Received: 24 Mar 2020 | Accepted: 30 Jul 2020 | Available online: 16 Oct 2020
DOI: http://dx.doi.org/10.18180/tecciencia.2020.29.2

TECCIENCIA

UNIVERSIDAD · ECCI

RESEARCH ARTICLE

Engineering

Modelos Gráficos Probabilísticos Aplicados al Análisis Espacial en R: Hurtos de Celulares en Bogotá

Probabilistic Graphical Models Applied to Spatial Analysis in R: Cell Phone Thefts in Bogotá

Danna Lesley Cruz Reyes [1,2]

[1]Departamento de Estadística, Universidade Federal de Minas Gerais, Minas Gerais, Brazil.

[2]Escuela de Medicina y Ciencias de la Salud (EMCS), Universidad del Rosario, Bogotá, Colombia.

Correspondence
Danna Lesley Cruz Reyes
Email: dcruzreyes@gmail.com

The publication of this journal is funded by Universidad ECCI, Bogotá-Colombia.

Editors: Robert Paul Salazar.

Editorial assistant : Luz Adriana Suáres Suáres.

Resumen. Los avances tecnológicos recientes permiten la recopilación, el procesamiento y el almacenamiento de información a gran escala. Esto ha determinado el comienzo del **big data**, donde el aumento de la información ha dado lugar a conjuntos de datos grandes y complejos que pueden ser potencialmente explotados para encontrar soluciones a problemas relevantes. Este trabajo tiene como objetivo explicar cómo los métodos estadísticos pueden analizar estos conjuntos de datos grandes y complejos, específicamente datos espaciales. Se realiza un análisis de dependencia espacial por medio de la un grafo que caracteriza la estructura espacial y un modelo ampliamente utilizado conocidos como autorregresivo condicionales (CAR por sus siglas en ingles). Estos modelos son útiles para obtener distribuciones conjuntas multivariadas de un vector aleatorio basado en especificaciones condicionales univariadas. Estas especificaciones condicionales se basan en las propiedades de Markov, de modo que la distribución condicional de un componente del vector aleatorio depende solo de un conjunto de vecinos, definido por el grafo. Los modelos autorregresivos condicionales son casos particulares de campos aleatorios de Markov. Finalmente, se explica como realizar estos análisis en R, incluyendo el manejo de grafos y los paquetes utilizados. Se realiza la estimación de los parámetros en R siguiendo la metodología bayesiana a un conjunto de datos que corresponde al robo de celulares en Bogotá.

keywords: Modelos atorregresivos condicionales, big data, programmación en R.

How to cite
Danna Cruz, **Probabilistic Graphical Models Applied to Spatial Analysis in R: Cell Phone Thefts in Bogotá**, TECCIENCIA, Vol. 15, No. 29, 9-22, 2020
DOI:http://dx.doi.org/10.18180/tecciencia.2020.29.2

[*] Equally contributing authors.

Abstract

Recent technological advances allow large-scale collection, storage and processing information. As a consequence textbf big data has become more important nowadays, since the increase in information has given rise to large and complex data sets that can be potentially exploited to find solutions to relevant problems. This work aims to explain how statistical methods can analyze these large and complex data sets, specifically spatial data. A spatial dependency analysis is carried out by means of a graph that characterizes the spatial structure and a widely used approach known as Conditional Auto-Regressive (CAR). These models are useful for obtaining multivariate joint distributions of a random vector based on uni-variate conditional specifications. These conditional specifications are based on the Markov properties. Hence, that the conditional distribution of a component of the random vector depends only on a set of neighbors defined by the graph. CAR models are particular cases of random Markov fields. Finally, it is explained how to carry out these analyzes in R language programming including the handling of graphs and the packages used. Finally, the parameters estimation in R is carried out following the Bayesian methodology to data corresponding to stolen cell phones in Bogotá-Colombia.

keywords: Conditional Auto-Regressive, big data, R programming.

1 | INTRODUCCIÓN

Los modelos estadísticos para analizar datos espaciales se dividen en dos clases generales: modelos geoestadísticos con soporte espacial continuo y modelos en una lattice, también llamados datos de área, donde los datos se producen en una cuadrícula (posiblemente irregular), con un conjunto enumerable de vértices o ubicaciones. Estos modelos autoregresivos se utilizan en muchos campos, incluyendo la cartografía de las tasas de infecciones [1], agricultura [2], econometría [3], ecología [4] y el análisis de imágenes [5]. En este trabajo, se presenta el modelo CAR como un ejemplo de los campos aleatorios gaussianos de Markov. El modelo CAR es un ejemplo dos campos aleatorios de Gaussian Markov [6]. Varios autores han analizado este modelo, mostrando las características de las covarianzas y las correlaciones dada una estructura espacial, por ejemplo, el modelo CAR produce variaciones no constantes en cada sitio, así como covarianzas desiguales entre regiones separadas por el mismo número de vecinos (ver [5], [7]), [8] estudió ampliamente la estructura de covarianza a priori que conlleva el modelo CAR. Encontró que la relación entre vecinos no parece tener explicación, ya que nodos con el mismo numero de vecinos y en vecindades similares tienen covarianzas diferentes. Luego, [7] en su articulo esclarece todos estas rarezas y concluye que el modelo CAR es afectado por sus vecinos de mayor orden.

El modelo CAR visualiza el dominio geográfico como un grafo no dirigido con un vértice en cada región y una arista entre dos vértices si las regiones correspondientes comparten un borde geográfico. Esto crea vecinos bien definidos para cada región, que se utilizan para definir la distribución conjunta o condicional. La distribución será la distribución multivariante normal.

El sistema de vecindad es un punto clave en los modelos autorregresivos (CAR) que se usan comúnmente en estadísticas espaciales. Para este caso, los grafos que apoyan la construcción del GMRF serán aquellos que expresen estas estructuras de vecindad. En este contexto, las aristas $\mathscr{E}$ en el grafo $\mathscr{B} = (\mathscr{G}, \mathscr{E})$, representan las conexiones en la estructura geográfica y, en consecuencia, definen los vecinos que se utiliza para modelar la dependencia espacial.

