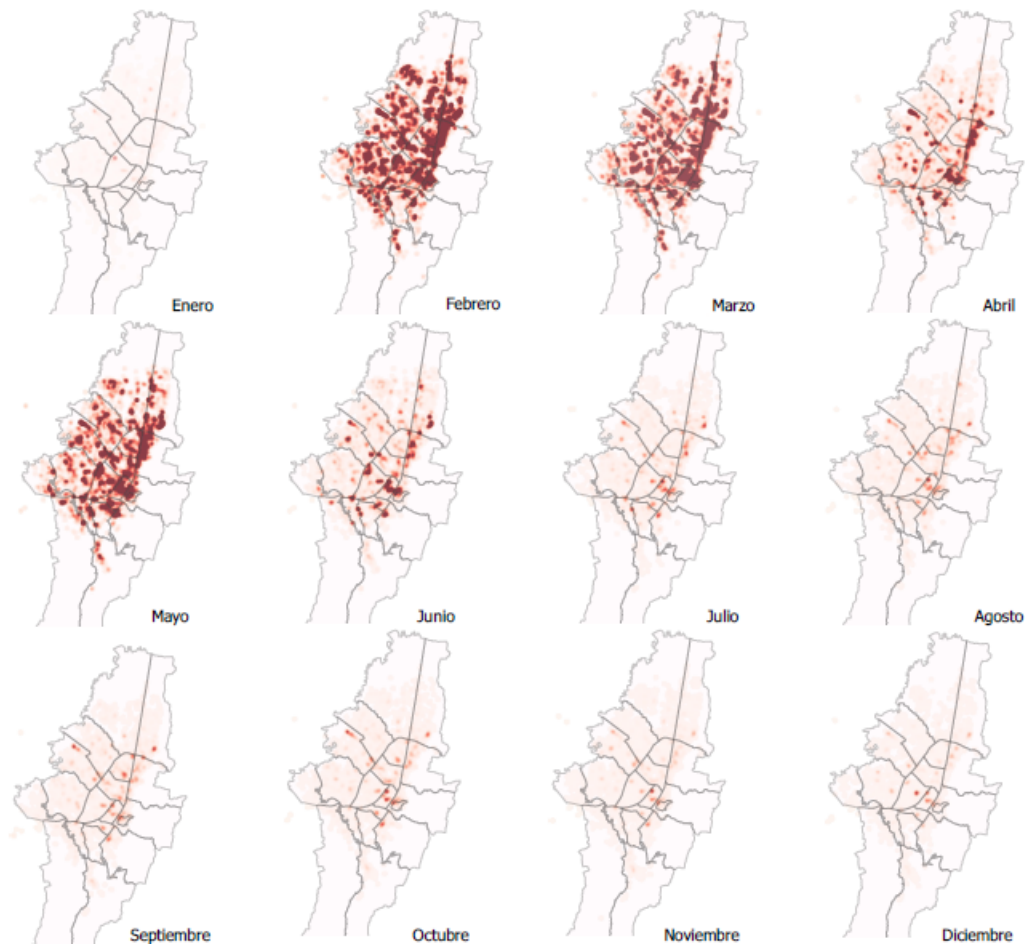


Figura 8: Mapa de calor de infracción C02 en los meses del año.

Hay puntos de la ciudad que siempre presentan esta infracción en todos los meses, como lo son el Aeropuerto Internacional el Dorado (ver Fig. 9).

en la metodología, se eliminan los valores nulos en estas dos variables



Figura 9: Acercamiento zona Aeropuerto Internacional el Dorado en el mes de febrero. Infracción C02.

Para poder hacer un modelo del comportamiento de los datos, es necesario identificar el tipo de patrón que representa la situación. La densidad kernel calcula la densidad de cada punto alrededor de cada celda del mapa, la Fig. 10 muestra la estimación de la densidad a lo largo de los meses y de la ciudad, donde hay puntos en donde la concentración es prácticamente nula y otros donde hay alta concentración de puntos referidos a la infracción. Se puede ver, además, de lugares específicos de la ciudad en donde permanece constante la concentración en todos los meses.

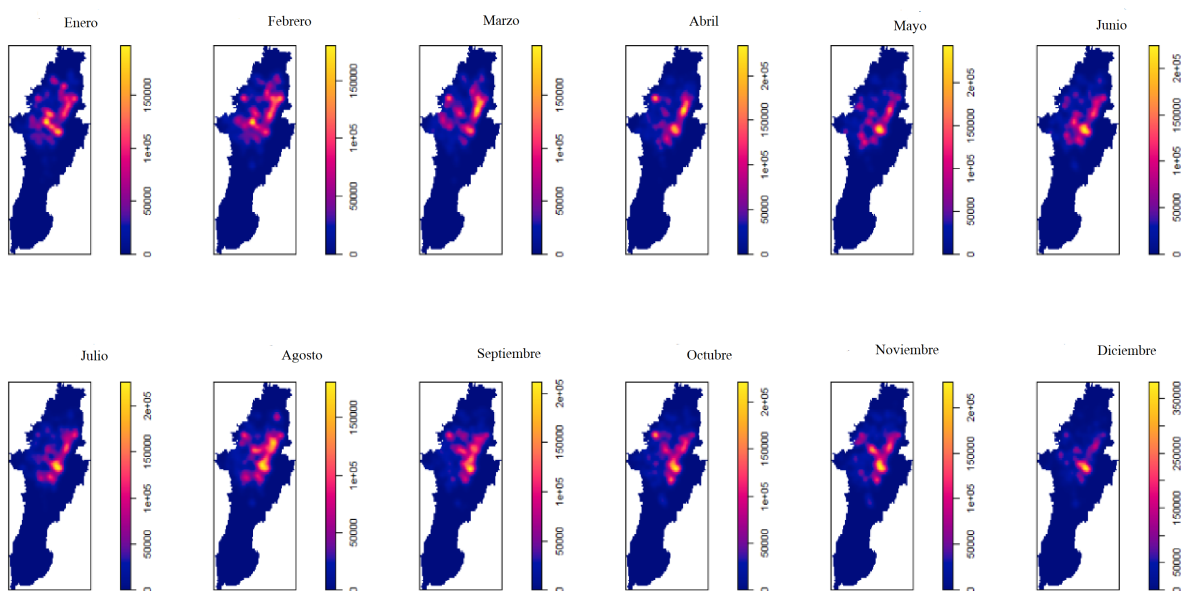


Figura 10: Estimación densidad por mes

Para identificar el tipo de patrón de un proceso, se puede usar la representación gráfica de la función K, la función K de Ripley es una función de λ^2 para procesos estacionarios e isotrópicos. También es conocida

como la medida reducida del segundo momento (Cressie, 1993), y está dada por:

$$k(d) = \frac{E(\text{Número de eventos adicionales a una distancia } d \text{ de un evento elegido aleatoriamente})}{\lambda}$$

