# Comparación de las metodologías: Modelo Lineal Generalizado mixto marginal espacial con varianza CAR bajo respuesta Poisson y Modelo Lineal Generalizado Poisson Log-lineal con distribución subyacente gausiana en el estudio de datos de área

Por:

Ricardo Alberto Borda Hernández


UNIVERSIDAD NACIONAL DE COLOMBIA

FACULTAD DE CIENCIAS

ESCUELA DE ESTADÍSTICA

MEDELLÍN - ANTIOQUIA

NOVIEMBRE DE 2011

# Comparación de las metodologías: Modelo Lineal Generalizado mixto marginal espacial con varianza CAR bajo respuesta Poisson y Modelo Lineal Generalizado Poisson Log-lineal con distribución subyacente gausiana en el estudio de datos de área

Por:

Ricardo Alberto Borda Hernández

Presentado como requisito parcial para optar al título de
MAGISTER EN ESTADÍSTICA

Director:
Rene Iral Palomino - M.Sc Estadística
Kennet Roy Cabrera Torres - M.sc(c) Estadística

UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS
ESCUELA DE ESTADÍSTICA
MEDELLÍN - ANTIOQUIA
NOVIEMBRE DE 2011

UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS
ESCUELA DE ESTADÍSTICA

Los jurados abajo firmantes certifican que han leido y recomiendan a la **Facultad de Ciencias** aprobar el trabajo de grado titulado "**Comparación de las metodologías: Modelo Lineal Generalizado mixto marginal espacial con varianza CAR bajo respuesta Poisson y Modelo Lineal Generalizado Poisson Log-lineal con distribución subyacente gausiana en el estudio de datos de área**" presentado por **Ricardo Alberto Borda Hernández** como requisito parcial para optar al título de **Magister en Estadística**.

Fecha: Noviembre de 2011

Director: Rene Iral Palomino - M.Sc Estadística
Kennet Roy Cabrera Torres - M.sc(c) Estadística

Jurados:

# Índice general

# Índice de cuadros

# Índice de figuras

## Resumen

La estadística espacial es una herramienta que permite analizar información a partir de la ubicación espacial de las observaciones. Áreas del saber como: la geología, la minería, las ciencias ambientales, las ciencias sociales, entre otras, hacen parte de áreas que hoy en día pueden utilizar esta valiosa herramienta. Dentro de los estudios sociales, la estadística espacial se puede convertir en una herramienta que proporcione información consolidada que de pistas sobre las dinámicas sociales y culturales de la población, convirtiéndose en un excelente complemento al trabajo cualitativo propio de esta área.

La Isonimia o Isonomia estudia la distribución de la población a partir del análisis de frecuencia y distribución de apellidos de los pobladores con el fin de establecer relaciones de parentesco y origen. Hasta el momento, en esta clase de estudios no se ha tenido en cuenta el componente espacial y por ende se desconocen las técnicas espaciales. Esta necesidad permitió conjeturar sobre qué metodologías de la estadística espacial podría modelar mejor conteos georeferenciados y encontrar tendencias que los estudios típicos de Isonimia y la Estadística clásica no pueden hallar y así modelar la distribución de los pobladores más comunes en el departamento de Antioquia. Al indagar en la literatura, se encontraron múltiples esfuerzos encargados de modelar datos georeferenciados de este tipo, destacándose los modelos tradicionales simultaneos autoregresivos (SAR) y los modelos condicionales autoregresivos (CAR) cuya estimación es realizada por máxima verosimilitud y que parten del supuesto de variable continua normal en los datos; también, es posible modelar los datos por medio de un Modelo lineal generalizado Mixto (MLGM) por medio de seudo verosimilitud y Cadenas de Markov de Monte Carlo (MCMC) que parten de que la respuesta se distribuya como miembro de la familia exponencial.

El objetivo principal de este estudio, se basa en comparar, por medio de simulación, algunas metodologías de estadística espacial que estén interesadas en el modelamiento y predicción de datos de conteo. Un segundo objetivo, es mostrar si los MLGM proporcionan un mejor ajuste de los datos que los modelos tradicionales SAR y CAR, si es así, en futuros estudios se podría omitir el elaborado cálculo de la matriz de vecindad y el campo de aplicación sería mas amplio debido a que la variable respuesta no se limita a ser distribuida normalmente si no como miembro de la familia exponencial.

Por último, los modelos expuestos serán aplicados al modelamiento de los procesos distribucionales del departamento de Antioquia por medio de la georeferenciación de los apellidos más frecuentes.

***Palabras clave:*** datos de área, datos georeferenciados, modelos SAR, modelos CAR, modelos lineales generalizados mixtos, isonímia.

## Abstract

Spatial statistics is a tool to analyze data from the spatial location of the observations. Knowledge areas as: geology, mining, environmental sciences, social sciences, among others, are part of areas that can now use this valuable tool. In social studies, spatial statistics can become a tool that provides consolidated information for clues to the social and cultural dynamics of the population, making it an excellent complement to their own qualitative work in this area.

The isonymy studies the distribution of the population from the analysis of frequency and distribution of surnames of the settlers to establish relations of kinship and origin. So far, this type of study has not been taken into account the spatial component and hence unknown space techniques. This need allowed to guess on what spatial statistical methodologies could better model georeferenced counts and find that studies trends isonymy typical and classical statistics can not find and so model the distribution of common people in the department of Antioquia. When asked in the literature, many efforts were responsible for georeferenced data model on this type, especially the traditional models simultaneous autoregressive (SAR) and conditional autoregressive models (CAR) which are estimated by maximum likelihood and based on the assumption that variable normal continuous data, too it, is possible to model the data using a generalized linear model Mixed (MLGM) using pseudo-likelihood and using Markov Chain Monte Carlo (MCMC) to presume that the response is distributed as a member of the exponential family.

The main objective of this study, based on the comparison, by simulation, some spatial statistical methodologies that are interested in modeling and prediction of count data. A second objective is to show whether MLGM provide better data fit than traditional models SAR and CAR, if so, future studies could skip the calculation of the matrix developed neighborhood and the scope would be wider because the response variable is not restricted to be normally distributed if not as a member of the exponential family.

Finally, the models on display will be applied to modeling processes of the department of Antioquia distributional through georeferencing most frequent surnames.

# Capítulo 1

# Introducción

El modelamiento espacial en datos de área fue inicialmente realizado por medio de modelos Simultáneos Espaciales Autoregresivos (SAR) y por medio de modelos Condicionales Autoregresivos Espaciales (CAR) con respuesta gausiana, siendo este un limitante en la modelación de otros tipos de respuesta. Schabenberger [19], muestra que las estimaciones por máxima verosimilitud se vuelven numéricamente complejas, si la distribución de la variable respuesta $Z(s_i)$ no es gausiana. Por ejemplo, si la varaible respuesta corresponde a conteos o a respuestas dicotómicas.

Schabenberger [18] y[19], expone un enfoque en el estudio de datos de área basados en Modelos Lineales Generalizados (MLG) con el objetivo de moverse fuera de la distribución gausiana y poder trabajar con respuesta cuya distribución sea miembro de la familia exponencial. Sin embargo, las inferencias típicas usadas con este modelo requieren una distribución multivarada y cuando los datos son espacialmente autocorrelacionados no se puede construir con facilidad esta distribución como el producto de las marginales, como si es el caso de datos independientes. Este problema puede ser solucionado usando la especificación condicional del Modelo Lineal Generalizado Mixto (MLGM), en donde los efectos aleatorios corresponden a un modelo subyacente con distribución gausiana y una función de covarianza definida por el autor o investigador.

Schabenberger [19] afirma que no es necesario limitarse a los modelos SAR y CAR y que se puede crear cualquier representación paramétrica, siempre y cuando las matrices resultantes sean definidas positivas y los parámetros sean estimables. El autor compara el ajuste y respuesta del modelo Poisson tradicional con el MLGM condicional espacial, el modelo Marginal Espacial MLG con varianza geoestadística y el modelo Marginal Espacial MLG con varianza CAR usando datos del estudio: “porcentaje de niños menores de seis años de edad con elevado nivel de plomo en la sangre para cada condado de Virginia en el 2000”,

concluyendo que el Modelo marginal con varianza CAR suaviza mejor los datos que los modelos tradicionales pero sin atreverse a generalizar esta conclusión.

De otro lado, Diggle y Ribeiro [7], extienden la idea de los MLG y MLGM a estudios geoestadísticos, siempre, bajo el supuesto de una distribución subyacente Guassiana y cuya respuesta observada está distribuida como miembro de la familia exponencial. Este enfoque permitió dar nuevas orientaciones a estudios de tipo geoestadístico a partir de estimaciones por máxima verosimilitud. Por otro lado, a partir de una distribución subyacente gausiana estacionaria se pueden extender propiedades inferenciales. En el estudio de conteos, Diggle y Ribeiro [7], propone el modelo Poisson Log-lineal cuya función de linqueo corresponde al logaritmo y la distribución condicional de cada $Y_i$ se distribuye en forma Poisson.

En este trabajo, se busca avanzar en la aplicación de los MLGM bajo un soporte discreto (estudios de datos de área) y respuesta Poisson, por medio de: en primer lugar, la modelación por máxima verosimilitud de los parámetros de los modelos tradicionales SAR y CAR; en segundo lugar, máxima verosimilitud penalizada para estimar los parámetros de un MLGM y en tercer lugar, Cadenas de Markov de Monte Carlo (MCMC) para estimar los parámetros de un modelo Log-lineal Poisson

En el Capítulo 2, se definen algunas nociones importantes de la teoría de los modelos SAR, CAR y MLGM que se van a tratar durante el desarrollo de este trabajo.

Las simulaciones y comparaciones entre los modelos SAR, CAR y MLGM serán desarrolladas en el Capítulo 3.

En el Capitulo 4, se modelan los 200 apellidos más frecuentes del departamento de Antioquia por medio de los modelos expuestos en el Capitulo 2 y se realiza una aplicación con el apellido más frecuente.

Finalmente, en el Capítulo 5 se presentan las conclusiones y recomendaciones del estudio.

# Capítulo 2

# Modelos espaciales

La estadística espacial es la reunión de un conjunto de metodologías apropiadas para el análisis de datos que corresponden a la medición de variables aleatorias en diversos sitios (puntos del espacio o agregaciones espaciales) de una región. De manera más formal se puede decir que la estadística espacial trata con el análisis de realizaciones de un proceso estocástico:

$$\left\{Z(w,s) : w \in \Omega, s \in D \subset \mathbb{R}^P\right\} \tag{2.1}$$

en el que $s$ representa una ubicación en el espacio euclidiano P-dimensional, $Z(s)$ es una variable aleatoria que representa la distribución de todas las posibles realizaciones $z(s)$ en la localización $s$.

La estadística espacial se subdivide en tres grandes áreas. La pertinencia de cada una de ellas está asociada a las características del conjunto $D$ de índices del proceso estocástico de interés. A continuación se mencionan dichas áreas y se describen las propiedades de $D$ en cada una de éstas.

**Geoestadística:** estudia datos de procesos estocásticos en el que el conjunto $D \subset \mathbb{R}^P$ es continuo. Por ejemplo:

$\left\{Z(s) : s \in D \subset \mathbb{R}^2\right\}$, donde $Z(s)$ mide el contenido de nitrógeno en sitios de una parcela experimental. En este caso los sitios pertenecen a $D \subset \mathbb{R}^2$.

$\left\{Z(s) : s \in D \subset \mathbb{R}^3\right\}$, donde $Z(s)$ corresponde al nivel de oxígeno en una represa; donde, las coordenadas geográficas y la profundida corresponden a las variables explicativas.