Los componentes del vector $\boldsymbol{\theta}$ son nodos del grafo. Supongamos que $\theta_1, ..., \theta_n$ sean las observaciones realizadas en las áreas $1, \ldots, n$. Denotemos por $j \sim i$ que el nodo j es vecino del nodo i. El término condicional, en el modelo CAR se usa porque cada elemento del proceso aleatorio se especifica condicionalmente en los valores de los nodos vecinos. En este articulo se presenta cuales comando pueden ser utiles para construir

mapas, grafos y las matrices de adyacencia necesarias para este tipo de análisis, a manera de ejemplo se presenta el mapa de Colombia y el uso de este modelo en R aplicado al caso de robos de celulares en Bogotá.

2 | CAMPOS DE MARKOV GAUSSIANOS

Los campos aleatorios son distribuciones multivariadas que, en general, se usan para describir la asociación espacial entre las variables $\boldsymbol{\theta}$. Un campo aleatorio de Markov extiende el concepto de cadena de Markov a un contexto espacial y supone que dicha distribución conjunta de $\boldsymbol{\theta}$ esta definida como sigue: sea un grafo $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ con n vértices $\mathscr{V}$ donde cada uno representa una de las componentes del vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ y $\mathscr{E}$ el conjunto de aristas que conecta dos vértices θ_i e θ_j. Asi,

$$f(\theta_i|\boldsymbol{\theta}_{-i}) = f(X_i|\boldsymbol{\theta}_{j\sim i}),$$

con $\boldsymbol{\theta}_{j\sim i}$ el vector formado por todos los componentes de $\boldsymbol{\theta}$ quienes son vecinos de i. En esta sección discutiremos brevemente los campos aleatorios gaussianos de Markov que a menudo se usan como una distribución *a priori* para efectos espaciales.

Un campo de Markov gaussiano aleatorio (GMRF por sus siglas en inglés) es un campo de Markov donde la distribución del vector aleatorio (de dimensión finita) y una distribución normal o gaussiana satisfacen los supuestos de independencia condicional. Se puede encontrar una discusión detallada de GMRF en ([6]) y ([7]).

Todos los resultados válidos para la distribución normal también serán válidos para un GMRF. Por lo tanto, en la siguiente sección presentamos los resultados más relevantes de la distribución normal multivariante. Después de definir formalmente un GMRF con todas las propiedades heredadas de la distribución normal, presentaremos la conexión entre el grafo $\mathscr{G}$ y los parámetros de la distribución normal multivariada $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$. Se mostrará que toda la información en el grafo se condensa en la matriz de covarianza $\boldsymbol{\Sigma}$ por medio de la matriz de precisión $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$, además, el vector promedio $\boldsymbol{\mu}$ no influirá en la estructura del vecindario del grafo.

3 | LA DISTRIBUCIÓN NORMAL MULTIVARIANTE.

Para facilitar la comprensión de los campos aleatorios gaussianos de Markov, revisamos la distribución normal multivariante y algunas de sus propiedades básicas.

Un vector aleatorio n-dimensional $\boldsymbol{\theta}_{n\times 1} = (\theta_1, \theta_2, \ldots, \theta_n)^T$, $n < \infty$ tiene una distribución n–variada con vector de medias $\boldsymbol{\mu}_{n\times 1}$ y matriz de covarianza $\boldsymbol{\Sigma}_{n\times n}$ si su función de densidad de probabilidad (pdf) toma la siguiente forma:

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\}, \quad \boldsymbol{\theta}\in\mathbb{R}^n. \tag{1}$$

Esta distribución será denotada por $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ donde $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ son tales que $\mu_i = E(\theta_i)$, $\boldsymbol{\Sigma}_{ij} = Cov(\theta_i, \theta_j)$, $\boldsymbol{\Sigma}_{ii} = Var(\theta_i)$ y $Corr(\theta_i, \theta_j) = \boldsymbol{\Sigma}_{ij}(\boldsymbol{\Sigma}_{ii}\boldsymbol{\Sigma}_{jj})^{-1/2}$.
Para presentar algunas propiedades de la distribución dada en (1), consideramos la siguiente partición: $\boldsymbol{\theta} = (\boldsymbol{\theta}_A, \boldsymbol{\theta}_B)^T$, $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, y

$$\boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_A\\ \boldsymbol{\mu}_B\end{pmatrix}, \quad \begin{bmatrix}\boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB}\\ \boldsymbol{\Sigma}_{AB} & \boldsymbol{\Sigma}_{BB}\end{bmatrix}$$

donde $\#(A) + \#(B) = n$. Suponiendo tal partición, algunas propiedades básicas de la distribución normal son:

- $\boldsymbol{\theta}_A \sim N(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_{AA})$ es la distribución marginal del vector $\boldsymbol{\theta}_A$ de orden $A \times 1$;
- $\boldsymbol{\Sigma}_{AB} = 0$ si y solo si $\boldsymbol{\theta}_A$ e $\boldsymbol{\theta}_B$ son independientes;
- La distribución condicional de $\boldsymbol{\theta}_A$ dado $\boldsymbol{\theta}_B$ es $N(\boldsymbol{\mu}_{A|B}, \boldsymbol{\Sigma}_{A|B})$ donde,

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}(\boldsymbol{\theta}_A - \boldsymbol{\mu}_B) \quad \text{y} \tag{2}$$

$$\boldsymbol{\Sigma}_{A|B} = \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}; \tag{3}$$

- Si $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y $\boldsymbol{\theta}' \sim N(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$, entonces $\boldsymbol{\theta} + \boldsymbol{\theta}' \sim N(\boldsymbol{\mu} + \boldsymbol{\mu}', \boldsymbol{\Sigma} + \boldsymbol{\Sigma}')$.

4 | DEFINICIÓN BÁSICA Y PROPIEDADES DE GMRF

Para construir un GMRF consideramos un grafo $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ con vértices n donde cada vértice representa uno de los componentes del vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ y las aristas conectan nodos que tienen algún tipo de asociación. Un GMRF supone que $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y que las aristas del grafo conectan nodos i y j si y solo si $\theta_i \not\perp \theta_j | \boldsymbol{\theta}_{-ij}$, es decir, si θ_i es independiente de θ_j, dados los componentes de $\boldsymbol{\theta}$ excepto θ_i y θ_j.