λ = Intensidad

$$\hat{K}(d) = \frac{\sum \sum_{i \neq j} I_d(d_{ij})}{\hat{\lambda}_n}$$

con:

- $\hat{\lambda} = n/A$
- Area de la región R
- Distancia entre el i-ésimo y el j-ésimo evento
-

$$I_d(d_{ij}) = \begin{cases} 1 & \text{si } d_{ij} > 0 \\ 0 & \text{en otro caso} \end{cases}$$

Bajo aleatoriedad

$$K(d) = \frac{\lambda \pi d^2}{\lambda} = \pi d^2$$

De donde

- Si $\hat{K}(d) < \pi d^2$ hay regularidad
- Si $\hat{K}(d) > \pi d^2$ hay agregación

Graficamente, la función K, se ve de la siguiente manera de acuerdo al tipo de proceso. (Ver Fig. 11)

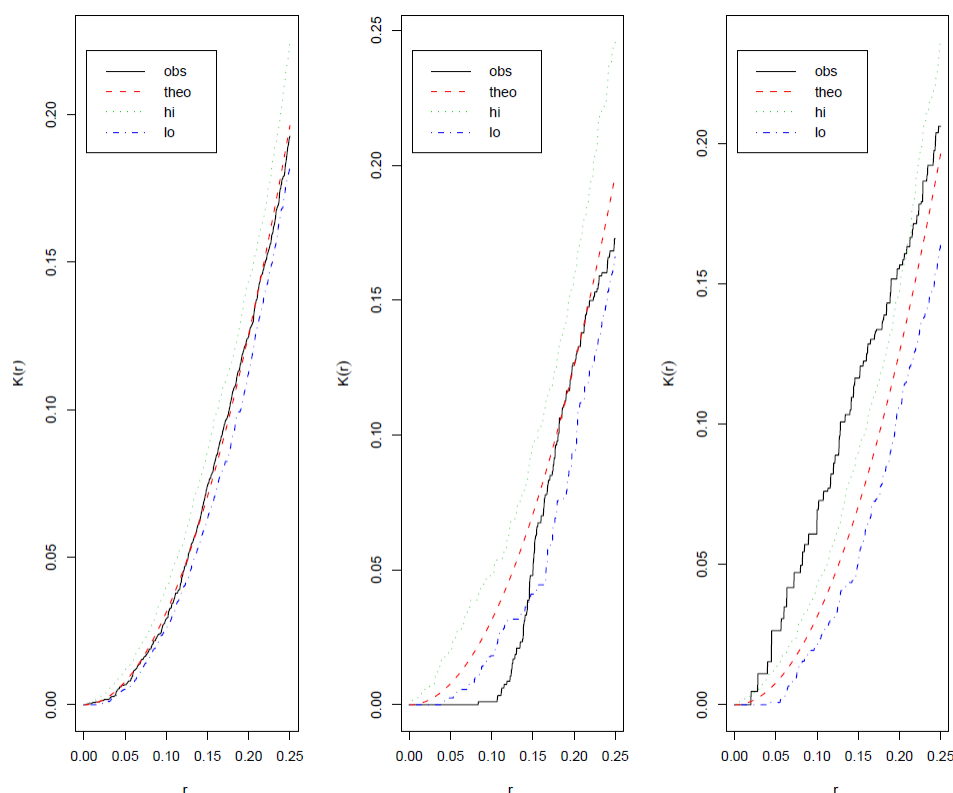


Figura 11: De izquierda a derecha, función K de acuerdo al tipo de patrón: aleatorio, regular y agregado

De acuerdo a lo anterior, la Fig. 12 sugiere un patrón agregado en todos los meses.

4.4. Análisis espacio temporal

Los puntos en un patrón puntual pueden llevar todo tipo de atributos. La información auxiliar adjunta a cada punto en el patrón se llama marca, y en ese caso, el patrón está marcado. Los datos también pueden tener covariables, que son aquellas que se tratan como explicativas¹⁰, en lugar de ser parte de la “respuesta”.

El ejemplo más simple de un patrón de puntos marcado surge cuando la marca adjunta a cada punto es un valor categórico (en este caso el mes en el que se impuso la infracción), una marca numérica es como una coordenada espacial adicional.

Para abordar este análisis, se consideraron varias alternativas: considerar el tiempo como una marca cualitativa, una marca cuantitativa y tratandolo desde su naturaleza como una dimensión más a partir de un análisis espacio temporal. El motivo principal del análisis estructurado de esta manera es indagar, a partir de los datos de las infracciones, acerca de las generalidades, ventajas y desventajas de cada alternativa.

¹⁰Para que una variable sea tratada como explicativa debe estar disponible para cualquier punto de la región, incluidos aquellos en donde no ocurrieron eventos

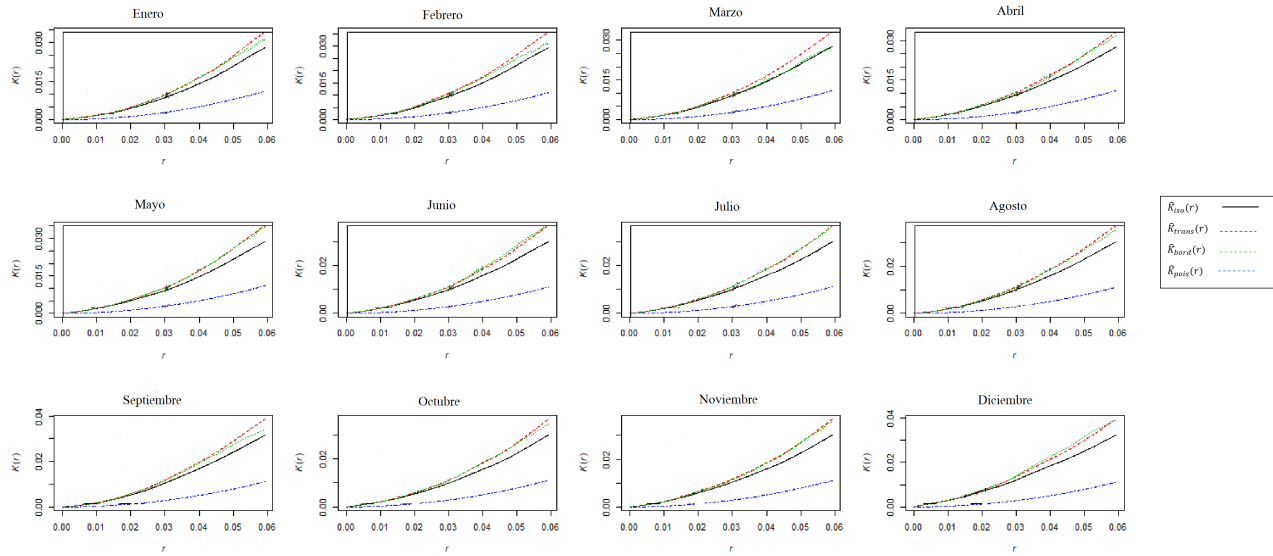


Figura 12: Estimación Función K para cada mes.