En los ejemplos anteriores, las coordenadas en donde se miden las variables son continuas.

**Datos de área o Lattices:** en este caso, el conjunto del proceso estocástico $D \subset \mathbb{R}^P$ es discreto y la selección de los sitios de medición depende del investigador (D fijo). Las ubicaciones de muestreo pueden estar regular o irregularmente espaciadas. Algunos ejemplos de datos de Lattices son:

$\{Z(s) : s \in D \subset \mathbb{R}^2\}$, donde $Z(s)$ es la variable aleatoria correspondiente a la tasa de mortalidad y los sitios son los municipios de Antioquia, es decir $D$ es el conjunto discreto formado por los municipios de Antioquia.

$\{Z(s) : s \in D \subset \mathbb{R}^2\}$, donde $Z(s)$ corresponde a la producción cafetera (en Kilogramos) y $D$ es el conjunto de todas las fincas productoras de café en el país.

El estudio de Lattices o datos de área no tienen como objetivo principal la interpolación espacial pero si el modelamiento de los datos. Las principales aplicaciones se encuentran en el campo epidemiológico. El interés de este estudio es hacer énfasis en datos de área (Lattices), siendo $D$ fijo y discreto y por lo tanto contable.

**Patrones puntuales:** la diferencia central del análisis de patrones puntuales con las técnicas geoestadísticas y el análisis de datos de área radica en el hecho de que el conjunto $D \subset \mathbb{R}^P$es aleatorio, es decir que la decisión al respecto de donde se hace la medición no depende del investigador. El proposito de análisis en estos casos es el de determinar si la distribución de los individuos dentro de la región es aleatoria, agregada o uniforme.

Estas definiciones fueron tomadas de Giraldo [9].

El modelo estadístico espacial descompone la variabilidad de $Z(s)$ en varias fuentes: una estructura determinística y uno o mas procesos aleatorios espaciales. Cressie[6] define el modelo:

$$Z(s) = \mu(s) + e(s) \tag{2.2}$$

$$e(s) = W(s) + \eta(s) + \varepsilon(s) \tag{2.3}$$

La variación de $Z(s)$ es expresada a través de la media determinística $\mu(s)$, la media puede depender de la localización espacial y de otras variables, $W(s)$ es llamada la variación suave a pequeña escala, se trata de un proceso estacionario con semivariograma $\gamma_w(u)$ cuyo rango

es mas pequeño que la mas grande distancias entre puntos. $\eta(s)$ es un proceso espacial con variograma $\gamma_\eta(u)$ que mide la variación a microescala, cuyo rango es mas pequeño que la mas pequeña distancia entre puntos, el variograma de $\eta(s)$ no puede ser modelado ya que los datos están disponibles a distancias menores al rango. La presencia de microescala es reflejada en el variograma de $Z(s)-\mu(s)$ como un efecto pepita. $\varepsilon(s)$, finalmente, es un proceso de ruido blanco que representa una medida de error.

Se asume que $e(s) \sim N(0,\sigma^2 I)$, cuando se necesite realizar inferencias de los parámetros y obtener un intervalo de confianza y realizar pruebas de hipótesis. Este supuesto no siempre se cumple, en la Sección (2.3) se considerá la formulación mas general a partir de los modelos lineales generalizados que permite relajar el supuesto distribucional y la linealidad de la función media.

A continuación se hará una presentación de los modelos espaciales que se utilizarán en el desarrollo de este trabajo.

## 2.1. Modelos Simultaneos Autoregresivos (SAR)

La idea central de los modelos SAR se centra a partir de una autoregresión espacial del vector de errores residuales, $e(s)$, en un modelo de regresión lineal con respuesta gausiana. Que es, el regresor $e(s)$ sobre todos los términos de error encontrados de la siguiente forma:

$$Z(s) = X(s)\beta + e(s) \tag{2.4}$$

$$e(s) = Be(s) + v \tag{2.5}$$

donde $B$ es una matriz simétrica de orden $n$ que contiene los parámetros de dependencia $b_{ij}$ entre las áreas $i$ y $j$ con $b_{ii} = 0$ (los valores $b_{ii}$ se asignan como cero asegurando que el valor de cada área no regrese sobre si misma).

Se asume que los errores residuales a partir de la autoregresión, $v_i, i = 1,\ldots,n$, tienen media cero y matriz de varianzas y covarianzas diagonal $\Sigma_v = diag[\sigma_1^2,\ldots,\sigma_n^2]$ (frecuentemente se considerada el mismo valor de $\sigma_v^2$). Si todos los valores de $b_{ij}$ son cero, esto ya no es una autoregresión y el modelo se reduce al tradicional modelo de regresión lineal con errores incorrelacionados. También se puede expresar el modelo autoregresivo como:

$$(I-B)(Z(s)-X(s)\beta) = v \tag{2.6}$$

donde **B** es una matriz que contiene los parámetros de dependencia $b_{ij}$ e $I$ corresponde a la matriz identidad con las dimensiones requeridas. Bajo el modelo expresado en la ecuación (2.4), $Z(s)$ se distribuye de acuerdo a una normal multivariada con media:

$$E[Z(s)] = X\beta; \tag{2.7}$$

y matriz de varianzas y covarianzas

$$\Sigma_{SAR} = V[Z(s)] = (I-B)^{-1}\Sigma_v(I-B')^{-1}. \tag{2.8}$$

Bajo el supuesto de que $(I-B)^{-1}$exista. La estructura de covarianza está determinada indirectamente por $B$ y la elección de $\Sigma_v$. El modelo en (2.8) fue introducido por Whittle [20], y a menudo aparece en la literatura como modelo simultaneo autoregresivo (SAR), donde el adjetivo "simultaneo" describe las n autoregresiones que ocurren simultameamente en cada localización de los datos.

Reparametrizando el modelo a partir de $B = \rho W$, donde $W$ es una matriz de proximidad espacial (explicada con mas detalle en la Sección 2.1.1), $\rho$ corresponde al parámetro de autocorrelación espacial; y asumiendo $\Sigma_v$ sobre un único parámetro $\sigma_s^2$ tal que $\Sigma_v = \sigma_s^2 I$, la ecuación (2.8) se convierte en la siguiente expresión:

$$\Sigma_{SAR} = V[Z(s)] = \sigma_s^2(I-\rho W)^{-1}(I-\rho W')^{-1} \tag{2.9}$$

Para que esté bien definido el modelo, se requiere que $(I-\rho W)^{-1}$ sea no singular (invertible). La rázon principal del uso de este modelo es la dificultad real de ajustar modelos con más parámetros. La forma usual de realizar la estimación e inferencia de los parámetros es por mínimos cuadrados generalizados y por máxima verosimilitud. Dado que que al realizar las estimación de $\rho$ por mínimos cuadrados ordinarios se encuentran problemas de inconsistencia (Whitte[20] y Ord[16]) es preferible trabajar por medio de máxima verosimilitud.

### 2.1.1. Matriz de pesos

Tanto los modelos SAR expuestos en esta Sección y los modelos CAR que se exponen en la Sección 2.2, requieren de la elaboración de la matriz de vecidad (a menudo se asume que esta

matriz es simétrica). Existen numerosas técnicas que permiten su elaboración y no existe una técnica que proporcione los vecinos óptimos; esta matriz habitualmente se le conoce como matriz de pesos, ponderaciones, retardos o contactos espaciales. El uso de la matriz de pesos espaciales permite al investigador la elección del conjunto de ponderaciones que él considere apropiado para cada fenómeno, lo que supone una mayor flexibilidad en la definición de la estructura de interdependencias de un sistema regional y permite considerar cuestiones como las barreras naturales o el tamaño de las regiones. Es más, cuando sea necesaria la consideración de hipótesis acerca del grado de vinculación existente entre áreas vecinas, deben utilizarse distintos conjuntos de ponderaciones que permitan contrastar dichas hipótesis.

En el presente estudio se tendrá en cuenta la matriz de vecindad $W_p$ que cuenta con las siguientes caracteristicas: la matriz $W_p$ recoge el efecto de la región $i$ sobre la región $j$ por medio de pesos o ponderaciones $w_{ij}$. El valor de $w_{ij} = 0$, indica la ausencia de autocorrelación espacial entre las observaciones $i$ y $j$; los elementos de la diagonal principal de la matriz de pesos serán iguales a cero; valores de $w_{ij} \neq 0$ es indicativo de existencia de una interacción espacial entre las observaciones $i$ y $j$ que podría ser expresada como simple contigüidad binaria (teniendo una frontera común) tal y como se muestra a continuación:

$$w_{ij} = \begin{cases} 1 & \text{si sitio } i \text{ y } j \text{ son vecinos} \\ 0 & \text{si sitio } i \text{ y } j \text{ no son vecinos} \end{cases} \tag{2.10}$$

Una completa recopilación de las diversas matrices de pesos se pueden encontrar en Chasco [4].

## 2.2. Modelos condicionales autoregresivos (CAR)

Los modelos condicionales autoregresivos se derivan de los modelos condicionales autoregresivos de series de tiempo y la especificación de modelos para el conjunto de distribuciones de probabilidad condicional de cada observación, $Z(s_i)$, dado los valores observados de todas las otras observaciones. Generalizando, se tiene el modelo $f(Z(s_i) \mid Z(s_{-i}))$, donde $Z(s_{-i})$ denota el vector de todas las observaciones de todos poligónos, excepto $Z(s_i)$. De forma similar para cada observación.

En series de tiempo, una secuencia de variables aleatorias $Y_1, Y_2 \ldots Y_T$ tiene la propiedad de Markov si la distribución condicional de $Y_{t+1}$ dado $Y_1, Y_2 \ldots Y_t$ es la misma que la distribución condicional de $Y_{t+1}$ dado $Y_t$, que es, la predicción mas allá del tiempo de $t$, solo requiere de la observación mas reciente.

Extendiendo esta idea al dominio espacial, asumiendo que $Z(s_i)$ depende solo del conjunto de vecinos, i.e, $Z(s_i)$ depende de $Z(s_j)$ solo si la localización de $s_j$ está en el conjunto de vecinos, $N_i$, de $s_i$; es decir, la observación $Z(s_i)$, únicamente depende de los vecinos de orden uno. En este caso, el proceso $Z(s)$ es llamado un campo aleatorio de Markov. Bajo el modelo condicional autoregresivo, se construye el modelo espacial autoregresivo (CAR) $f(Z(s_i) \mid Z(s_j), s_j \in N_i)$. Si cada una de estas distribuciones condicionales es gausiana, entonces estas distribuciones se puede modelar así:

$$E(Z(s_i) \mid Z(s)_{-j})) = x(s_i)'\beta + \sum_{i=1} c_{ij}(Z(s_i) - x(s_i)'\beta) \tag{2.11}$$

$$Var(Z(s_i) \mid Z(s_{-i})) = \sigma_i^2, i = 1, \ldots, n \tag{2.12}$$

donde las $c_{ij}$ denotan los parámetros de dependencia espacial similares a $b_{ij}$, con $c_{ii} = 0$ y diferentes de cero solo si $s_j \in N_j$.