Teorema 4.1. *Si $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{Q})$, entonces para $i \neq j$,*

$$\theta_i \perp \theta_j | \boldsymbol{\theta}_{-ij} \Leftrightarrow Q_{ij} = 0.$$

La demostración se puede encontrar en [6]. Por lo tanto, este resultado establece que los componentes no nulos de $\boldsymbol{Q}$ determinan la relación de vecindad presente en $\mathscr{G}$. Esto implica que cualquier distribución normal con una matriz de covarianza definida positivamente también es un GMRF y viceversa. Por lo tanto, un GMRF se define formalmente de la siguiente manera:

Definición 4.2. Un vector aleatorio $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^T \in \mathbb{R}^n$ es llamado GMRF correspondiente a un grafo $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ con media $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{Q} > 0$, si y solamente si la función de densidad de probabilidad de $\boldsymbol{\theta}$ tiene la siguiente forma:

$$\pi(\boldsymbol{\theta}) = (2\pi)^{-n/2}|\boldsymbol{Q}|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^T \boldsymbol{Q}(\boldsymbol{\theta} - \boldsymbol{\mu})\right),$$

donde la matriz $\boldsymbol{Q}$ cumple la siguiente condición:

$$Q_{ij} \neq 0 \Leftrightarrow \{i, j\} \in \mathscr{E}, \forall i \neq j.$$

Si $\boldsymbol{Q}$ es una matriz completamente densa, entonces $\mathscr{G}$ está completamente conectado, es decir, el vértice está conectado a todos los demás vértices del grafo. En este artículo, consideramos el caso donde $\boldsymbol{Q}$ es esparsa, es decir, la mayoría de las componentes de la matriz son cero.

Teorema 4.3. *Sea un grafo $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ que representa $\boldsymbol{\theta}$ un GMRF, con media $\boldsymbol{\mu}$ y la matriz de precisión $\boldsymbol{Q}$ simétrica y definitiva positiva. Luego, la distribución de cada componente θ_i de $\boldsymbol{\theta}$, dado el vector $\boldsymbol{\theta}_{-i}$ formado por todos los componentes de $\boldsymbol{\theta}$, excepto θ_i es una distribución normal tal que:*

$$\begin{aligned} E(\theta_i|\theta_{-i}) &= \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij}(\theta_j - \mu_j), \\ Prec(\theta_i|\theta_{-i}) &= Q_{ii}, \\ Corr(\theta_i, \theta_j|\boldsymbol{\theta}_{-ij}) &= -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, \quad i \neq j, \end{aligned}$$

donde $i \sim j$ denota que el nodo j es vecino de nodo i.

5 | ESPECIFICACIÓN DE GMRF A TRAVÉS DE CONDICIONALES COMPLETAS

Una alternativa a la construcción de un GMRF es considerar las condicionales completas de las distribuciones $\pi(\theta_i|\theta_{-i})$. Este enfoque pionero se adoptó en ([9]) para construir los modelos autorregresivos condicionales, conocidos como modelos *CAR*. Suponga que la distribución condicional completa de θ_i dada $\boldsymbol{\theta}_{-i}$ para todos $i = 1, \ldots, n$ es una distribución normal con media y precisión dadas respectivamente por:

$$E(\theta_i|\boldsymbol{\theta}_{-1}) = \mu_i - \sum_{j:j\sim i} \beta_{ij}(\theta_j - \mu_j), \tag{4}$$

$$Prec(\theta_i|\boldsymbol{\theta}_{-i}) = \kappa_i, \tag{5}$$

con $\kappa_i > 0$, donde la suma en (4) está sobre todos los j de i vecinos. El siguiente teorema nos muestra que esta densidad conjunta de $\boldsymbol{\theta}$ existe y que es única.

Teorema 5.1. *Si θ_i, dado $\boldsymbol{\theta}_{-i}$, $\forall i = 1, \ldots, n$, tiene una distribución normal, con media y precisión dadas en (4) y (5), por lo que la distribución conjunta de $\boldsymbol{\theta}$ es un GMRF con vector de medias $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$ y matriz de precisión $\boldsymbol{Q}$, de modo que*

$$Q_{ij} = \begin{cases} \kappa_i\beta_{ij} & i \neq j \\ \kappa_i & i = j \end{cases} \tag{6}$$

en que $\kappa_i\beta_{ij} = \kappa_j\beta_{ji}$, $i \neq j$.

La demostración de este resultado se puede encontrar en ([6]) y la demostración de los resultados de la aplicación de la Expansión Brook se muestra a continuación y cuya demostración se puede encontrar en ([7]).

Teorema 5.2. *Si $f(\theta_1|\boldsymbol{\theta}_{-1}), f(\theta_2|\boldsymbol{\theta}_{-2}), \ldots, f(\theta_n|\boldsymbol{\theta}_{-n})$ son las n distribuciones condicionales completas de una densidad conjunta $f(\boldsymbol{\theta})$ y si $\boldsymbol{x}$ es una configuración de referencia cualquiera tal que $f(\boldsymbol{x}) > 0$, entonces la densidad conjunta $f(\boldsymbol{\theta})$ se puede obtener menos de una constante de integración de la siguiente relación:*

$$\frac{f(\boldsymbol{\theta})}{f(\boldsymbol{x})} = \frac{f(\theta_1|\theta_2, \ldots, \theta_n)}{f(x_1|\theta_2, \ldots, \theta_n)} \frac{f(\theta_2|x_1, \theta_3, \ldots, \theta_n)}{f(x_2|x_1, \theta_3, \ldots, \theta_n)} \cdots \frac{f(\theta_n|x_1, \ldots, x_{n-1})}{f(x_n|x_1, \ldots, x_{n-1})}$$

$$\prod_{i=1}^{n} \frac{f(\theta_i|\theta_1, \ldots, \theta_{i-1}, x_{i+1}, \ldots, x_n)}{f(x_i|\theta_1, \ldots, \theta_{i-1}, x_{i+1}, \ldots, x_n)}.$$