4.4.1. Análisis espacial con marcas cualitativas

Cuando se toma el tiempo como una marca cualitativa, se considera el mes en el que se presentó la infracción, este análisis no es el más ideal, pero se abordó como una opción.

La idea central de este análisis es estudiar el comportamiento del patrón puntual en cada uno de los meses, a través de estimaciones de la intensidad en cada mes.

Una gráfica que vale la pena resaltar es la estimación de la función K, con base en cada mes del año. (ver Fig. 13). La diagonal muestra la función K ordinaria para cada uno de los meses. Las discrepancias entre las estimaciones de $\hat{K}_{ij}(r)$ y el punto de referencia πr^2 sugieren fuertemente que el patrón no es Poisson. Además, las gráficas fuera de la diagonal muestran una concordancia entre $\hat{K}_{ij}(r)$ y πr^2 , la cual es consistente con la dependencia entre cada mes.

4.4.2. Análisis espacial con marcas cuantitativas

Según Baddeley (2015), un conjunto de datos con latitud, longitud y los registros del tiempo de ocurrencia de cada evento, puede ser tratado como un proceso espacio temporal o, como un proceso marcado con dos dimensiones en el espacio.

En este caso, se considera el tiempo como una marca cuantitativa, tomando el número del día del año en el que se cometió la infracción, así que, la variable tiempo toma valores de 1 a 365.

Inicialmente, se construye el histograma para observar el comportamiento de la variable (ver Fig. 14)

La Fig. 15, muestra para cada ubicación y, la marca adjunta del vecino más cercano de y en x , es decir, se calcula la intensidad del patrón con base en la marca. La escala de color de cada pixel u representa la media espacialmente suavizada $\hat{t}(u)$. La gráfica sugiere una fuerte concentración de infracciones tipo C02 en los primeros días del año en la parte central derecha del mapa, se puede observar que hay áreas en donde hay

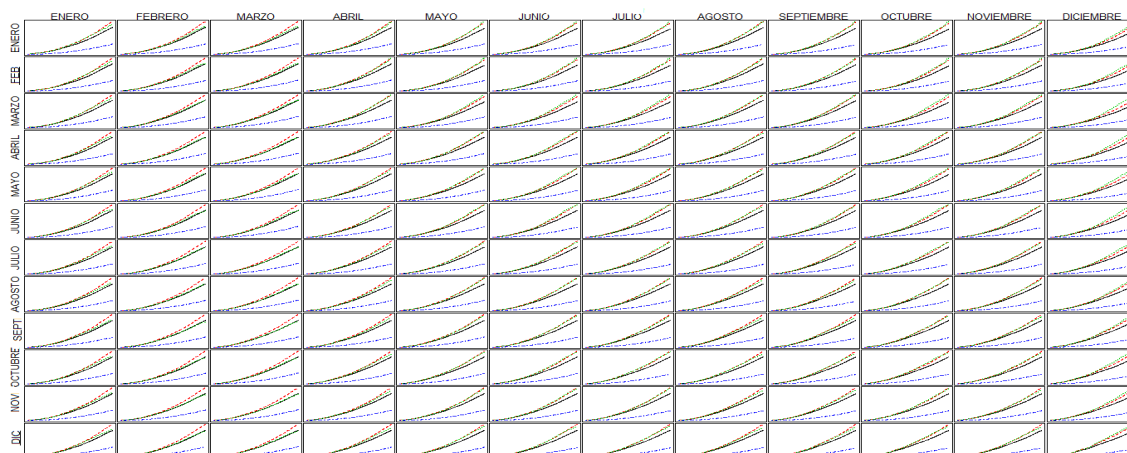


Figura 13: Estimación función K cruzada para marcas cualitativas.

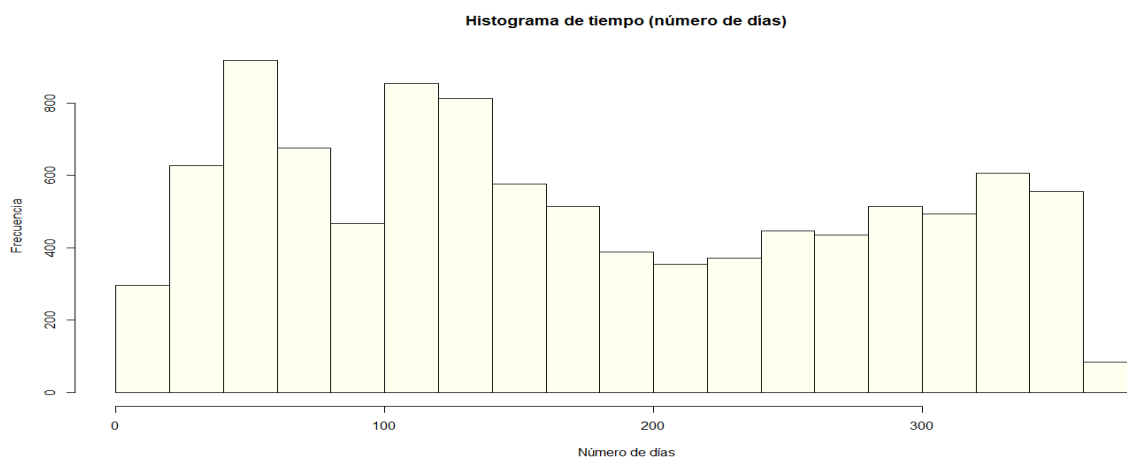


Figura 14: Distribución de la variable tiempo.

una concentración de infracciones en temporadas específicas del año. Por ejemplo, en la localidad de Santa Fe, la mayoría de vehículos estacionados en sitios prohibidos se capturaron en los primeros 50 días del año, al igual que en un sitio específico de la localidad de Ciudad Bolívar. En las cercanías de la localidad de San Cristóbal y al norte de la localidad de Usme se evidencia que las infracciones se capturaron en los últimos días del año.