Si se asume que las distribuciones condicionales son gausianas, con media condicional y varianzas dadas por la ecuación (2.11) y (2.12) respectivamente, las condiciones requeridas por el teorema de Hammersley Cliffort no son demasiado restrictivas y Besag[1] muestra que esta distribución condicional permite construir una distribución conjunta multivariada valida con media y varianza:

$$E[Z(s)] = X(s)\beta \tag{2.13}$$

$$\Sigma_{CAR} = V[Z(s)] = (I - C)^{-1}\Sigma_c \tag{2.14}$$

donde $\Sigma_c = diag[\sigma_1^2, \ldots, \sigma_n^2]$, asegurando que la matriz de varianzas y covarianzas sea simétrica. El poder construir una distribución conjunta, permite la realización de inferencias y el modelamiento de los datos espacial, este proceso se facilita bajo normalidad, caso que no ocurre en otras distribuciones.

Los modelos CAR de primer orden (Homocedásticos) tienen la siguiente media marginal y varianza:

$$E[Z(s)] = X(s)\beta \tag{2.15}$$

$$V[Z(s)] = \sigma_c^2(I - \rho W)^{-1} \tag{2.16}$$

en donde $Z(s)$ es un vector de variables aleatorias medidas en los sitios $s$,$\beta$ es un vector de parámetros de dimensión $k$ a estimar, $X(s)$ es una matriz de tamaño $nxk$, $\rho$ mide la autocorrelación espacial y $W$ corresponde a la matriz de vecindades de orden $n$. La estructura y

grado de autocorrelación es determinada conjuntamente a partir de la matriz $(I - \rho W)$; esta debe ser invertible.

Los modelos SAR solo son definidos para datos Guasianos multivariados y extendidos a los modelos CAR con algunos inconvenientes. Cressie [6] discute las condiciones necesarias para la construcción de la distribución conjunta de probabilidad a partir de una especificación condicional. El punto central de estas condiciones es el teorema de Hammersley Clifford, el cual garantiza que el modelo condicional es coherente con la distribución conjunta de probabilidad (Besag [1]).

## 2.3. Modelo Lineal Generalizado Mixto en el estudio de datos de área

El modelo lineal generalizado clásico introducido por Nelder y Wedderburn [15], proporciona una herramienta unificada para el análisis de diferentes tipos de datos. En diferentes momentos se han extendido los modelos lineales generalizados clásicos a datos dependientes que tienen el propósito de describir la dependecia espacial. Los usados en forma mas extensa son los modelos marginales (Liang y Zeger [14]) y los modelos mixtos (Breslow y Clayton [2]). Estos modelos son usados para modelar datos longitudinales no-gausianos y extendidos al contexto espacial.

El modelamiento espacial en datos de área, inicialmente fue realizado mediante los modelos SAR y CAR; Shamberberger, expone que se pueden crear otras representaciones paramétricas diferentes a los modelos SAR y CAR siempre y cuando las matrices resultantes sean definidas positivas y los parámetros sean estimables.

### 2.3.1. Modelo marginal con Varianza CAR

Shamberberger [19], bajo el criterio de poder crear cualquier representación paramétrica siempre y cuando las matrices resultantes sean definidas positivas y los parámetros sean estimables, adopta un enfoque de modelamiento que permite eludir limitaciones como el supuesto de normalidad en la variable respuesta y en las estimaciones de los parámetros. El autor, parte de un modelo MLG marginal:

$$E[Z(s)] = \mu(s) \tag{2.17}$$

donde, $Z(s)$ es una variable aleatoria que representa la distribución de todas las posibles realizaciones en la localización $s$. La función de enlace es:

$$g(\mu(s)) = log\{\mu(s)\} \tag{2.18}$$

La expresión (2.17) corresponde al promedio de un proceso gausiano dado a partir de la expresión:

$$log\{\mu(s)\} = log\{n_i\} + \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{2.19}$$

donde: $n_i$ corresponde al número de unidades por región; $\beta_0$, $\beta_1$ y $\beta_2$ son los parámetros del modelo, $x_1$ y $x_2$ corresponde a los ejes de coordenadas este-oeste y norte-sur respectivamente. La matriz de varianza y covarianza de este proceso gausiano multivariado se define a partir de la siguiente expresión:

$$VAR[Z(s)] = \sigma_0^2 V_\mu + \sigma_1^2 V_\mu^{1/2} R(\theta) V_\mu^{1/2} \tag{2.20}$$

$$R(\theta) = (I - \rho W)^{-1} \tag{2.21}$$

en donde $\sigma_0^2$ y $\sigma_1^2$ miden la sobredispersión de los datos, $V_\mu$ corresponde a una matriz diagonal de orden $n$ con los términos de la varianza $V(\mu(s_i))$ sobre la diagonal, $V_\mu^{1/2}$ es una matriz diagonal de orden $n$ con los términos de la raiz cuadrada de la varianza $\sqrt{V(u_i)}$. $R(\theta)$ denota la matriz de correlación entre las observaciones parámetrizado por el vector $\theta$, compuesto por: matriz identidad $I$, coeficiente de correlación espacial $\rho$ y matriz de pesos de pesos $W$ generada a partir de la ecuación (2.13).

Por último, se asume que los datos son dependientes condicionados sobre un proceso espacial subyacente $\{S(s) : s \in D\}$; $S(s)$, es un campo aleatorio gausiano con media cero y función de covarianza $\sigma_S^2 \rho S(s_i - s_j)$; $S$ también puede ser interpretado como un efecto aleatorio sobre la localización $s$. Dado $S(s)$, $Z(s)$ se distribuye como miembro de la familia exponencia.

#### 2.3.1.1. Estimación de los parámetros de un MLGM por medio de máxima verosimilitud penalizada

Wolfinger y O'Connell[21] proponen un acercamiento denominado pseudo-verosimilitud como una flexible y eficiente forma de estimar los parámetros desconocidos en un modelo lineal generalizado mixto (MLGM). La estimación por Seudo-verosimilitud consiste en asumir $\beta$ conocido y estimar la varianza $\sigma^2$ y los párametro de autocorrelacion espacial $\theta$ usando máxima verosimilitud (ML) o máxima verosimilitud penalizada (REML). En una segunda fase, se asume $\theta$ conocido y se estima los valores de $\beta$ usando ML o mínimos cuadrados generalizados (EGLS), y se itera hasta que haya convergencia. Este acercamiento en valido para MLGM y MLG marginal.

La idea detrás del acercamiento pseudo-verosimil es linealizar el problema con el fín de poder utilizar máxima verosimilitud y máxima verosimilitud penalizada para realizar las estimaciones y las inferencias de los parámetros. Esta metodología usa la expansión de primer orden de Taylor de la función de linqueo. Los seudo-datos son construidos a partir de la siguiente expresión:

$$v_i = g(\hat{\mu}_i) + g'(\hat{\mu}_i)(Z(s_i) - \hat{\mu}_i) \tag{2.22}$$

donde $g'(\hat{\mu}_i)$ es la derivada de primer orden de la función de linqueo con respecto a $\mu$, evaluando en la actual estimación de $\hat{\mu}$. Para aplicar máxima verosimilitud y máxima verosimilitud penalizada, es necesario tener la media y la matriz de varianzas y covarianzas de los pseudo-datos $v$. Condicionando sobre $\beta$ y $S$, asumiendo $VAR[Z(s)/S]$ tienen la forma de la ecuación 2.20, y usando algunas aproximaciones descritas en Wolfinger y O´Connell [21], esto puede ser derivadas de la forma tradicional:

$$E(v \mid \beta, S) = X\beta + S \tag{2.23}$$

$$VAR(v \mid \beta, S) = \Sigma_{\hat{\mu}} \tag{2.24}$$

$S$, es un campo aleatorio gausiano con media cero y función de covarianza $\sigma_S^2 \rho S(s_i - s_j)$. Con

$$\Sigma_{\hat{\mu}} = \sigma^2 \hat{\Psi}^{-1} V_{\hat{\mu}}^{1/2} R(\theta) V_{\hat{\mu}}^{1/2} \hat{\Psi}^{-1} = \hat{\Psi}^{-1} \Sigma(\hat{\mu}, \theta) \hat{\Psi}^{-1} \tag{2.25}$$

donde la matriz $\hat{\Psi}$es una matriz diagonal de orden $n$ con elementos en la diagonal $\left[\frac{\partial \mu(s_i)}{\partial \eta(s_i)}\right]$ y es evaluada en $\hat{\mu}$.Los momentos marginal de los pseudo-datos son:

$$E(v) = X\beta \tag{2.26}$$

$$VAR(v) = \Sigma_s + \Sigma_{\hat{\mu}} \equiv \Sigma_v \tag{2.27}$$

y $\Sigma_s$ tiene el $(i, j)$ elemento igual a $\sigma_s^2 \rho_s(s_i - s_j)$. Esto puede ser considerado como un modelo de regresión lineal con errores autocorrelacionados espacialmente con la media (de los pseudo-datos $v$) es lineal en $\beta$, por lo tanto, si estamos asumiendo que $\hat{\mu}$ es conocida (o por lo menos no dependa de $\beta$) cuando se desee estimar $\beta$, y que $\beta$ sea conocido cuando se quiera estimar $\theta$, se puede maximizar el log-verosimilitud analíticamente con buen rendimiento por medio del las ecuaciones de mínimos cuadrados:

$$\hat{\beta} = (X'\Sigma_v^{-1}X)^{-1}X'\Sigma_v^{-1}Xv \tag{2.28}$$

$$\hat{\sigma^2} = \frac{(v - X\hat{\beta})'(\Sigma_v^*)^{-1}(v - X\hat{\beta})}{n} \tag{2.29}$$

$$\hat{S} = (\Sigma_S\Sigma_v^{-1})(v - X\beta) \tag{2.30}$$

La matriz $\Sigma_v^*$ en 2.24 es obtenida por factorización residual desde $\Sigma_\mu$y $\Sigma_s$. Sin embargo, ya que $\Sigma_{\hat{\mu}}$ depende de $\beta$ se itera como sigue:

- Obtener una estimación inicial de $\hat{\mu}$ desde los datos originales. Una estimación de un modelo lineal generalizado no espacial a menudo funciona bien.
- Calcular los pseudo-datos a partir de la ecuación 2.22.
- Usar ML con los psudo-datos para obtener las estimaciones de los parametros de autocorrelación espacial, $\theta$, y $\sigma_s^2$ en $\Sigma_v$.
- Usar esta estimación para calcular la estimación por mínimos cuadrados generalizados (Que son también la estimación por máxima pseudo-verosimilitud) de $\beta$ y $\sigma^2$ a partir de las ecuaciones 2.23 y 2.24 y la predicción $S$ se hace a partir de la ecuación 2.25.
- Actualizar la estimación de $\mu$, usando $\hat{\mu} = g^{-1}(X\hat{\beta} - \hat{S})$.
- Finaliza, repitiendo estos pasos hasta que converja.

La estimación por pseudo-verosimilitud se puede trabajar por medio del sofware estadistico R mendiante la función glmmPQL desarrollado por Pinheiro [17]; esta función depende de

las librerías: *lme* y *nlme*. La implemetanción de está metodología esta expuesta en Bivand [3].