5.1 | Un ejemplo de Campos de Markov Gaussianos: el modelo CAR

Considere una región geográfica que está dividida en subregiones n indexadas por enteros $1, 2, \ldots, n$. Suponga que esta colección de subregiones está dotada de un sistema de vecindad $\{V_i : i : 1, \ldots, n\}$, donde V_i denota la colección de subregiones que, en un sentido bien definido, son vecinos de la subregión i. En términos geográficos,

$$V_i = \{j : \text{subregiones } i \text{ y } j \text{ que comparten frontera}\}, \quad \text{para} \quad i \in \{1, 2, \ldots, n\},$$

El modelo CAR supone que las distribuciones condicionales completas para efectos aleatorios correspondientes a las variables de respuesta para cada $y_1, \ldots, y_n$, definidas como $\theta_1, \ldots, \theta_n$ y dadas por $\boldsymbol{\theta} = \theta_i|\theta_{-i}$, $i = 1, \ldots, n$,

siguen distribuciones normales con media y precisión dadas respectivamente en (4) y (5) suponiendo que $\beta_{ij} = \rho_{\mathscr{G}}/d_i^{\mathscr{G}}$, si $i \sim j$, y $\beta_{ij} = 0$ caso contrario en que $\kappa_i = d_i^{\mathscr{G}}/\boldsymbol{\Sigma}_{\mathscr{G}}^2$, esto es,

$$\theta_i|\theta_{-i} \sim N(\mu_i + \rho_{\mathscr{G}}\overline{\theta_i}, \frac{\boldsymbol{\Sigma}_{\mathscr{G}}^2}{d_i^{\mathscr{G}}}), \tag{7}$$

donde $\boldsymbol{\Sigma}_{\mathscr{G}}^2/d_i^{\mathscr{G}}$ es la variante condicional de $\theta_i|\theta_{-i}$, $\rho_{\mathscr{G}}$ es una constante de proporcionalidad, $d_i^{\mathscr{G}}$ es el número de vecinos del nodo i en el grafo $\mathscr{G}$, la media de los vecinos del nodo i es $\overline{\theta_i} = \sum_{\mathscr{E}^{\mathscr{G}}}(d_i^{\mathscr{G}})^{-1}(\theta_j)$ y $\mathscr{E}^{\mathscr{G}} = \{(i,j) \in E(\mathscr{G}) : j \sim i\}$ es el conjunto de aristas que pertenecen al grafo $\mathscr{G}$.
Sea la matriz de adyacencia:

$$A_{\mathscr{G}} = \begin{cases} 0, & \text{si} \quad i = j \\ 1, & \text{si} \quad i \sim j \\ 0, & \text{si} \quad i \not\sim j \end{cases}$$

tome $M_{\mathscr{G}} = diag\{d_1^{\mathscr{G}}, d_2^{\mathscr{G}}, \ldots, d_n^{\mathscr{G}}\}$. No es inmediato que (7) conduzca a una distribución conjunta completa de $\boldsymbol{\theta}$. [9] usa la expansión de Brook y muestra que cuando la matriz $(M_{\mathscr{G}} - \rho_{\mathscr{G}} A_{\mathscr{G}})^{-1}$ es definida positiva y simétrica, la distribución conjunta de $\boldsymbol{\theta}$ es:

$$\boldsymbol{\theta} \sim N(\mathbf{0}, (\boldsymbol{\Sigma}_{CAR}^{\mathscr{G}})^{-1}),$$

donde $(\boldsymbol{\Sigma}_{CAR}^{\mathscr{G}})^{-1} = \boldsymbol{\Sigma}_{\mathscr{G}}^2(M_{\mathscr{G}} - \rho_{\mathscr{G}} A_{\mathscr{G}})^{-1}$. Para que la matriz de covarianza sea definida positiva, es necesario que $\rho_{\mathscr{G}} < \frac{1}{\lambda_1}$ donde λ_1 es el valor propio más pequeño de la matriz $M_{\mathscr{G}}^{-1/2} A_{\mathscr{G}} M_{\mathscr{G}}^{-1/2}$. La demostración se puede encontrar en ([10]).
En resumen, el modelo CAR esta definida por una estructura de correlación inducida por el grafo de la siguiente forma: la región de estudio V se divide en n unidades de área sobre el conjunto de regiones $\{V_i : i : 1, \ldots, n\}$ que están vinculados a un conjunto correspondiente de respuestas $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ distribuidas condicionalmente como:

$$y_j|\theta_j \sim N(\mu + \theta_j, \boldsymbol{\Sigma}_y^2), \quad \text{e} \quad (\theta_1, \ldots, \theta_n)^t \sim N(\mathbf{0}, (\boldsymbol{\Sigma}_{CAR}^{\mathscr{G}})^{-1}), \quad \text{para cada} \quad j \in \{1, \ldots, n\}.$$

Así, el patrón espacial en la respuesta está modelado por $\boldsymbol{\theta}$. En particular, su uso es fundamental en los modelos bayesianos.

5.2 | Definición de las matrices del modelo CAR en R: División política de Colombia

En esta sección se muestra un ejemplo para formular las matrices correspondientes al modelo CAR. Además, se muestra como diseñar estas matrices en `R` a partir de mapas y, a manera de ejemplo, se utiliza el mapa de Colombia teniendo como base el formato de `ggplot2`. En primer lugar, se cargan los siguientes paquetes:

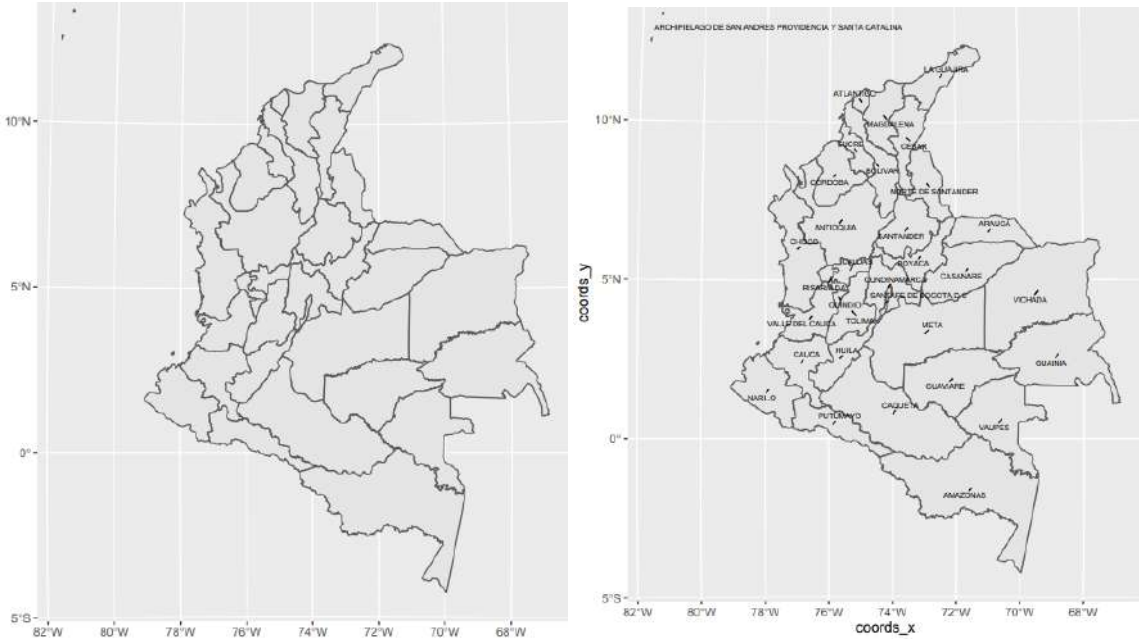
```
library(viridis)
library(ggrepel)
library(ggplot2)
library(spdep)
```

Luego, en la internet, se debe buscar el archivo tipo shapefile (`.shp`) del mapa y los atributos deseados. Por ejemplo, se buscó el mapa de Colombia con la división política en la página web[1] llamado `depto.shp`. En esta página web, una gran cantidad de mapas de Colombia con diferentes características, incluyendo el mapa con las Fronteras Marinas de Colombia y la División municipal de Colombia, además de otros mapas de Bogotá y Sudamérica. Otras páginas web con descargas libres y con gran cantidad de mapas y atributos de Colombia

[1] https://sites.google.com/site/seriescol/shapes

son: Sistema de Información Ambiental para Colombia, SIAC [2] y la página web de los Datos Abiertos de Bogotá a través del ViveLab Bogotá de la Universidad Nacional de Colombia[3] entre otras.

Después de descargar el archivo *.shp*, se importa a R con los siguientes comandos:



(a) Mapa generado sin atributos (b) Mapa generado con la división política de Colombia

FIG. 1 Mapa con la división política de Colombia

```
coL_Dep1 <- readOGR("depto.shp")
col_Dep <- st_as_sf(coL_Dep1)
```

El comando `st_as_sf` convierte un objeto desconocido en un objeto `shapefile`, para que pueda ser utilizado sin problemas en las funciones del paquete `ggplot2`. El siguiente comando genera el mapa de Colombia sin atributos (1a).

```
ggplot(col_Dep) + geom_sf()
```

Otra opción es crear variables asociadas a alguna característica, como sería el caso de asignar los nombres de los departamentos a alguna variable, esto se puede realizar utilizando el comando `mutate`. Una forma de hacerlo, es generando puntos ubicados en el centro de los unidades, estos puntos se llaman **centroides**. Para generarlos en R, se crean las variables: `centroid, coords, coords_xy coords_y` utilizando los siguientes comandos:

```
coords <- st_centroid(st_geometry(col_Dep), of_largest_polygon=TRUE)
ind <- row.names(col_Dep)
col_Dep <- col_Dep %>% mutate(centroid = map(geometry,
 st_centroid), coords = map(centroid,  st_coordinates),
 coords_x = map_dbl(coords, 1), coords_y = map_dbl(coords,
2))
```

Así, la Figura (1b) fue generada con las variables mencionadas anteriormente con `ggplot` y el paquete `ggrepel` para generar los textos:

```
ggplot(data = col_Dep) +
geom_sf()+
```

[2] http://www.siac.gov.co
[3] https://datosabiertos.bogota.gov.co

```
geom_text_repel(mapping = aes(coords_x,
coords_y, label = NOMBRE_DPT), size = 2,
min.segment.length = 0)
```

Ahora, para construir la matriz de adyacencia debemos construir el grafo de vecindad de los departamentos de Colombia en el modelo CAR, se consideran como vecinos a cualquier par de departamentos que comparten limite geográfico. Para hacer este grafo, primero se debe guardar las coordenadas de las unidades del mapa con el comando `st_centroid`:

```
coords <- st_centroid(st_geometry(col_Dep), of_largest_polygon=TRUE)
ind <- row.names(col_Dep)
suppressPackageStartupMessages(require(deldir))
```

Luego, se debe identificar el conjunto de vecindarios. Los archivos tipo `shapefile` son objetos usualmente representados por vectores, que consisten en la descripción de la geometría o forma (*shape*) de los objetos, algunos están descritos por medio de `polígonos` y otros por `multipolígonos`, se debe prestar atención a estas características. Si el archivo esta descrito por medio de `polígonos`, la función que se debe utilizar para generar el conjunto de vecindarios es `poly2nb` y si el archivo esta descrito por medio de multipolígonos, se debe utilizar `tri2nb`, ambas funciones pertenecen al paquete `spdep`. El archivo `depto.shp` esta descrito por medio de multipolígonos, así:

```
col.tri.nb <- tri2nb(coords, row.names=ind)
col.tri.nb
```

```
## Neighbour list object:
## Number of regions: 33
## Number of nonzero links: 178
## Percentage nonzero weights: 16.34527
## Average number of links: 5.393939
```

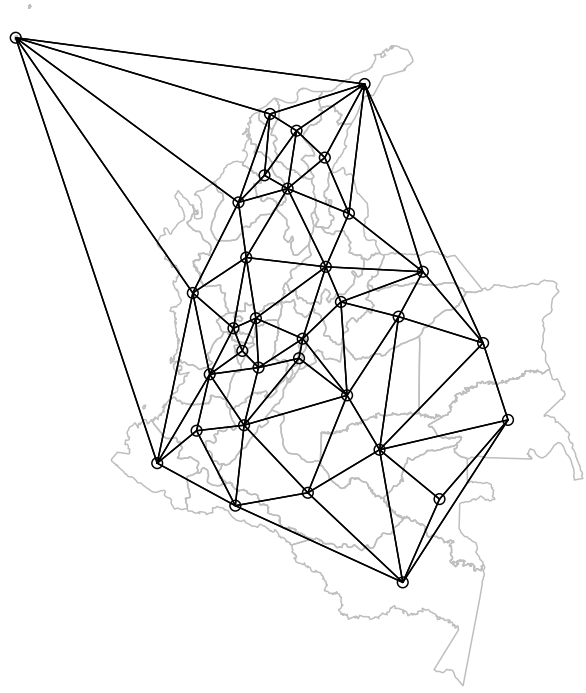


FIG. 2 Mapa de la división política de Colombia con el grafo de vecindad por adyacencia superpuesto.