La Fig. 16, muestra la función marca correlación, este dejar ver como entre infracciones cercanas hay una alta correlación entre los días en que fueron capturadas, es decir, cuando las infracciones se capturan en sitios cercanos, al parecer se capturaron sobre la misma temporada del año, en cambio a distancias largas la correlación es casi nula, si la infracción fue capturada en partes lejanas de la ciudad, al parecer fue en una temporada diferente del año. Vale la pena aclarar que la función no es una correlación en el sentido estadístico usual, esta puede tomar cualquier valor real no negativo, el valor 1 sugiere “correlación débil”.

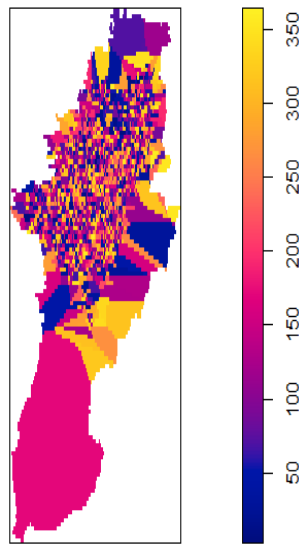


Figura 15: Gráfico de vecinos más cercanos según marca cuantitativa.

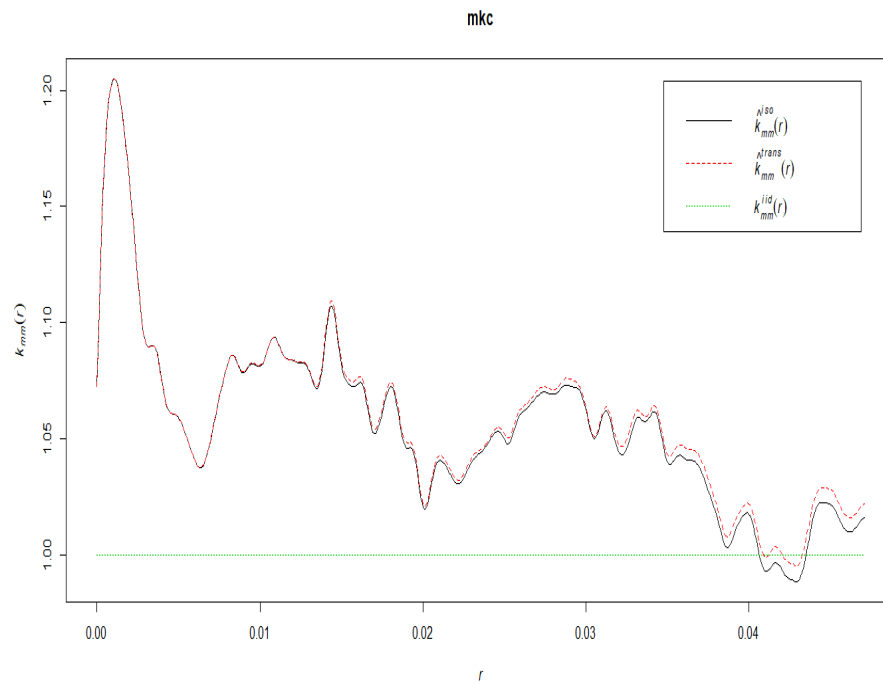


Figura 16: Función marca correlación del proceso usando el tiempo como marca cuantitativa.

Para identificar el tipo de proceso, también se recurre a las estimaciones de las funciones K y L, se propone

la función K ponderada por marca

$$K_f(r) = \frac{1}{\lambda \mathbb{E}[f(M, M')]} \mathbb{E} \left[\sum_{x_j \in \mathbf{X}} f(t(u), t(x_j)) \mathbf{1}\{0 < \|u - x_j\| \leq r\} \quad u \in \mathbf{X} \right] \quad (29)$$

donde $t(u)$ y $t(x_j)$ denotan los valores de las marcas en los puntos u y x_j , y nuevamente M y M' son marcas aleatorias independientes que tienen la misma distribución que las marcas en el proceso puntual. La normalización está diseñada para que bajo aleatoriedad $K_f(r) = K(r)$, la función K usual. La lectura de estas funciones es la misma que las funciones K y L usuales.

Ahora bien, las estimaciones de las funciones K y L (Fig. 17) sugieren que el proceso es agregado.

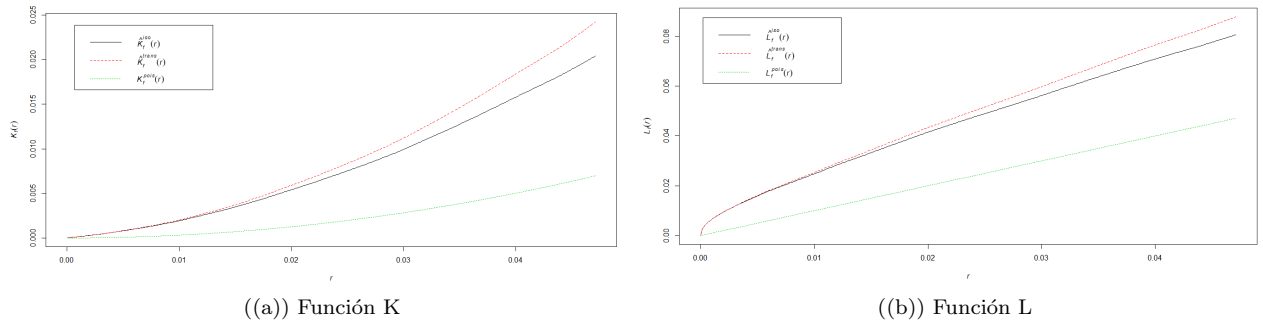


Figura 17: Funciones K y L del proceso con marca cuantitativa.

4.4.3. Análisis espacio temporal

En este análisis se consideran los datos de forma espacio temporal, es decir, los datos disponibles son puntos $(x_i, t_i) : i = 1, \dots, n$, con $x \in \mathbb{R}^2$ y $t \in \mathbb{R}^+$

Las funciones STIKHAT y PCFHAT proporcionan estimaciones espacio temporales de la función k inhomogénea y la función par correlación.

La función STIKHAT (K_{ST}) caracteriza las propiedades de segundo orden y la intensidad de segundo orden de un proceso puntual espacio temporal estacionario ponderado y puede ser como una medida espacio temporal de agregación o regularidad. Para cualquier proceso puntual espacial temporal Poisson inhomogéneo con intensidad acotada lejos de cero, con $u = \|x - xt\|$ y $v = \|t - t'\|$, $K_{ST}(u, v) = \pi u^2 v$. Valores de K_{ST} mayores que $\pi u^2 v$ indican agregación y acumulación espacial y separaciones temporales menos que u y v , mientras que $k(u, v) < \pi u^2 v$ indica regularidad. La función STIKHAT puede ser usada para probar agrupación e interacción espacio temporal.