## 2.4. Modelo lineal generalizado Geoestadístico

El modelo lineal generalizado Geoestadístico es el mismo modelo lineal generalizado mixto orientado a datos geoestadísticos. Parte de esta clase de modelos es un proceso gausiano estacionario $S(s)$. Un proceso estocástico $S(s)$ es un modelo gausiano si la distribución conjunta de $S(s_1), S(s_2)..., S(s_n)$ es gausiana multivariada para cualquier conjunto de localizaciones $s_i$. El proceso es estacionario si la esperanza de $S(s)$ es la misma para todos los $s$, la varianza de $S(s)$ es la misma para todos los $s$ y la correlación entre $S(s)$ y $S(s^{'})$ depende sólo de la distancia euclídea entre $s$ y $s^{'}$, $\mu = \|s - s^{'}\|$**.** Típicamente, la naturaleza de la superficie $S(s)$, llamada señal, es de interés científico pero no podemos medir directamente esta superficie. En general se supone que $\left\{S(s) : s \in R^2\right\}$ es un proceso gausiano con media $\mu$, varianza $\sigma^2 = Var\{S(s)\}$ y función de correlación $\rho(u) = Corr\{S(s), S(s')\}$

El segundo componente en el modelo lineal Geoestadístico generalizado es la descripción estadística de los datos, generando un mecanismo condicional sobre la señal $S(s)$. Esta parte del modelo sigue un modelo lineal generalizado clásico como lo describe McCullagh y Nelder [14], con $S(s)$ como el equivalente al predictor lineal. Explícitamente, condicionado en $S(.)$, $Z(s)$ son variables aleatorias mutuamente independientes cuyas esperanzas condicionales, $u_i = E[Z(s_i) \mid S(.)]$ son determinadas como:

$$h(\mu_i) = S(s_i) + \sum_{k=1}^{p} \beta_k X_k(s_i) \tag{2.31}$$

donde $h(.)$ es una función conocida denominada función de enlace (link), el valor $X_k(.)$ corresponde a variables explicativas espaciales y $\beta_k$ son parámetros desconocidos de la regresión espacial. Los términos de lado derecho son llamados predictores lineales del modelo. La distribución condicional de cada $Z(s_i)$ dado $S(.)$ es llamada la distribución del error y corresponde a los efectos aleatorios de un modelo mixto.

### 2.4.1. Modelo log-lineal Poisson

El modelo Log-lineal Poisson, es un modelo lineal generalizado cuya función de enlace es el logaritmo y la distribución condicional de cada $Z(s_i)$ es Poisson. La media local de los conteos Poisson es determinada por el valor de un no-observable, proceso estocastico de valor real. En la forma mas simple del modelo, los $Z(s_i)$ son conteos condicionales independientes Poisson, con valor esperado condicional $\mu_i$, dado por:

$$log(\mu_i) = \beta + S(s_i) \tag{2.32}$$

donde, $S(\cdot)$ es un proceso gausiano estacionario con media cero, varianza $\sigma^2$ y función de correlación $\rho(u)$. En el modelo Poisson, a diferencia del modelo lineal gausiano, la varianza condicional de $Z(s_i)$ dado $S(s_i)$ no es un parámetro libre, pero es forzado a ser el mismo en el valor esperado condicional de $Z(s_i)$.

En la práctica, se puede encontrar evidencia adicional de variabilidad en los datos, a menudo llamada una variación extra-Poisson, que no es una estructura espacial. En este caso, una extensión del modelo es incluir un efecto pepita dentro del predictor lineal (entendiéndose por efecto pepita a la discontinuidad que presenta el semivariograma en el origen). La distribución condicional de $Z(s_i)$ es modelada aun como una distribución Poisson con valor esperado condicional $\mu_i$:

$$log(\mu_i) = \beta + S(s_i) + Z_i \tag{2.33}$$

Las $Z_i$ son mutuamente independientes $N(0, \tau^2)$. Esta extensión del modelo permite separar dos componentes de la varianza pepita, la cual fue generalmente indistinguible en los modelos lineales gausinos: la variación Poisson inducida por esquema de muestreo, análoga a la anterior interpretación del efecto pepita como medida de error y un componente espacial incorrelacioando análogo a la interpretación alternativa del efecto pepita a pequeña escala de variación espacial. Lo anterior se expone con mayor detalle en Diggle, P y Ribeiro[7]

### 2.4.2. Correlación Matérn

La correlación espacial entre $S(s)$ y $S(s')$ se medirá a partir de la correlación Matérn. Esta función se muestra a continuación:

$$p(u) = \left\{2^{\kappa-1}\Gamma(\kappa)(u/\phi)^{\kappa}K_{\kappa}(u/\phi)\right\} \tag{2.34}$$

donde, $K_{\kappa}(u/\phi)$ es la función modificada de Besel de orden $\kappa$, con $\kappa > 0$, el cual corresponde al parámetro de forma que determina el análisis de suavidad del proceso subyacente $S(s)$.

$\phi > 0$ es el parámetro de escala, el valor de $\phi$ depende directamente del rango $\alpha$ que será descrito con mas detalle en la subSección (2.1.4.3)

Cuando el valor de $\kappa = 0.5$, la función de correlación Matérn se reduce a la función de autocorrelación exponencial:

$$\rho(u) = exp(-u/\phi) \tag{2.35}$$

donde $u$ mide la distancia entre las observaciones; la función exponencial se caracteriza por representar datos espaciales de menor correlación a cortas distancias. Un valor de $\kappa = 1.5$ la función de correlación Matern tiende a la función de correlación gausiana, cuando el valor de $\kappa \to \infty$, la función de correlación Matérn se convierte en una función gausiana de la siguiente manera:

$$\rho(u) = exp\left\{(-u/\phi)^2\right\} \tag{2.36}$$

donde $u$ mide la distancia entre las observaciones. Esta función se caracteriza por representar correlaciones mas altas que la función exponencial a cortas distancias. En la Figura (2.1) se fijó un rango $\alpha = 200$ km, en esta gráfica se muestran los valores de $\kappa = 0.5$ (función exponencial), $\kappa = 1.5$ (tiende a la función gausiana) y $\kappa \geq 2.5$ (función gausiana); los valores de $\phi$ dependen de la ecuación (2.19). Con los valores de $\kappa$ expuestos se evaluará el alcance de esta metodología cuando el proceso presenta bajas y fuertes correlaciones espaciales. Estas funciones de correlación son mostradas a continuación:

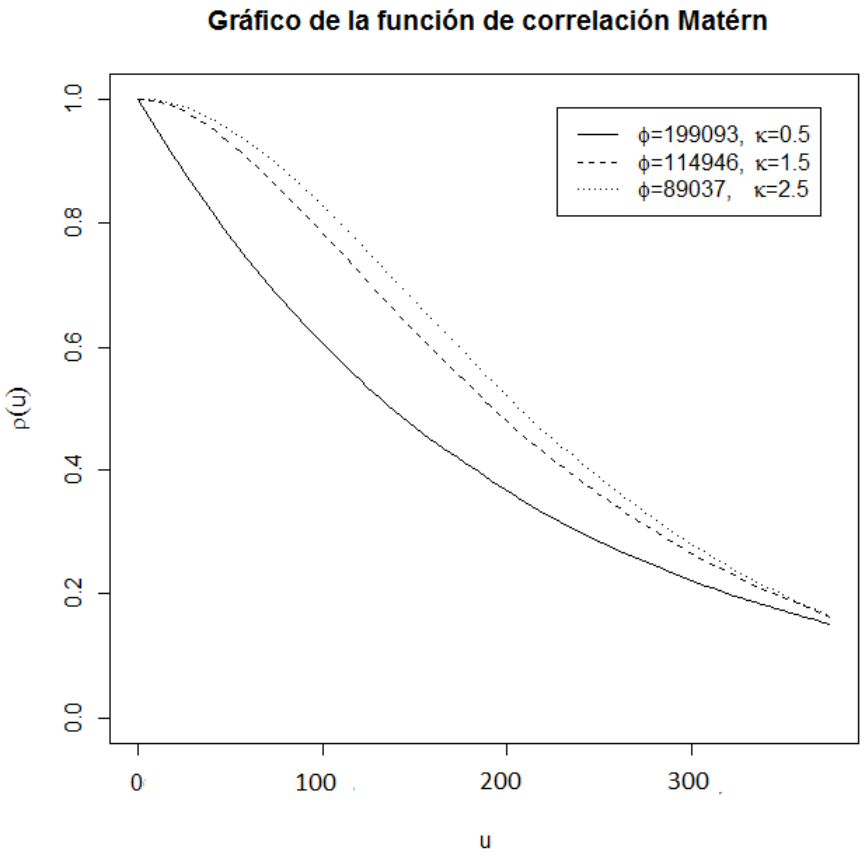


Figure 2.1: Variación de la función de correlación Matern a partir de los valores de $\kappa$ de 0.5 (función exponencial), $\kappa$ de 1.5(tienden a la gausiana) y $\kappa$ de 2.5 (función gausiana) en km.

### 2.4.3. Rango

El valor de rango representa la distancia a partir de la cual dos observaciones son independientes. Handcock y Wallis (1994) sugieren una reparametrizacion de la función de correlación Matern a partir de que $\kappa$ y $\phi$ sean lo mas cercano a un par ortogonal de $\kappa$ y $\alpha$. El rango $\alpha$ se define como:

$$\alpha = 2\phi\sqrt{\kappa} \tag{2.37}$$

Los resultados gráficos, adaptaciones a la simulación y una explicación mas detallada será explicada en la Sección (3.2).

### 2.4.4. Estimación de los parámetros de un modelo Log Poisson

La estimación de los parámetros y predicción es realizada dentro de la metodologías Bayesiana. Para la inferencia Bayesiana, la forma usual de enfrentar estas dificultades consiste en usar los métodos de Monte Carlo, en particular MCMC, para generar muestras de la parte posterior requerida o distribuciones predictivas. Para la explicación de esta metodología se definirá la siguiente notación:

$\theta$ es el conjunto de parámetros que define la estructura de covarianza del modelo de forma similar a la Sección 2.3.1

$\beta$ corresponde a las párametros de regresión

$S(.)$ determina la esperanza condicional de $Y$

$S$ vector de valores de $S(s_i)$ en las localizaciones de $s_i$

$Y$ vector correspondiente de medias de $Y_i$

$S^*$corresponde a la predicción de la localización $s$

Partiendo de que los datos de $Y$ son fijos, un solo ciclo del algoritmo MCMC genera la primera muestra desde $[S \mid \theta, \beta, Y]$, luego desde $[\theta \mid S]$, y finalmente desde $[\beta \mid S, Y]$.

La segunda etapa en el ciclo a su vez puede dividirse en una secuencia de muestras desde la distribución condicional univariada $[S_i \mid S_{-i}, \theta, \beta, Y]$, donde $S_{-i}$ denota el vector $S$ con la eliminación del i-esimo elemento. Alternativamente, el vector $S$ puede ser actualizado en un solo paso. En principio, el algoritmo repite este proceso varias veces a partir de los valores iniciales $\theta$, $\beta$ y $S$, generando finalmente muestras desde $[\theta, \beta, S \mid Y]$ y por lo tanto, simplemente ignorando los valores de la muestra de $S$, desde los requerimientos posteriores de $[\theta, \beta \mid Y]$.

Diggle y Ribeiro [8] crean y utilizan el paquete geoRglm de R para ajustar los datos simulados a un modelo Log Poisson por medio de MCMC.

## 2.5. Aplicación del variograma a datos de área

Zhao[21], investiga qué sucede al aplicar el variograma al utilizarlo en la modelación de correlación espacial y predicción en datos de Lattices. Usualmente, cuando el variograma es utilizado en datos de área, se asigna un punto arbitrario dentro de cada polígono. La preocupación sobre esta asignación arbitraria reside en que la estimación del variograma depende de la distancia entre las observaciones y que la correlación se estaría midiendo a partir de asignaciones arbitrarias dentro de cada poligono. Zhao[21], como experimento,

hace multiples asignaciones aleatorias de los datos dentro de cada polígono y calcula los varigramas empiricos con el objetivo de medir que sucede al hacer variaciones aleatorias dentro de cada polígono. El experimento se desarrolla tanto para polígonos regulares como polígonos irregulares. El estudio determinó entre otras conclusiones, que se encuentra muy poca diferencia en la detección de correlación espacial al tomar el centro del polígono o cualquier otro punto colocado al azar dentro de cada área.