El mapa en la Figura (2) de la división política de Colombia con el grafo de vecindad por adyacencia superpuesto es generado por medio de los siguientes comandos:

```
plot(st_geometry(col_Dep), border="grey")
plot(col.tri.nb, coords, add=TRUE)
shpnb.mat <- nb2mat(col.tri.nb, style="B",zero.policy=TRUE)
```

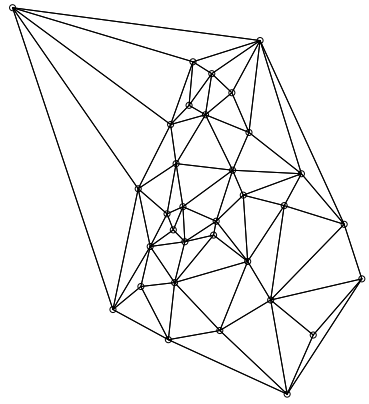
Y finalmente, para construir la matriz de adyacencia, la función `nb2mat`, permite convertir el conjunto de vecindarios en una matriz de adyacencia:

```
shpnb.mat <- nb2mat(col.tri.nb, style="B",zero.policy=TRUE)
```

Así, en la Figura (3b) podemos ver el grafo asociado al mapa de Colombia y en la Figura (3a) las primeras diez componentes de la matriz de adyacencia; esta matriz es de tamaño 33×33, ya que Colombia se divide administrativa y políticamente en 33 divisiones: 32 departamentos y un distrito capital, Bogotá.

	Antioquia	Atlántico	Bogotá	Bolivar	Boyacá	Caldas	Caquetá	Cauca	Cesar	Córdoba	...
Antioquia	0	0	0	1	0	1	0	0	0	1	
Atlántico	0	0	0	0	0	0	0	0	0	1	
Bogotá	0	0	0	0	0	0	0	0	0	0	
Bolivar	1	0	0	0	0	0	0	0	1	1	
Boyacá	0	0	0	0	0	0	0	0	0	0	
Caldas	1	0	0	0	0	0	0	0	0	0	
Caquetá	0	0	0	0	0	0	0	0	0	0	
Cauca	0	0	0	0	0	0	0	0	0	0	
Cesar	0	0	0	1	0	0	0	0	0	0	
Córdoba	1	1	0	1	0	0	0	0	0	0	
⋮											

(a) Primeras diez componentes de la matriz de adyacencia



(b) Grafo del mapa político de Colombia

FIG. 3 Matriz de adyacencia y grafo del mapa político de Colombia

5.3 | Ejemplo modelo CAR en R: hurto de celulares en Bogotá

Para este ejemplo, vamos a suponer que los valores y son el conjunto de datos de hurto de celulares del 01 de enero al 30 de noviembre del año 2018, estos fueron descargados de la página web de Datos Abiertos de Bogotá a través del ViveLab Bogotá de la Universidad Nacional de Colombia[4], el archivo se llama `"scat.shp"` y es importado a `R` por medio de los siguientes comandos:

```
shp <- readOGR("scat.shp")
```

La base de datos contiene el número de hurtos de celulares en cada uno de los 1168 barrios de Bogotá. Para graficar estos datos se propone dos métodos, uno con el comando propio de `R` y el otro con el paquete `leaflet`. El siguiente código genera la Figura 5a.

```
data.frame(shp)
par(mai=c(0,0,0,0))
plot(shp, col=2:7)
xy <- coordinates(shp)
points(xy, cex=0.2, pch=20, col='white')
```

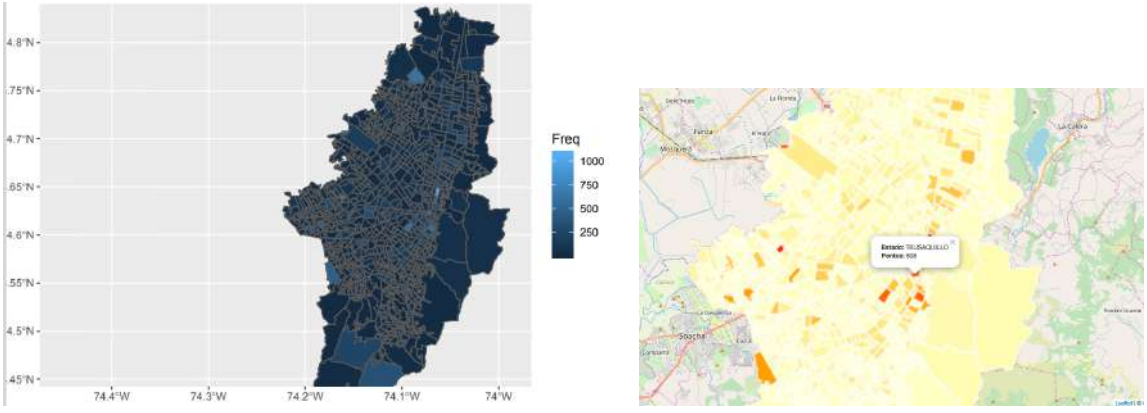
[4] https://datosabiertos.bogota.gov.co

`Leaflet` es una biblioteca *JavaScript* de código abierto muy popular para mapas interactivos. Se puede crear un mapa usando `Leaflet` llamando a la función `Leaflet()` y luego agregando capas al mapa usando funciones de capa, llamadas `layer`. Por ejemplo, podemos usar `addTiles()` para agregar un mapa de fondo, `addPolygons()` para agregar polígonos y `addLegend()` para agregar una leyenda. Podemos emplear una variedad de mapas de fondo.