Con base en lo anterior, la Fig. 18, indica la presencia de un patrón agregado.

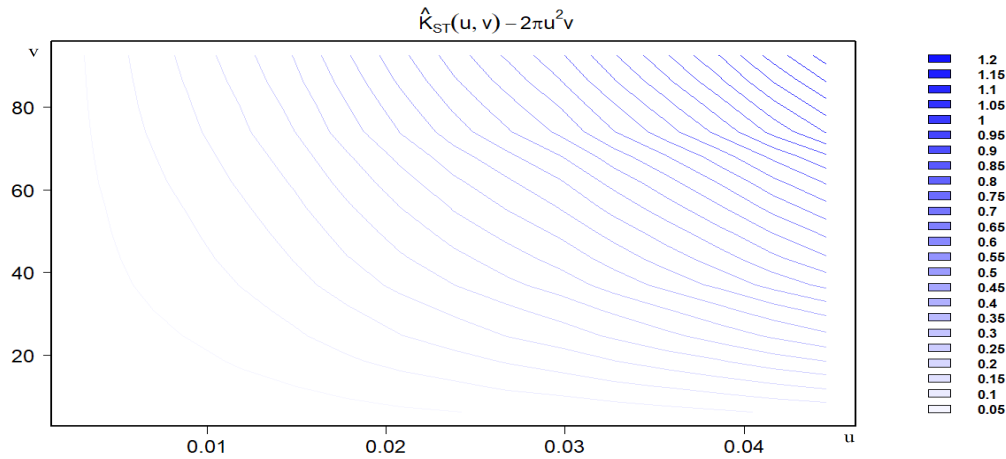


Figura 18: Función STIKHAT.

4.4.4. Modelo y predicción espacio temporal

Según (Taylor et al. 2013), en el marco del modelado de un proceso de Cox con intensidad estocástica se tiene $R(s, t) = \lambda(s)\mu(t) \exp\{\mathcal{Y}(s, t)\}$, donde, $\mathcal{Y}(s, t)$ es un proceso Gaussiano y λ y μ son las estimaciones de la intensidad espacial y temporal, respectivamente. Haciendo uso del paquete Lgcp, la estimación de λ y μ son no paramétricas, a través de una estimación de densidad kernel bivariada simple y un suavizado, respectivamente.

En este punto, las observaciones, X , son tratadas como contenido de celdas en la grilla la versión discreta de Y será denotada como Y , con Y una colección de variables aleatorias, las propiedades de Y implican que Y tiene una densidad Gaussiana multivariada con matriz de covarianza apropiada σ .

Inicialmente, la ventana poligonal en la que se realizará el modelo es:

$[-74.22358, -73.98653] \times [4.269664, 4.836779]$

- Estimación de las componentes espacial y temporal.

La componente espacial fija $\lambda(s)$ representa la intensidad espacial de los eventos, promediados a lo largo del tiempo y escalados para integrarse a 1 sobre la ventana de observación W . La componente temporal fija $\mu(t)$ es el número promedio de eventos en W por unidad de tiempo.

Para la componente temporal, se hace una estimación no paramétrica a través de una regresión local Lowess, la cual es un tipo de ajuste mediante suavizados y una técnica de regresión local ponderada.

- Estimación de parámetros

Después de estimar $\lambda(s)$ y $\mu(t)$, el siguiente paso es estimar los parámetros de la covarianza del proceso $\mathcal{Y}$. El paquete Lgcp hace uso del método de mínimo contraste para estimar σ , ϕ y θ presentes en la ecuación 27. Los parámetros del modelo son $\sigma = 3.500414$, $\phi = 0.0043585$ y $\theta = 0.1514745$.

- Predicción

Para hacer la predicción se hace uso de la función lgcpPredict, la cual hace uso de cadenas de Monte Carlo: MCMC (MALA¹¹) para producir muestras y resumir estadísticas de la distribución predictiva

¹¹Metropolis-adjusted Langevin algorithm.

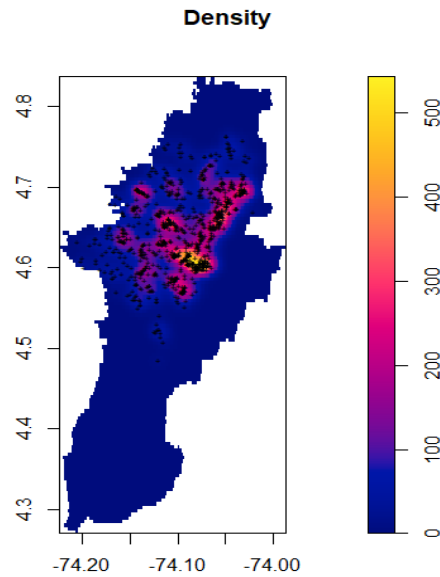


Figura 19: Estimación kernel de $\lambda(s)$.

del proceso discretizado Y . Los parámetros para correr el algoritmo son:

- $T = 50$, tiempo de interés en la predicción.
- $\text{laglength} = 5$, longitud del rezago
- $\text{cellwidth} = 0.01$, ancho de celda
- $\text{mala.length} = 80.000$, número de iteraciones
- $\text{burnin} = 20.000$, número de iteraciones a desechar antes de empezar
- $\text{retain} = 100$, presenta una acción cada 100 iteraciones
- adaptivescheme = parámetro de ajuste h del algoritmo MALA, en este caso se usará la función `andrieuthomsh`¹²,
- $\text{gridmeans} = \text{MonteCarloAverage}(\text{"exceed"})$ ¹³

¹²Ver (Andrieu & Thoms 2008).

¹³Calcula los valores esperados del MCMC en la función `lgcpPredict`.