.

# Capítulo 3

# Comparación de los Modelos SAR, CAR y MLGM por medio de simulación

En este Capitulo se presenta una comparación por simulación entre los modelos tradicionales SAR y CAR, y la aplicación de los MLGM expuestos en Schabenberger [19] y " Modelo Log-Poisson" basados en Diggle y Ribeiro [7] en cuanto al ajuste y predicción de los modelos. Las últimas metodologías parten de la aplicación de un MLGM como se expone en el Capitulo 2. Mientras que los modelos SAR y CAR miden la proximidad espacial a partir de una matriz de pesos y la correlación espacial es medida mediante el coeficiente $\rho$; Schabenberger [19] y Diggle y Ribeiro [7] miden la estructura y correlación a partir de una función de correlación espacial: familia Matérn o la familia de potencia. Previamente, Diggle y Ribeiro [7] mostran una aplicación similar bajo el modelo Log Poisson en un campo aleatorio regular y los métodos son únicamente aplicados a datos Geoestadísticos; Shamberberger, propone varias metodologías bajo la condición de que las matrices resultantes sean definidas positivas y mostró una aplicación bajo datos reales concluyendo que el modelo marginal con varianza CAR suaviza mejor las observaciones. Se espera que la metodología propuesta por Schabenberger [19] y Diggle y Ribeiro [7] sean capaces de lograr el mejor ajuste de los datos y en futuros estudios no haya necesidad de construir dicha matriz.

La simulación se soporta sobre un dominio $D$, discreto y por la tanto contable, tal y como se explica en la introducción del Capitulo 2. El dominio $D$, está determinado por los municipios del departamento de Antioquia (polígonos irregulares); en cada municipio se cuenta con el número de personas con el apellido $i$ (valor obtenido por medio las bases del SISBEN); este valor corresponde a una realización proveniente de una distribución Poisson $Y_i$. El campo aleatorio (departamento de Antioquia) está ubicado entre la coordenada mínima (707054.8, 1105895) y la coordenada máxima (992352.3, 1456702), la distancia mínima entre el centroide de dos municipios es de casi 5 km y la distancia máxima entre el centroide de dos

municipios es de 400 km aproximadamente.

Para el desarrollo de la simulación se partió de: 1000 realizaciones $z(s_i)$ ($z(s_i)$ es un valor escalar asociado con la coordenada $s_i$ o a un poligono o municipio) proveniente de un modelo Log Poisson expuesto en la ecuación 2.33, luego, se realizó el ajuste de los parámetros de los modelos tradicionales SAR y CAR por medio de máxima verosimilitud y se ajustaron los parámetros de MLGM por medio de seudo verosimilitud. De forma similar, se generaron 1000 realizaciones de un modelo Marginal con varianza CAR expuesto en la ecuación 2.19 (esta metodología parte de la construcción de una matriz de pesos), con cada simulación, se realizó el ajuste de los parámetros de los modelos tradicionales SAR y CAR por medio de máxima verosimilitud y se ajustaron los parámetros de un modelo Log Poisson por medio de MCMC. Finalmente, se compara el desempeño de las cuatro metodologías por medio del ECM y se concluye bajo que condiciones, estas metodologías son capaces de modelar conteos correlacionados espacialmente.

Para el desarrollo de la simulaciones fueron adaptados los programas de Peter J. Diggle y Paulo J. Ribeiro[7] en el caso del modelo Log-Poisson a reticulas irregulares, y la estimación por máxima verosimilitud penalizada fue realizada mendiante la libreria glmmPQL desarrollado por Pinheiro [17].

## 3.1. Ajuste de datos MLGM por medio los modelos SAR, CAR y Modelo Log Poisson

Inicialmente, se simularon 1000 realizaciones en cada poligono provenientes de un Modelo marginal con varianza CAR expuesto en la Sección 2.3.1. La simulación de los datos fue algo compleja debido a que inicialmente se tiene que construir una matriz de pesos de tamaño acorde al dominio $D$ (en este caso, el tamaño de la matriz es de $125 * 125$). Una vez definida la matriz de vecindad, en la ecuación 2.20, se fijaron los párametros $\sigma_0^2 = \sigma_1^2 = 1$ y el valor de $\rho$ se fijó como 0.8 y la diagonal de la matriz $V$ se fijo en la simulación en forma constante como 4.01 (Valor de $\sigma^2$ obtenido a partir del modelo CAR para el apellido Gómez). La ecuación 2.20, corresponde a la matriz de varianzas y covarianza de una distribución normal multivariada con vector de medias cero. Para la simulación de los datos se debe garantizar la invertibilidad de la expresión $(I - \rho W)$; en segundo lugar, se debe garantizar que la ecuación 2.20 sea definida positiva, situaciones que limitaban la simulación. Mas adelante, se explicarán algunas limitaciones encontradas en el modelo marginal con varianza CAR expuesto en Schanbenberger [21].

Finalmente, cada simulación fue modelada a partir de los modelos SAR, CAR y Modelo Log Poisson expuestos en las secciones: 2.1, 2.2, 2.3 y 2.4 respectivamente con el fin de determinar cuál de estas metodologías lograba ajustar mejor los datos y por ende mostraba un ECM menor. Uno de los objetivos en esta Sección, es evaluar la capacidad que tiene las MCMC para encontrar los parámetros de un modelo Log Poisson expuesto en Diggle, P y Ribeiro[7]. Para lograr este objetivo, se pretendió inicialmente partir de tres funciones de correlación apriorí provenientes de una función de correlación Matérn expuesta en la Sección 2.4.2: función exponencial ($\kappa = 0.5$), gausiana (1.5) y gausiana (2.5).

Al realizar las estimaciones por medio de las MCMC, se pudo determinar que el supuesto asumido en la correlación Mátern no afectaba en gran medida los resultados de las estimaciones. La Tabla 3.1, muestra que al variar el valor de $\kappa$, las estimaciones de $\beta$ y $\sigma^2$ son muy similares.

| Parámetro | $\kappa = 0{,}5$ | $\kappa = 1{,}5$ | $\kappa = 2{,}5$ |
|---|---|---|---|
| $\beta$ | (1.55, 4.34) | (1.92, 4) | (2.08, 3.86) |
| $\sigma^2$ | (46.81. 83.71) | (27.18, 48.52) | (21.12, 37.66) |
| Rango | (279642.3, 281420) | (280032.6, 281676.6) | (280013.4, 281433.3) |
| *ECM* | (8.33, 1169.66) | (9, 746.24) | (9, 225.46) |

Table 3.1: Estimación por intervalo al 95% de confianza de los parámetros de un modelo Log Poisson a partir de datos generados de un modelo marginal con varianza CAR

Al comparar los rangos, se puede detectar que las tres funciones de correlación apriori asignadas en los MCMC, no detectan una correlación espacial más allá de 282 km y no es menor a 280 km. A continuación, se muestra gráficamente la estimación de rango bajo los tres escenarios:

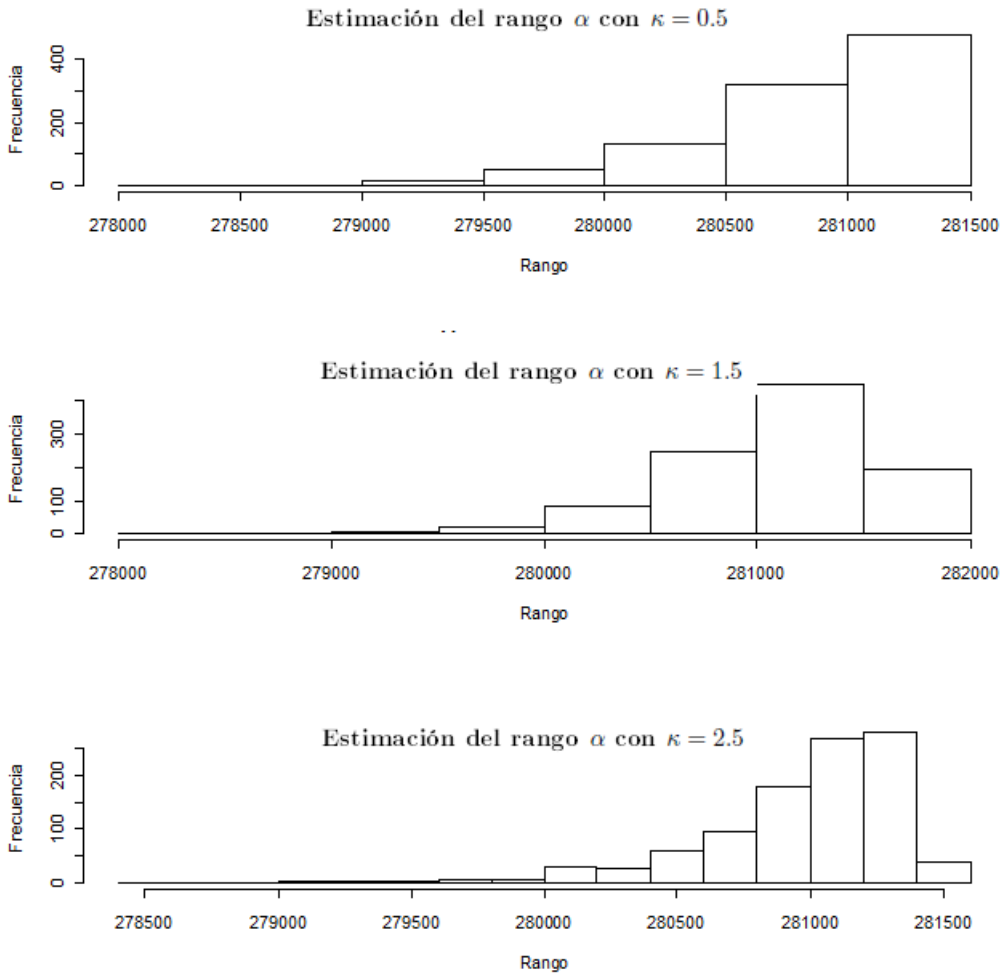


Figura 3.1: Histograma de frecuencia para la estimación del rango

Adicionalmente, los ECM obtenidos al modelar bajo una función de correlación gausiana ($\kappa = 2{,}5$) son mucho más pequeños que los ECM obtenidos bajo las funciones de correlación: exponencial ($\kappa = 0{,}5$) y gausiana ($\kappa = 1{,}5$) tal y como se muestra en la Tabla 3.1. La Figura 3.1, muestra que la estimación del rango es practicamente la misma al partir de valores de $\kappa = 0{,}5$ , $\kappa = 1{,}5$ y $\kappa = 2{,}5$.