Se puede crear el mapa de Bogotá con una escala de colores dada por la paleta `"YlOrRd"` del paquete `RColorBrewer` de la siguiente manera. Primero, transformamos el mapa al formato requerido por el paquete `Leaflet()`, hacemos esto usando la función `st_transform()` del `paquete sf`. En las Figuras (5a) y (5b) se presentan los mapas generados de Bogotá con las funciones descritas anteriormente y en las Figuras (4a) y (4b) se presenta los mismos mapas pero centrados en la zona urbana de Bogotá.

```
st_crs(map)
map <- st_transform(map, 4326)
pal <- colorBin("YlOrRd",domain = NULL, n=5)
state_popup <- paste0("<strong>Estado: </strong>",
  map$SCANOMBRE,
   "<br><strong>Pontos: </strong>",map$Freq)

leaflet(map) %>%
addTiles() %>%
addPolygons(
color = "white", fillColor = ~ pal(Freq),
fillOpacity = 1, popup = state_popup
) %>%
addLegend(pal = pal, values = ~Freq, opacity = 1)
```



(a) Mapa generado con la función **plot** (b) Mapa generado con la función **leaflet**

FIG. 4 Distribución de hurtos de celulares en el centro Bogotá

Antes de empezar a realizar el análisis con el modelo CAR, se puede calcular una medida de autocorrelación espacial. La estadística más utilizada para identificar autocorrelación espacial es el índice de Moran desarrollado por Patrick Alfred Pierce Moran, ([9]). Esta estadística es utilizada para realizar un test que permite evaluar la significancia de la correlación espacial bajo la hipótesis nula de que hay cero autocorrelación espacial presente en la variable. Este test de Moran creará una medida de correlación entre -1 y 1 en donde 1 determina una correlación espacial positiva perfecta 0 significa que nuestros datos están distribuidos aleatoriamente y -1 representa autocorrelación espacial negativa. En `R`, la implementación es muy sencilla, utilizando el comando `moran.test`. Para esto debemos generar un objeto de tipo `listw` (matriz de pesos):

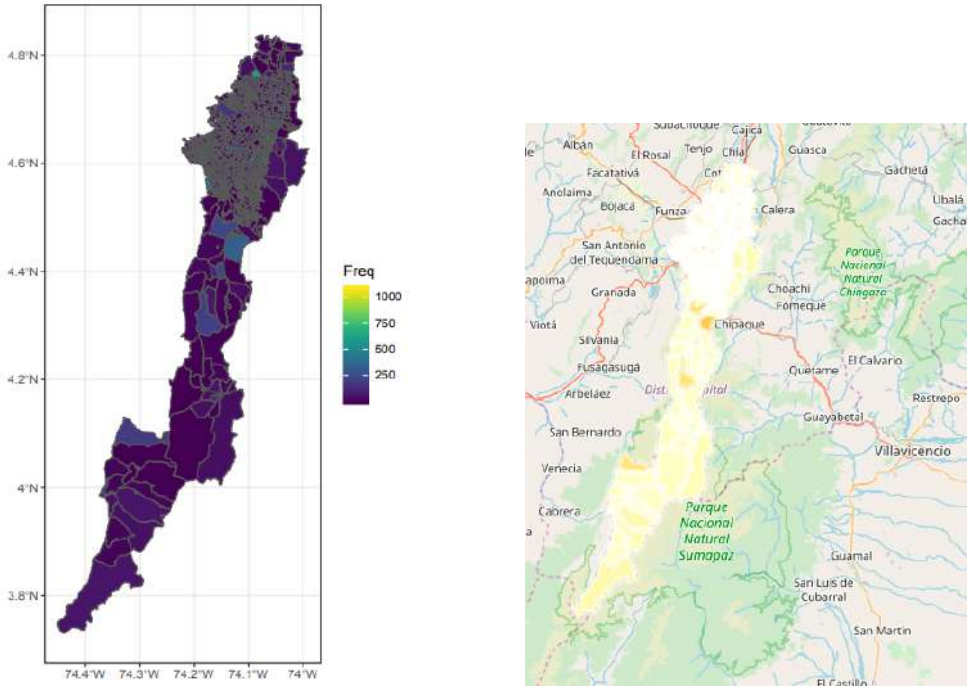
```
wl <- nb2listw(w)
moran.test(shp$Freq, listw = wl)
```

```
        Moran I test under randomisation

data:  shp$Freq
weights: wl

Moran I statistic standard deviate
= 8.7392, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation        Variance
0.1530097208     -0.0008568980     0.0003099895
```

El estadístico I de Moran confirma la existencia de autocorrelación espacial positiva en la distribución de hurtos de celulares. En efecto, con un valor estimado de 0.1530 y un p-valor de $2.2e-16$, debe rechazarse la hipótesis de aleatoriedad espacial y aceptar la presencia de autocorrelación espacial positiva. También nos muestra que existe una leve relación de correlación positiva, versus una expectativa de una leve relación negativa.



(a) Mapa generado con la función **plot** (b) Mapa generado con la función **leaflet**

FIG. 5 Distribución de hurtos de celulares en Bogotá

Así, con la evidencia estadístico I de Moran, se aplica el modelo CAR. Como no se considera variables regresoras, el modelo apenas va a describir el comportamiento espacial bajo el supuesto de que el número de hurtos se distribuye Normal. De la misma forma que en el mapa de Colombia, se genera el grafo y la matrices correspondientes para utilizar el modelo CAR. Así, con estos valores ajustamos el modelo:

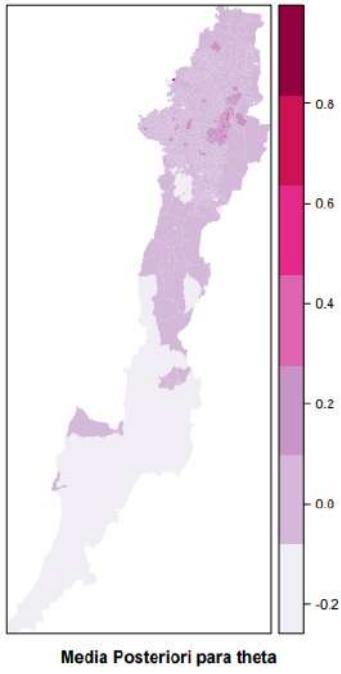
$$y_j|\theta_j \sim N(\mu+\theta_j, \mathbf{\Sigma}_y^2), \quad \text{e} \quad (\theta_1,...,\theta_n)^t \sim N(\mathbf{0}, (\mathbf{\Sigma}_{CAR}^{\mathscr{G}})^{-1}), \quad \text{para cada} \quad j \in \{1,\ldots,1168\},$$

con distribuciones *a priori* dadas por:

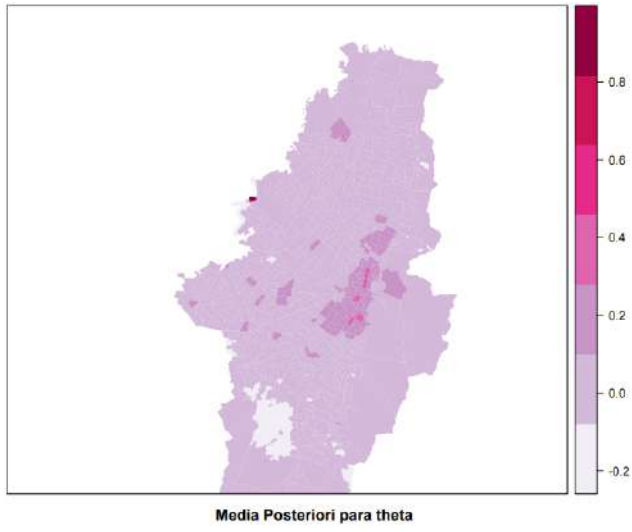
$$\begin{aligned} \Sigma_y^2, \Sigma_{\mathscr{G}}^2 &\sim InvGamma(0.1, 0.1), \\ \mu &\sim N(0, 100), \\ \rho &\sim Uniforme - Discreta(\{0.500, 0.501, \ldots, 0.999\}) \end{aligned}$$

y utilizando métodos MCMC se estima los parámetros $\Sigma_y^2, \Sigma_{\mathscr{G}}^2, \mu$ y ρ. En este artículo profundiza en este tema, pero se puede encontrar estos métodos en ([7]) o ([10]). En las Figuras (6a) y (6b)se presenta la media de los efectos aleatorios $\boldsymbol{\theta}$ *a posteriori*.

La estimación *a posteriori* de los valores de $\boldsymbol{\theta}$ mostrados en las Figuras (6a) y (6b) nos muestra un resultado relevante y es que existe evidencia de una autocorrelación espacial, esto es, el número de robos de celulares de Bogotá está asociado con elvecindario. Además, muestra que la variable se puede modelar siguiendo una distribución normal, ya que esta describe el comportamiento espacial mostrando valores de $\boldsymbol{\theta}$ mayores en la zona del centro de Bogotá, donde existe un mayor número de hurtos de celulares.



(a) Media *a posteriori* para θ en Bogotá



(b) Media *a posteriori* para θ en el área urbana Bogotá

FIG. 6 Media *a posteriori* para θ en Bogotá

Conclusión

Este documento contiene una descripción estadística de los hurtos de celulares registrados en los barrios de Bogotá, mostrando que existe dependencia espacial. Se mostró también como se pueden realizar estos análisis en R, presentando los comandos y los paquetes que pueden ser empleados para este tipo análisis. La idea es que las personas puedan utilizar estas herramientas en sus propios estudios, así como descargar información de internet, mapas, datos, ya que todo esto se encuentra disponible. Se efectúa la ubicación espacial de cada delito sobre el mapa de Bogotá: las mayores concentraciones delictivas se presentan con mayor intensidad en el centro de Bogotá y en algunos barrios al sur de la ciudad.

Sin embargo, este es apenas un análisis descriptivo, como complemento de las cifras de criminalidad, se

puede considerar variables regresoras que expliquen el comportamiento de hurtos de celulares. También se puede cambiar la distribución para la variable respuesta a una más adecuada, como la distribución Poisson, que podría ajustarse mejor al modelo. Todos estos análisis se pueden realizar por medio del uso del modelo CAR con una extensión natural modelando la media por medio de variables regresoras y cambiando la distribución Normal por una distribución Poisson.

REFERENCIAS

[1] P. Elliott and D. E. Wartenberg, "Spatial epidemiology: Current approaches and future challenges," in *Environmental health perspectives*, 2004. **DOI:** 10.1289/ehp.6735

[2] J. Besag and D. Higdon, "Bayesian analysis of agricultural field experiments," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 4, pp. 691–746, 1999. **DOI:** 10.1111/1467-9868.00201

[3] J. Lesage and R. Pace, "Introduction to spatial econometrics. crc press, boca raton, fl," *Introduction to Spatial Econometrics*, vol. 1, 01 2009. **DOI:** 10.1201/9781420064254

[4] O. Arslan and O. Akyürek, "Spatial modelling of air pollution from pm10 and so2 concentrations during winter season in marmara region (2013-2014)," *International Journal of Environment and Geoinformatics*, pp. 1 – 16, 2018. **DOI:** 10.30897/ijegeo.412391

[5] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 48, no. 3, pp. 259–302, 1986. **DOI:** 10.1111/j.2517-6161.1986.tb01412.x

[6] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, vol. 104 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall, 2005. **DOI:** 10.1201/9780203492024

[7] R. Assunção and E. Krainski, "Neighborhood dependence in bayesian spatial models," *Biometrical Journal*, vol. 51, no. 5, pp. 851–869, 2009. **DOI:** 10.1002/bimj.200900056

[8] M. M. Wall, "A close look at the spatial structure implied by the CAR and SAR models," *Journal of Statistical Planning and Inference*, vol. 121, no. 2, pp. 311–324, 2004. **DOI:** 10.1016/S0378-3758(03)00111-3

[9] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 192–236, 1974. **DOI:** 10.1111/j.2517-6161.1974.tb00999.x

[10] S. Banerjee, B. P. Carlin, and A. E. Gelfand, *Hierarchical Modeling and Analysis of Spatial Data*, vol. 101. 01 2004. **DOI:** 10.1201/9780203487808