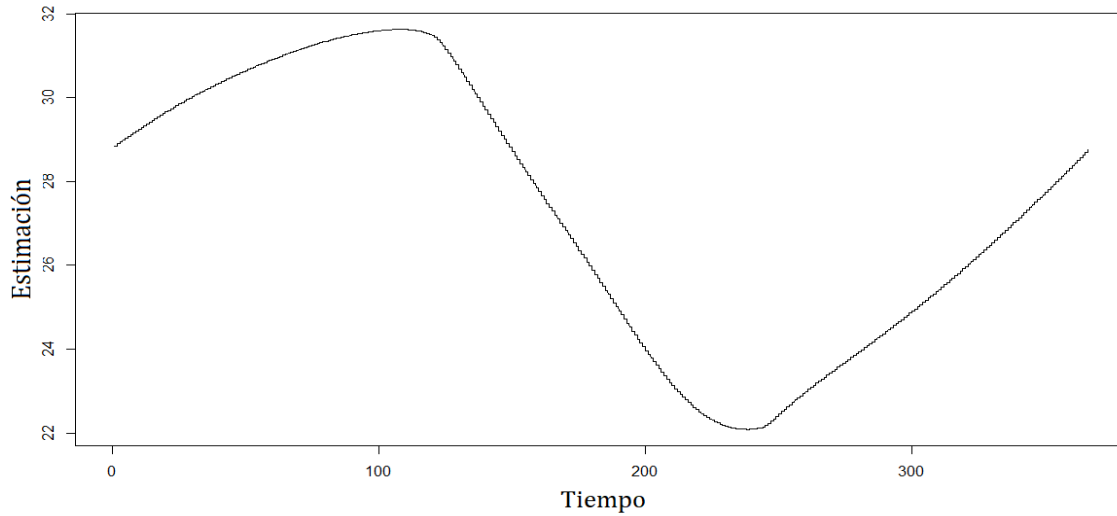
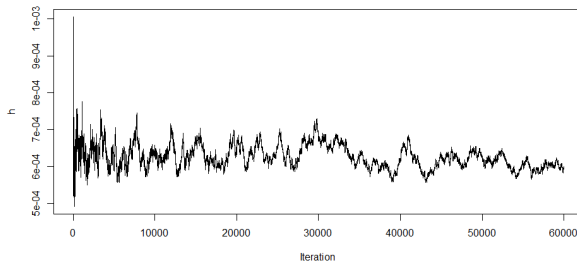
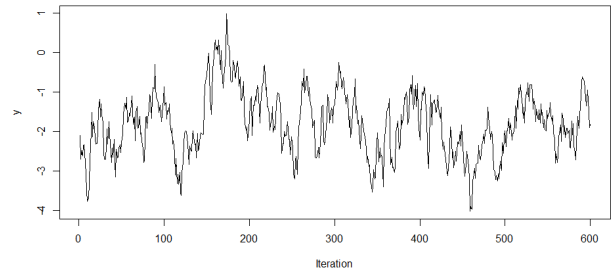


Figura 20: Estimación componente temporal de $\mu(t)$.



((a)) Rezago 5.



((b)) Rezago 10.

Figura 21: MCMC gráficas diagnóstico. Izquierda: Gráfica de valores de h tomados del algoritmo. Derecha: Gráfica de la cadena salvada de las celdas [6; 32].

Los diagnósticos de MCMC para la cadenas se basa en la salida completa de los datos volcados al disco. El gráfico de la Fig. 21 muestra los valores de h para el periodo de no quemado de la cadena, el algoritmo adaptativo se inició con $h = 1$, que rápidamente convergió a alrededor de $h = 0.0006$.

Para evaluar la autocorrelación del proceso en todas las celdas, se realizan las gráficas correspondientes en diferentes rezagos del proceso. En la figura 22 se puede observar que hay una alta auto correlación en el rezago 5, pero es bastante baja en toda la ventana de observación en el rezago 25. También se puede observar como disminuye la auto correlación a lo largo de todo el polígono a medida que aumenta el rezago. (Ver Fig. 22)

La Fig. 23 muestra el riesgo relativo medio de Monte Carlo, es decir, la media de $\exp\{Y\}$. Se puede ver como en $T = 50$, hay mucha más probabilidad de ocurrencia de la infracción C02 en algunos puntos de la ciudad que en otros, indicando como en coordenadas puntuales se pueden capturar carros estacionados en sitios prohibidos.

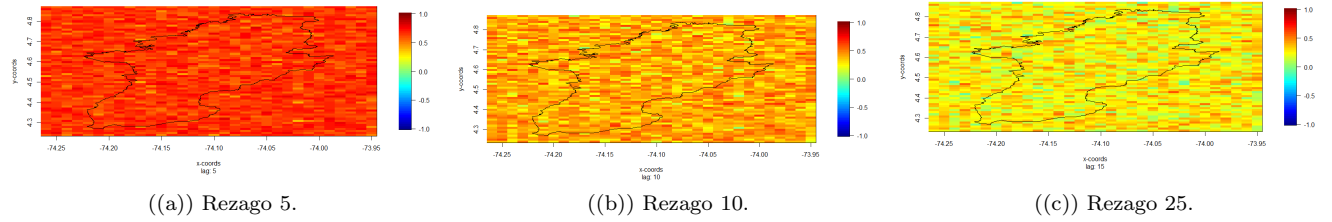


Figura 22: MCMC diagnóstico, izquierda to derecha: gráficos de autocorrelación de rezagos 5, 10 y 25 por celda.

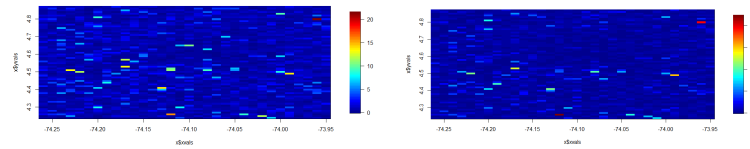


Figura 23: Gráficas del riesgo relativo medio de Monte Carlo (izquierda) y errores estándar asociados (derecha) en $T = 50$.

Cuando se corre el algoritmo en los tiempos $T = 50$, $T = 80$ y $T = 120$, es posible ver como cambian los riesgos relativos de la infracción en la misma ventana de visualización.