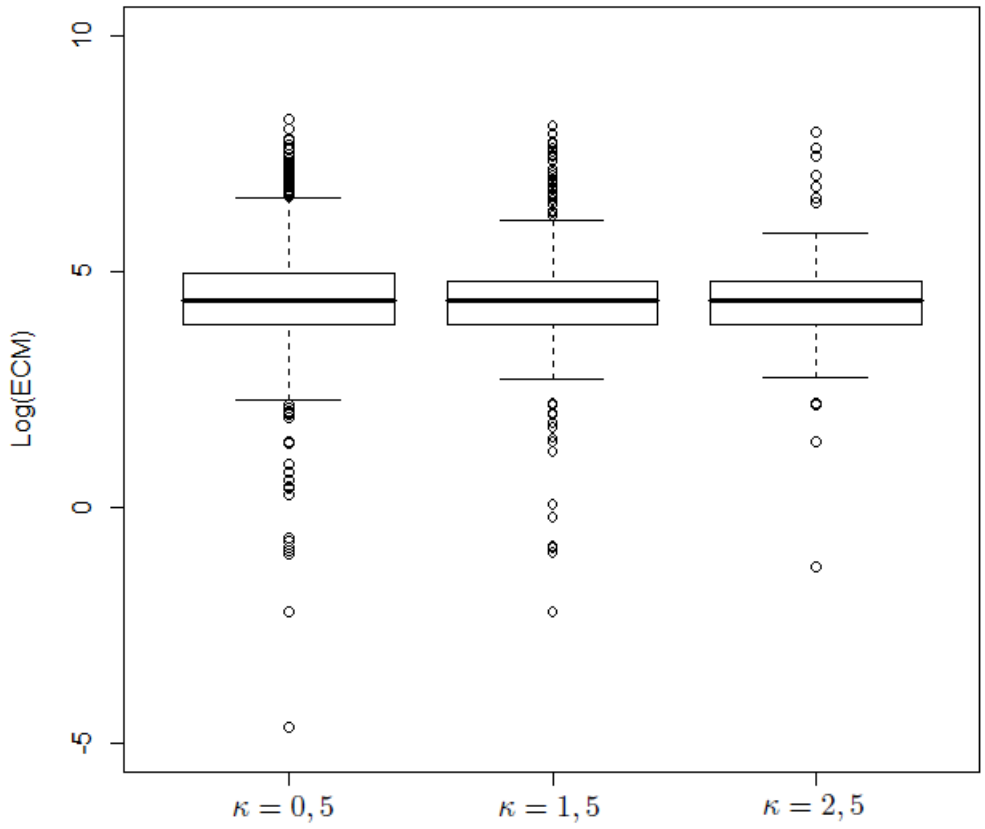


Figure 3.2: Logaritmo del ECM del ajuste de datos de un modelo marginal con varianza CAR a partir de un modelo Log Poisson con $\kappa = 0.5, \kappa = 1.5$ y $\kappa = 2.5$

Este primer acercamiento en el análisis espacial de datos de área a partir del modelo Log Poisson expuesto en Diggle, P y Ribeiro[7], permite dar indicios de que la función de de correlación apriori no influye en forma significativa sobre las estimaciones de los parámetros. Bajo esta conclusión, se decidió estimar los parámetros del modelo Log Poisson unicamente con la función de correlación gausiana.

### 3.1.1. ECM obtenido por medio los modelos SAR, CAR y Modelo Log Poisson

En esta Sección, se comparará el ajuste de los datos generados por medio de un modelo Marginal con varianza CAR a partir de los tradicionales modelos: SAR, CAR y modelo Log Poisson que tiene función de correlacion Mátern con $k = 2.5$ (función gausiana). El objetivo que se quiere desarrollar en esta Sección, es verificar si es posible modelar datos espaciales provenientes de una modelo con matriz de pesos $W$ tal como el que se expone en la Sección 2.3.1 a partir de modelos de estadística espacial tradicionales (SAR y CAR) y a partir de los

modelos lineales generalizados mixtos.

Las ventajas en la modelación de datos espaciales de área a partir de MLGM, se basa en no tener que partir restrictivamente de una variable respuesta distribuida en forma normal si no en poder ampliar el rango de aplicación a variables respuestas generadas a partir de distribuciones de la familia exponencial. Una segunda ventaja en la modelación de datos espaciales de área a partir de los MLGM es el no tener que construir tediosas matrices de vecindad. En esta Sección, se evaluará el ajuste de estos modelos.

Cada una de las 1000 realizaciones generadas de un modelo Marginal con varianza CAR fueron ajustados a partir de los modelos expuestos en párrafos anteriores. Los resultados de los errores cuadráticos medios se muestran a continuación:

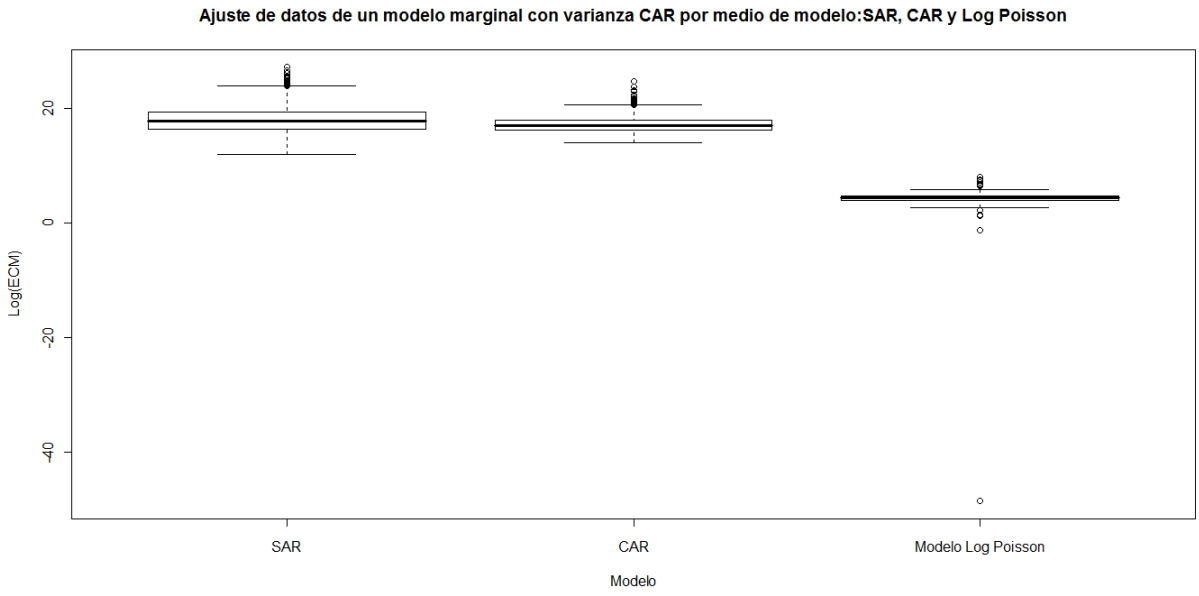


Figure 3.3: Logaritmo del ECM del ajuste de datos de un modelo Marginal con varianza CAR a partir de los modelos: SAR, CAR y Modelo Log Poisson

En la Figura (3.3) se muestra el comportamiento de los ECM obtenidos a partir de los modelos expuestos. Los ECM obtenidos a partir de los modelos tradicionales SAR y CAR, muestran que estos modelos no se adecuaron bien en la modelación de los datos provenientes de un modelo marginal con varaianza CAR debido a su alta magnitud y a su variabilidad; situación contraria, se puede apreciar en la estimación de los parámetros del Modelo Log Poisson hecha a partir de MCMC. A continuación se muestra con mas detalle el comportamientos de los ECM obtenidos a partir del ajuste del modelo Log Poisson.

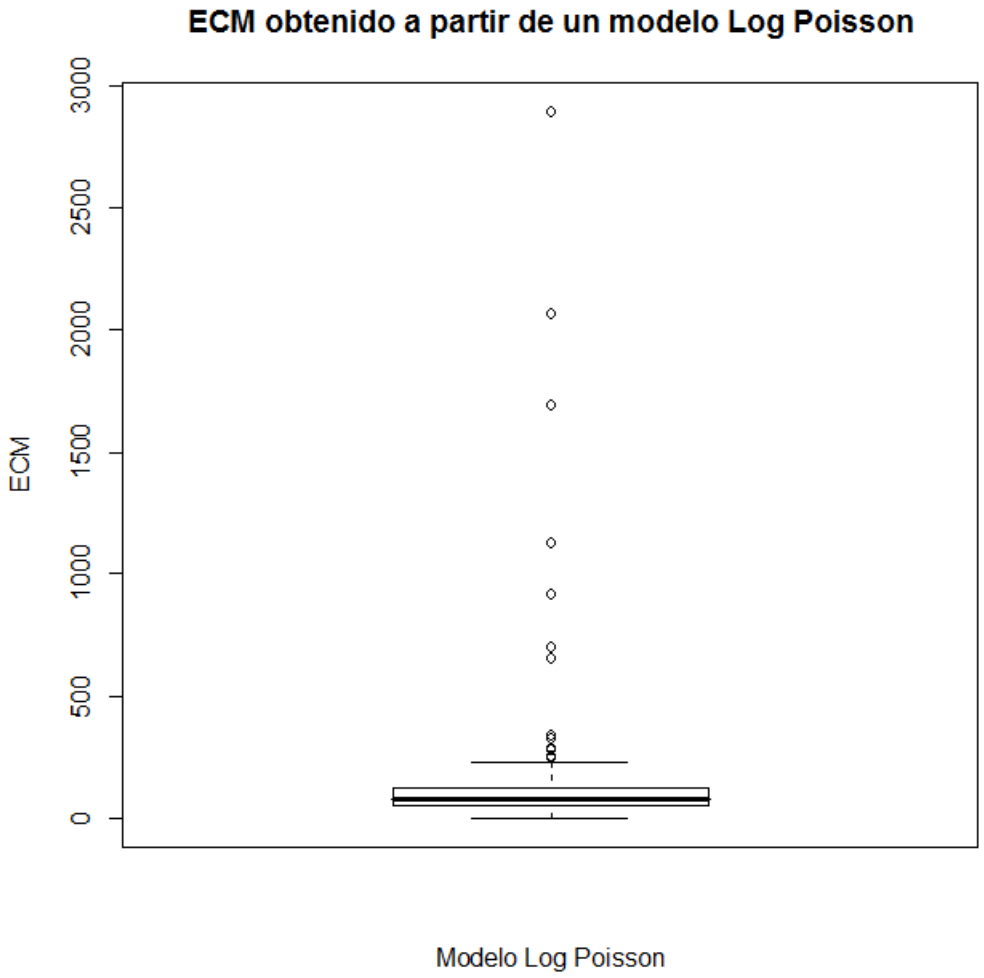


Figure 3.4: Ajuste de datos de un modelo marginal con varianza CAR por medio de un modelo Log Poisson

El 99.6% de los ECM al ajustar las realizaciones al modelo Log Poisson no son superiores a 1000, mientras que todos los ECM obtenidos por medio de los modelos SAR y CAR registraron ECM superiores a 1000. Lo anterior indica que las MCMC logran ajustar los datos a un Modelo Log Poisson con una mayor precisión que los métodos tradicionales. Aunque los modelos expuestos ajustan parámetros diferentes, estos coinciden en la modelación de los datos y la interpolación en lugares en donde no se pudo observar la variable en estudio.

## 3.2. Ajuste de datos Log Poisson por medio los modelos SAR, CAR y MLGM

A diferencia de la Sección 3.1, el modelo Log Poisson, a partir del cual se generan los datos no posee una matriz de pesos espaciales. El modelo lineal generalizado Log Poisson parte de medir la relación espacial a partir de la distancia Euclideana. Esta metodología parte de un modelo mixto con efectos aleatorios $S(s)$ y efectos fijos $\beta$. $S(s)$, corresponde a una distribución gausiana multivariada con función de correlación Matérn tal como se explica en la subSección (2.4.1); $S(s)$ no es observable en el contexto aplicado. En la simulación se parte de que se conoce el proceso $S(s)$ y que sumado con una constante $\beta$ corresponde al logaritmo del promedio de una variable aleatoria $Z(s_i)$ que se distribuye Poisson en cada uno de los 125 polígonos irregulares (cada polígono corresponde a un municipio del departamento de Antioquia).

A partir de la variable aleatoria Poisson cuyo promedio corresponde a $e^{\beta+S(s_i)}$ se determinó una realización en cada uno de los 125 polígonos. Una vez estandarizado el número de habitantes en una escala entre 0 a 1000 habitantes, se calcula el promedio de personas con el apellido Gómez por municipio (este apellido es el mas frecuente en el departamento de Antioquia y por medio del modelo SAR se registró una de las mayores correlaciones espaciales), encontrando un promedio de 36.92 personas por cada 1000 habitantes. Al reemplazar en la ecuación (2.32): $e^{\beta+S(s_i)} = 36.92$ y al aplicar logaritmo en ambos lados de la igualdad se encontró que $\beta + S(s_i) = 3.61$; finalmente se fijo $\beta = 2$ y el valor promedio de $S(s_i) = 1.61$, debido a que si $\beta$ es pequeño, la variación Poisson domina a $S(s_i)$; en la simulación, se asignó un valor de $\beta$ grande de tal manera que $S(s_i)$ domine los valores generados en la simulación, lo anterior permite partir de un modelo con alta correlación. Dentro de la señal se asigna un valor $\sigma^2$ pequeño de tal forma que los valores simulados presenten una baja variabilidad y sean similares a una muestra obtenida a partir de la aplicación. Para evaluar el alcance en el ajuste de los parámetros de un modelo SAR, CAR y MLGM de datos obtenidos a partir del modelo Log Poisson se tomaron 1000 realizaciones obtenidas a partir de este modelo con estructura de correlación Matérn con $\kappa = 2.5$ y $\phi = 281560$. Los resultados de las estimaciones se muestran a continuación:

| Modelo | Intervalo de confianza al 95 % del ECM |
|---|---|
| SAR | (450.54, 8775.67) |
| CAR | (438.99, 8430.68) |
| MLGM | (19.6, 151.89) |

Table 3.2: Estimación por intervalo al 95% del ECM para los modelos SAR, CAR y MLGM

La Tabla 3.2 muestra los ECM obtenidos a partir del máxima verosimilitud penalizada en el caso de los MLGM y por máxima verosimilitud en el caso de los modelo tradicionales SAR y CAR. Los ECM también fueron representados en la Figura 3.5.

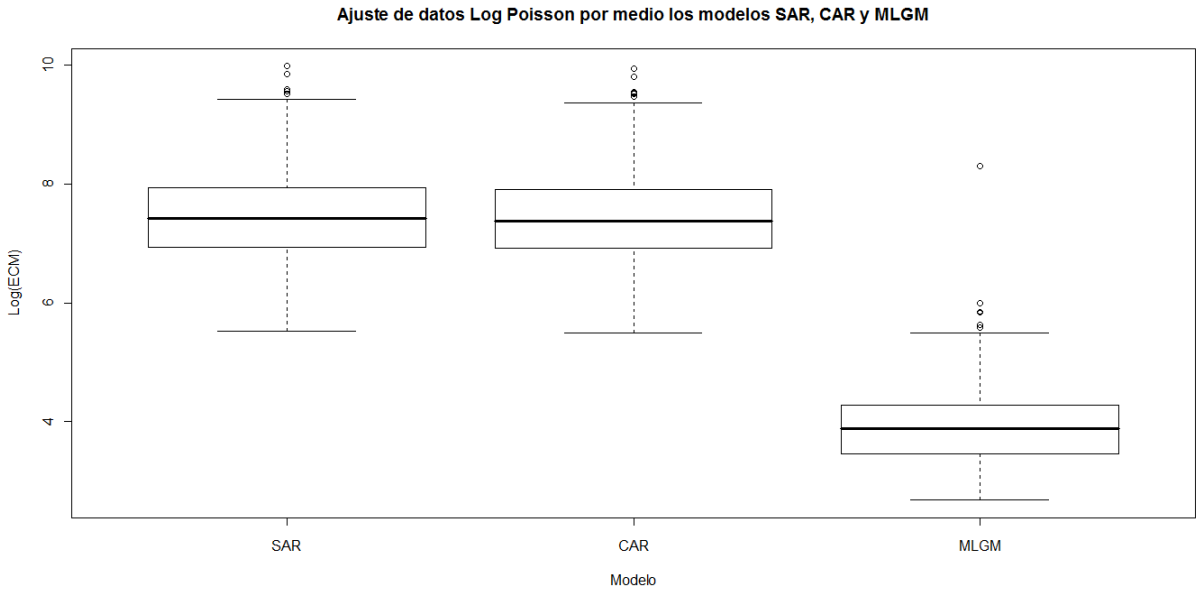


Figure 3.5: Logaritmo del ECM al ajustar datos Log Poisson por medio los modelos SAR, CAR y MLGM

Los intervalos de confianza que se muestran en la Tabla 3.2 y la Figura 3.5 dan fuerte evidencia de que la estimación de los parámetros hecha a partir de máxima verosimilitud penalizada muestra el mejor ajuste en los datos. En el caso de modelar datos espaciales, estos resultados pueden dar evidencia de que máxima verosimilitud penalizada en MLGM logra un mejor ajuste de los datos. A continuación se muestra el ECM obtenido a partir de MLGM a una escala de 0 a 500.

Los 91.3 % de los ECM del MLGM ajustados por máxima verosimilitud penalizada son inferiores a 100. Situación muy inusual en los modelos SAR, CAR e incluso en la modelación hecha por MCMC expuesto en la Sección anterior; en este último modelo, se encontró que el 62.4 % de los ECM fue inferior a 100. Aunque, el 91.3 % y 62.4 % se obtuvieron a partir de realzaciones distintas, estos resultados pueden dar evidencia de que máxima verosimilitud penalizada logra obtener un mejor ajuste en los datos que MCMC. Para corroborar esta información, en el Capitulo 4 se presentará la modelación de 200 realizaciones que no provienen de ningún modelo y que previamente se verificó la presencia de correlación espacial

por los modelos SAR y CAR. Con estos datos, se podrá refutar o no la anterior conclusión y establecer cual de las dos metodologías podría modelar mejor datos espaciales de área.

## 3.3. Dificultades al trabajar con MLGM

En primer lugar, al modelar un MLGM por máxima verosimilitud penalizada se encontraron problemas de convergencia en el algoritmo al partir de una función de correlación exponencial y gausiana. Indagando en el programa se encontró que los valores iniciales dependían de un modelo lineal generalizado y no siempre, este modelo puede generar valores iniciales que faciliten la optimización de los parámetros del modelo. Ante esto, surgen los siguientes interrogantes: ¿Cuáles son los valores iniciales que se deben asignar al algoritmo? ¿Se tienen problemas de mal condicionamiento? ¿Bajo la estimación de qué parámetro se obtiene la solución?

Al indigar en la red, se encontró abierta una discusión similar en la lista de R. En ella, se parte de que al cambiar el orden los datos, las estimaciones cambian gradualmente y que al realizar un segundo cambio en las observaciones se presenta problemas de convergencia similares a los obtenidos en este trabajo. Spenser Graves, afirma que la solución puede ser compleja y demorada. Ver cibergrafía [C1]

Una solución, propuesta a partir de este trabajo, es modelar los datos por medio MCMC enunciado por Diggle, P y Ribeiro[7], siempre y cuando la respuesta se distribuya como miembro de la familia exponencial cuando el algoritmo por máxima verosimilitud penalizada muestre problemas de convergencia debido a la no invertibilidad de las matrices de varianzas y covarianzas. Cabe aclarar, que el modelo Log Poisson se está extendiendo a partir del contexto continuo y que se debe tener cuidado con la interpolación de resultados, sin embargo, Zhao[21], investiga qué sucede al aplicar el variograma al utilizarlo en la modelación de correlación espacial y predicción en datos de Lattices.

En segundo lugar, Schabenberger [19], define la matriz de varianza y covarianzas (2.18) como una matriz diagonal con los términos de la función de varianza en la diagonal; contradictoriamente, en la página de la editorial CRC PRESS (ver cibergrafía [C2]), el autor expone el algoritmo bajo una solución generada como la inversas de la función de varianzas en la diagonal. Bajo este escenario, la solución del algoritmo encuentra valores óptimos rapidamente; caso contrario sucede al trabajar con la inversa de las funciones de varianza en la diagonal y/o cambiar los valores iniciales. Se encontró, que aun modificando los datos del autor el algoritmo presentaba problemas de convergencia.

# Capítulo 4

# Aplicación de las metodologías expuestas en datos reales

Previamente, Gómez y Muñeton[10] han estudiado los procesos distribucionales de la población en el departamento de Antioquia a partir del análisis de frecuencias y distribución de apellidos de sus pobladores, mediante el cual se han establecido relaciones de parentesco y origen; este tipo de investigación es conocida como estudios de Isonimia. Los estudios de Isonimia se han limitando únicamente al estudio de frecuencias y distribución de apellidos de sus pobladores sin tener en cuenta el comportamiento espacial a partir de la georeferenciación de la variable de estudio. Este interrogante permitió crear conjeturas sobre que métodos de la estadística espacial podrían responder estas preguntas y encontrar tendencias que los estudios típicos de Isonimia y estadística clásica no han podido hallar.

En forma adicional, a las simulaciones expuestas, se seleccionaron los 200 apellidos más frecuentes del departamento de Antioquia, el número de personas con el apellido *i* fue georeferenciado dentro de cada municipio y se ajustó cada apellido bajo los modelos: SAR, CAR, modelo Log Poisson y el Modelo Lineal Generalizado mixto. El objetivo de este capítulo, es corroborar los resultados obtenidos en las simulaciones a partir de datos reales que no provengan de ningún modelo. En segundo lugar, se quiere mostrar, como la estadística espacial es una potente herramienta que proporciona información consolidada sobre las dinámicas sociales y culturales de las poblaciones Antioqueña.

Para el desarrollo de esta inquietud, se parte de las bases de datos del SISBEN en donde se discrimina la población sisbenizada por municipio y apellidos; esta información es georeferenciada tomando como coordenada el centroide de cada municipio. Una vez georeferenciada la información, se procederá a modelar los datos bajo las técnicas expuestas en el Capítulo

2. La variable respuesta corresponde al número de personas con el apellido *i* en cada uno de los municipios de Antioquia.