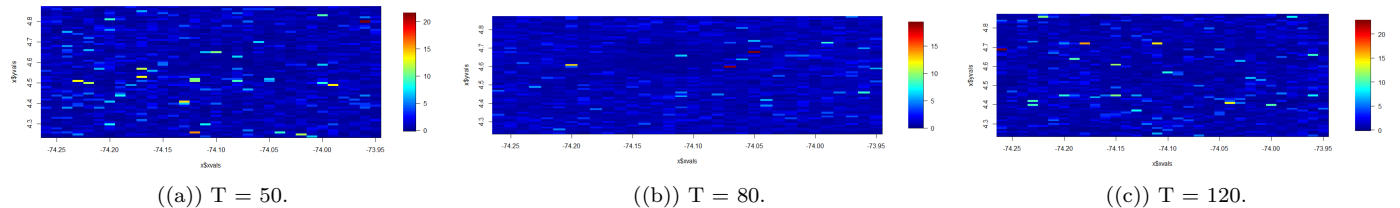
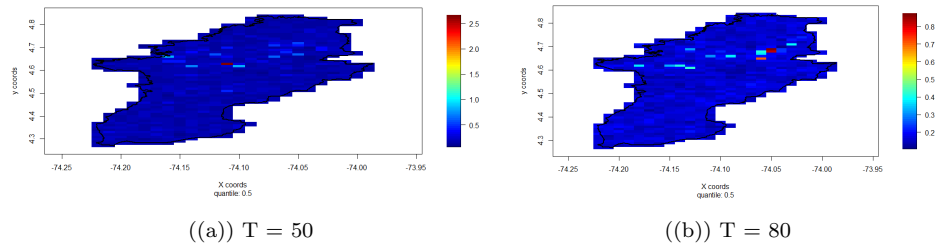
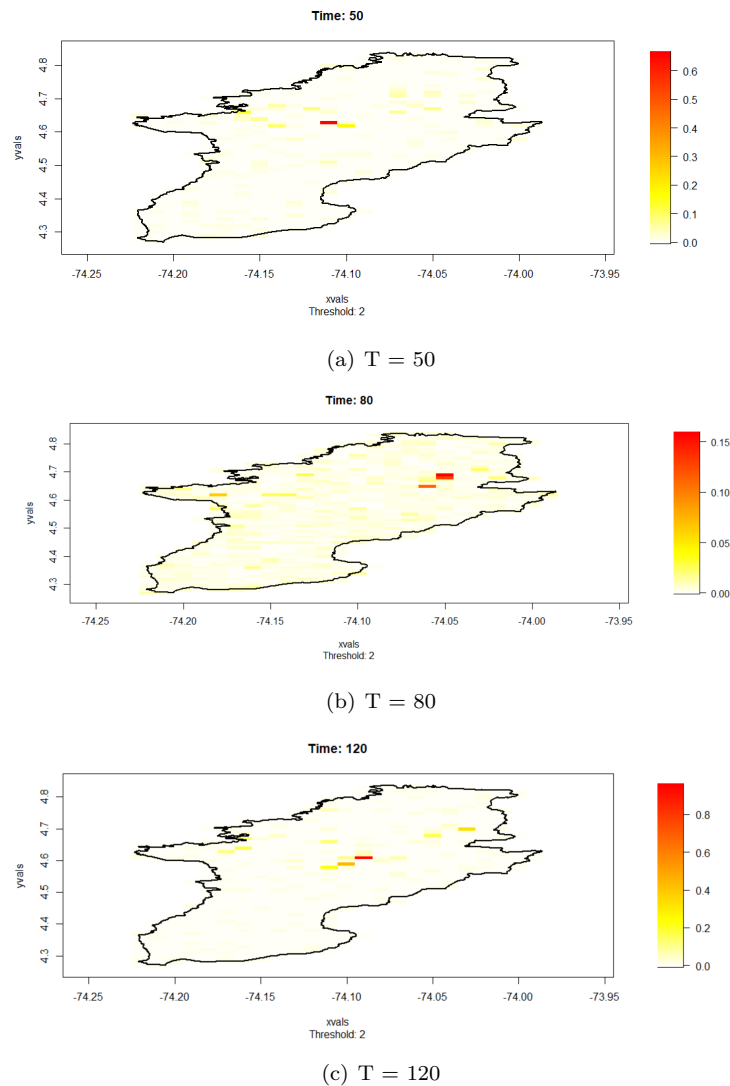


Figura 24: Media del riesgo relativo en diferentes tiempos de predicción.

En la Fig. 24, es posible ver la mediana del riesgo relativo en los tiempos $T = 50$ y $T = 80$, la cual indica, para el caso de $T = 50$, que el 50% de los riesgos relativos están por debajo de 2.5, y en ese punto específico de la ciudad hay 2.5 veces más de probabilidad de ocurrencia de la infracción con respecto a otros punto de la ciudad.

La Fig. 26 muestra la probabilidad de que el riesgo relativo sea mayor a 2, es decir, para $T = 120$, la probabilidad más alta está alrededor de 0.8, lo que quiere decir que en ese punto específico de la ciudad el riesgo de que ocurra una infracción tipo C02 es el doble a que ocurra en otro punto de la ciudad, con la probabilidad mencionada. Lo cual es de gran utilidad para saber en que puntos de la ciudad el riesgo de ocurrencia de la infracción es mayor en comparación con otros.

Figura 25: Mediana del riesgo relativo en $T = 50$ y $T = 80$ Figura 26: Probabilidad (riesgo relativo) > 2

5. Conclusiones y discusión

- Según un informe de la compañía estadounidense INRIX, Bogotá quedó en el primer puesto de las ciudades con más tráfico vehicular, esto debido a múltiples factores como son, el estado de las vías, el crecimiento del parque automotor y la cantidad de infracciones de tránsito que se cometen diariamente en la ciudad. Este trabajo brinda a la entidad encargada de la movilidad una herramienta para identificar puntos de la ciudad donde hay agregación de infracciones, el trabajo podría extenderse para estudiar infracciones relevantes para la ciudad, identificando no solo los puntos de calor en donde actualmente se concentran las infracciones, sino un pronóstico para determinadas temporadas del año. Con respecto al método elegido para abordar el problema, en el trabajo se abordaron cuatro metodologías con el fin de mostrar por qué aquella que muestra la interacción espacio tiempo es la más adecuada, aunque el análisis espacial presenta como una buena opción, es necesario llevar el análisis un paso más adelante incorporando el tiempo como una dimensión más, pues esta resulta ser relevante en el comportamiento de la infracción estudiada; en todos los análisis el patrón resulta siendo un patrón agregado, lo que indica que las infracciones si presentan un comportamiento específico en puntos de la ciudad, y lo más interesante es ver como también hay una influencia del tiempo en ese patrón agregado, pues quiere decir que en temporadas del año también hay una agregación de puntos a lo largo de la ciudad.
- Luego de ver las diferentes opciones para abordar un análisis espacio temporal, se puede concluir que la mejor opción es considerar el tiempo como una dimensión más, pues es aquella que concibe el tiempo desde su naturaleza. Aunque las opciones que consideran el tiempo como marca arrojan resultados de como se comporta el patrón, realmente no tienen en cuenta la dependencia temporal del fenómeno.
- Los meses que presentan mayor cantidad de infracciones son los meses de febrero y septiembre, y aquellos con menor cantidad de infracciones son los meses de enero, junio, julio y diciembre, que corresponden a la temporada vacacional y temporadas de fin e inicio de año.
- El proceso encontrado es agregado en algunas partes de la ciudad bajo la infracción C02 como son, la zona del aeropuerto internacional El Dorado, los alrededores de la estación de La Sabana, los corredores de las principales vías de la localidad de Chapinero como la cra. 15, la cra. 13 y la cra. 11., y la zona industrial de Puente Aranda, en muchas de las zonas mencionadas el costo del parqueadero es bastante alto, además de la ausencia de parqueaderos por la zona industrial, causa que las personas decidan estacionarse en las calles.
- El modelo presentado a través de la cadena de Monte Carlo, permite ver como se comporta el proceso en ambas componentes, en las Fig. 23 y 24, se muestra cual es la probabilidad de ocurrencia del evento en la ventana de estudio pero considerando la distribución espacial y el comportamiento temporal. Esta predicción se hace a tiempos futuros para ver, en dicho tiempo que podría suceder con el evento, lo cual es una de las grandes ventajas del trabajo.
- Acerca del análisis puramente espacial, se identificaron varios puntos de la ciudad en donde en todo el año hay alta recurrencia de las infracción de tránsito estudiada, uno de esos puntos es el aeropuerto El Dorado, este se debe ver en gran medida al alto costo del parqueadero, adicional que no hay otros parqueaderos públicos cerca. Otro punto importante donde ocurre esto, es el sector de la plaza de Paloquemao. En la localidad de Chapinero, el corredor de la carrera 15 desde la calle 72 hasta la calle 85 casi en todos los meses hay alta ocurrencia de estas infracciones, esto seguramente por el costo elevado de los parqueaderos en ese sector.
- Como se mencionó al inicio del trabajo, la gran mayoría de trabajos en un área similar no consideraron el tiempo como una tercera dimensión, en este trabajo se pudo evidenciar como ya existen las herramientas para poder trabajar modelos considerando las 3 dimensiones, y además obtener predicciones y pronósticos.

- Como trabajo futuro, se puede hacer una extensión del modelos haciendo uso de covariables y utilizando inferencia bayesiana también a partir del paquete Lcgp de R, los cuales permiten hacer diferentes modelos y establecer la influencia de las covariables en modelos espacio temporales de patrones puntuales, teniendo en cuenta que una de las dificultades del trabajo es el garantizar el ajuste del modelo debido a que se elaboró uno solo, lo ideal es poder tener una diversidad de modelos con el fin de comparar el ajuste y elegir el mejor.

Agradecimientos

Primeramente a la vida y a Dios por darme la oportunidad de estudiar y permitirme crecer profesionalmente, pues gracias a la maestría he encontrado grandes ofertas laborales y he tenido la oportunidad de crecer. A Andrés Rico, quien me acompañó en todo el proceso y decidió apoyarme, incluso en los momentos difíciles. A mi hermana por ayudarme a cumplir mis sueños y siempre ser incondicional, y finalmente a mis amigas, que han sido uno de los grandes pilares de mi vida.

Referencias

- Andrieu, C. & Thoms, J. (2008), ‘A tutorial on adaptive mcmc’, *Statistics and Computing* **18**.
- Baddeley, A., Rubak, E. & Turner, R. (2015), *Spatial Point Patterns: Methodology and Applications with R*, Chapman Hall/CRC Interdisciplinary Statistics, CRC Press.
- Bermudez Arias, S. C. (2016), Metodología para la evaluación espacio temporal de la accidentalidad vial en Bogotá: caso Avenida Boyacá, PhD thesis, Universidad Nacional de Colombia-Sede Bogotá.
- Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V. & Pebesma, E. J. (2008), *Applied spatial data analysis with R*, Vol. 747248717, Springer.
- Blazquez, C. A. & Celis, M. S. (2013), ‘A spatial and temporal analysis of child pedestrian crashes in santiago, chile’, *Accident Analysis Prevention* **50**, 304–311.
*<https://www.sciencedirect.com/science/article/pii/S0001457512001534>
- Castro, F. (2011), ‘Análisis espacial de los accidentes de tránsito en el cantón de pococí’, *Revista Geográfica de América Central* **2**, 1–43.
- Diggle, P. J., Moraga, P., Rowlingson, B., Taylor, B. M. et al. (2013), ‘Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm’, *Statistical Science* **28**(4), 542–563.
- Diggle, P., Rowlingson, B. & Su, T. (2005), ‘Point process methodology for onâline spatioâtemporal disease surveillance’, *Environmetrics* **16**, 423 – 434.
- El Espectador (2021), ‘<https://bit.ly/31R74R9>, 10 de Mmarzo de 2021’.
- Escobar, F. Á. C. (2013), ‘Análisis espacial de los accidentes de tráfico en bogotá dc fundamentos de investigación’, *Perspectiva Geográfica* **18**(1), 9–38.
- Gabriel, E., Rowlingson, B., Diggle, P. et al. (2013), ‘stpp: an r package for plotting, simulating and analyzing spatio-temporal point patterns’, *Journal of Statistical Software* **53**(2), 1–29.
- Giraldo, R. (2009), ‘Estadística espacial’, *Bogotá: Universidad Nacional de Colombia* .
- Jiang, X., Abdel-Aty, M. & Alamili, S. (2014), ‘Application of poisson random effect models for highway network screening’, *Accident Analysis Prevention* **63**, 74–82.
*<https://www.sciencedirect.com/science/article/pii/S0001457513004399>
- Moncada, I. (2018), Análisis espacio temporal de los accidentes de tránsito en Bogotá utilizando patrones puntuales, PhD thesis, Universidad Nacional de Colombia-Sede Bogotá.
- Muertes en eventos de transporte, Colombia, 2019* (2019), url<https://www.medicinalegal.gov.co/cifras-estadisticas/forensis>.
- Steenberghen, T., Aerts, K. & Thomas, I. (2010), ‘Spatial clustering of events on a network’, *Journal of Transport Geography* **18**(3), 411–418. Tourism and climate change.
*<https://www.sciencedirect.com/science/article/pii/S0966692309001240>
- Taylor, B. M., Davies, T. M., Rowlingson, B. S. & Diggle, P. J. (2013), ‘lgep: An R package for inference with spatial and spatio-temporal log-Gaussian Cox processes’, *Journal of Statistical Software* **52**(4), 1–40.
*<http://www.jstatsoft.org/v52/i04/>
- Taylor, B. M., Davies, T. M., Rowlingson, B. S. & Diggle, P. J. (2015), ‘Bayesian Inference and Data Augmentation Schemes for Spatial, Spatiotemporal and Multivariate Log-Gaussian Cox Processes in R’, *Journal of Statistical Software* **63**(i07).
*<https://ideas.repec.org/a/jss/jstsof/v063i07.html>