## 4.1. Modelación del número de personas con el apellido *i* por medio de los modelos tradicionales SAR y CAR

En primer lugar, cada realización fue ajustada a partir de los modelos SAR y CAR, tomando la matriz de pesos $W$ definida en la ecuación 2.10. Tanto en los modelos SAR y CAR es posible calcular el índice de correlación espacial. Este indice, permitirá justificar el correcto uso de los modelos espaciales. Estos cálculos son presentados a continuación:

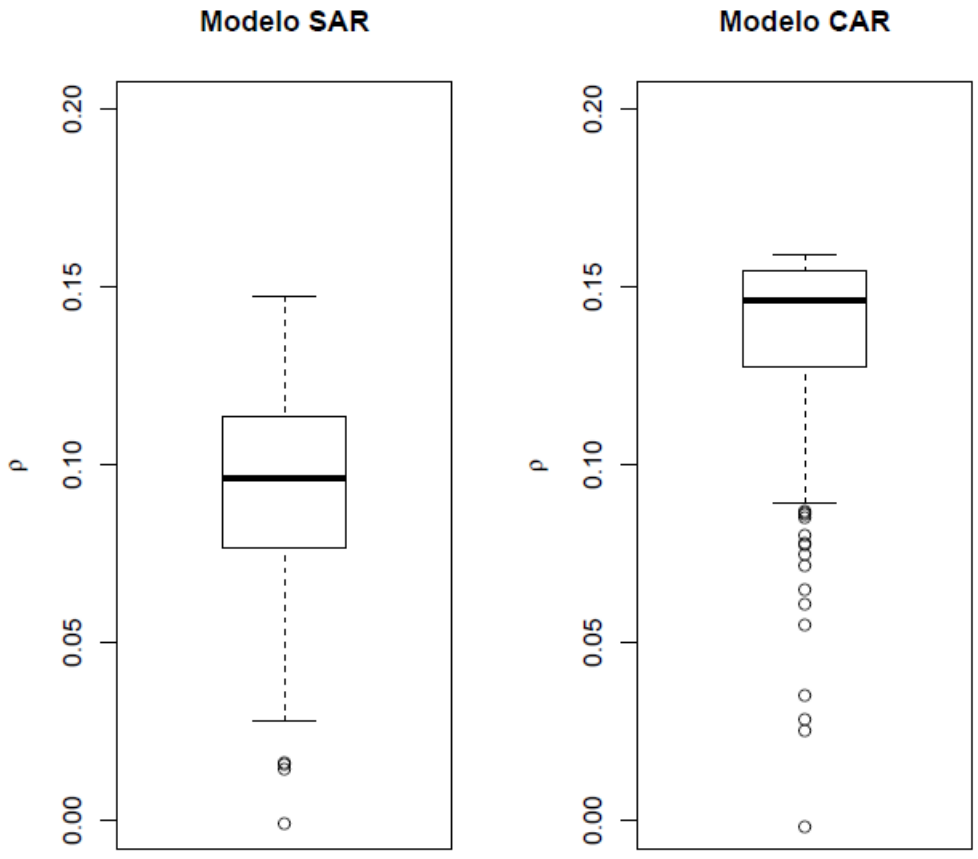


Figure 4.1: Estimación de $\rho$ en los modelos SAR y CAR a partir de los 200 apellidos mas frecuentes del departamento

Aunque, los coeficiente de correlación espacial obtenidos en la Figura 4.1 no son comparables debido a que provienen de diferentes modelos; si es posible apreciar que ambos modelos detectan la presencia de correlación espacial al modelar el número de personas con el apellido *i*; lo anterior da evidencia de que la distribución espacial de los antioqueños no es aleatoria

y que obedece a algún patrón de distribución espacial.

La significancia estadística de estos resultados se puede observar en la Figura 4.2; en este gráfico, se muestra la distribución de los valores p obtenidos para $\rho$. La escala de la Figura 4.2 se muestra como el logaritmo del valor p con el fin de visualizar mejor la información.

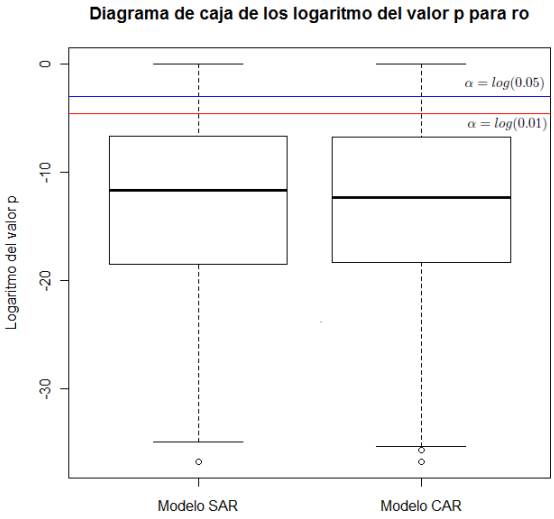


Figure 4.2: Valor p para $\rho$

En el modelo SAR, bajo la hipotesis nula $H_0$:$\rho$=0, se detectó que en el 91.46 % de los apellidos se rechazó $Ho$ baja una significacia del 5 % (valores inferiores a la linea azul gruesa); también se encontró que en el 84.43 % de los apellidos se rechazo $Ho$ a una significacia del 1 % (valores inferiores a la linea roja delgada) tal y como lo muestra la Figura 4.2. Como en la mayoría de los apellidos se rechazó $Ho$, se puede concluir que la distribución poblacional de los antioqueños no se comporta en forma aleatoria y que responde a patrones de distribución espacial.

## 4.2. Comparación de los modelos tradicionales SAR y CAR, El modelo Log-Poisson y MLGM

Para comparar los modelos SAR, CAR, Log Poisson y MLGM se ajustaron las 200 realizaciones (apellidos más frecuentes) a estos modelos y finalmente se cálcularon los ECM. Los resultados aparecen a continuación:

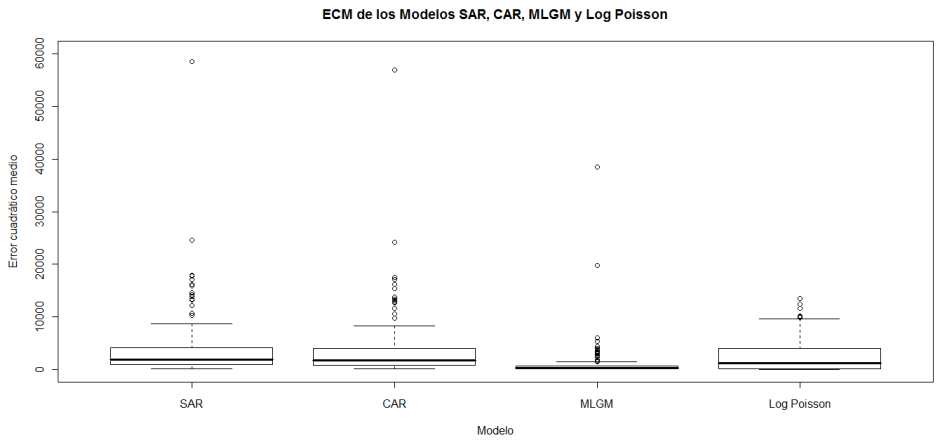


Figura 4.3: ECM en los modelos expuestos al ajustar los 200 apellidos mas freccuentes de Antioquia

En la Figura 4.3 se observa la estimación de los ECM. El ajuste generado por los modelos SAR y CAR muestra la existencia de sobrestimaciones hechas a partir de su modelación. En segundo lugar, los ECM generados en el ajuste de cada apellido al modelo Log Poisson son inferiores a los obtendios por medio de los modelos tradicionales SAR y CAR tal y como lo muestra la Figura 4.3. Por último, se observa que el MLGM generó los ECM mas bajos con la presencia de sobreestimaciones en algunos apellidos a pesar del excelente ajuste que muestra el MLGM, se encontraron apellidos como Cuesta y Palacio que al ser modelados bajo el MLGM mostraron ECM superires a 80.000, la modelación de los demás apellidos presentó errores cuadráticos muy bajos.

## 4.3. Aplicación con respecto al número de personas con el apellido Gómez

En esta Sección se mostrará una aplicación con respecto al número de personas por municipio que se apellidan Gómez. El objetivo del Capitulo es mostrar un análisis mas detallados acerca de las técnicas expuestas en capítulos anteriores.

### 4.3.1. Análisis exploratorio

En la Figura 4,4 se puede apreciar el comportamiento del número de personas con el apellido Gómez en el departamento de Antioquia. La información obtenida en cada municipio fue

llevada a una escala de 0 a 1000 habitantes con el fin de evitar la influencia de municipios sobre poblados, colores que tienden a blanco en el mapa indican una frecuencia baja de personas con este apellido y el colores oscuros indica una alta frecuencia de personas con el apellido Gómez.

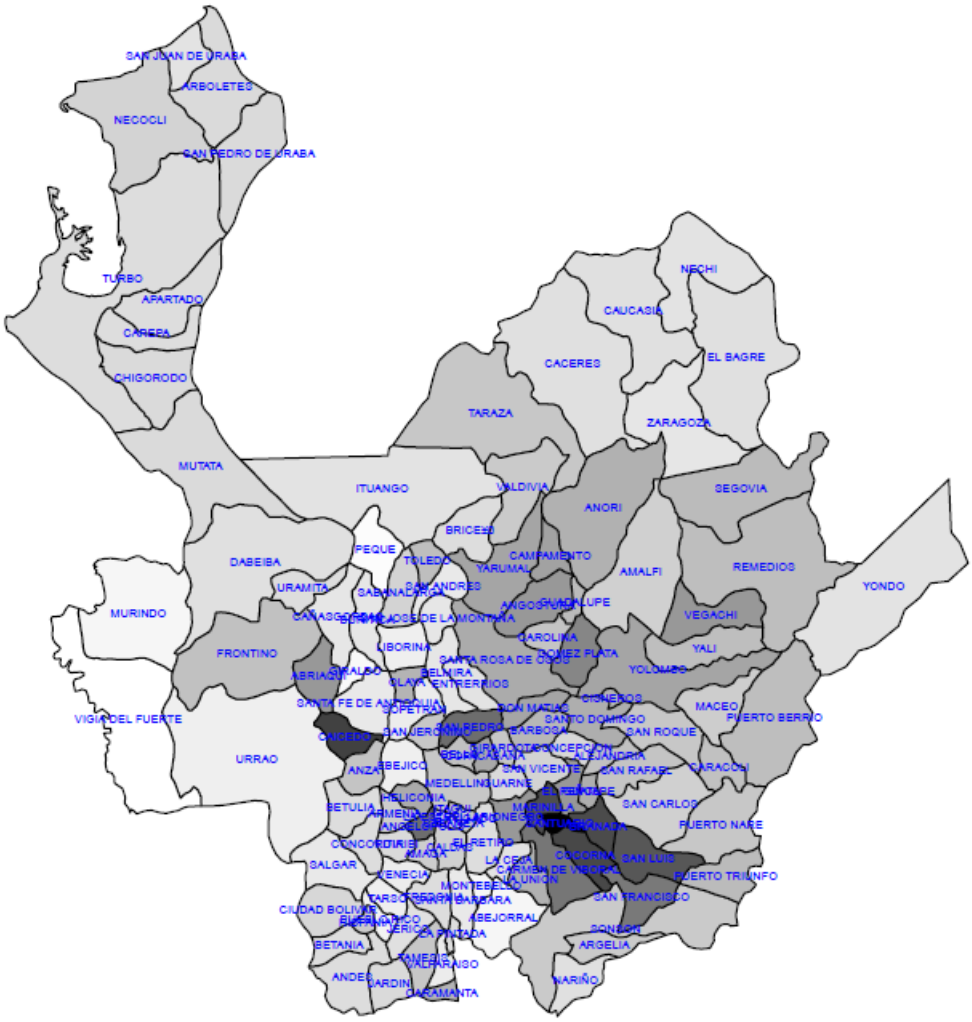


Figure 4.4: Número de personas con el apellido Gómez en el departamento de Antioquia a una escala de 1000 habitantes

En la Figura 4.4 se puede apreciar que las personas apellidadas Gómez se presentan con mayor frecuencia en la subregión de Oriente, Nordeste y Magdalena medio principalmete. En el caso del oriente Antioqueño, el número de personas con el apellido Gómez se presenta con mayor frecuencia en el municipio del Santuario con un total de 82 personas por cada 1000 habitantes; también se destaca que los municipios ubicados en el sur oriente de este municipio toman valores muy similares del Santuario; estos municipios son: Cocorná, Granada y San Luis con 59, 58 y 58 personas por cada 1000 habitantes respectivamente. De forma similar, sucede en municipios ubicados en el nor-occidente de este municipio tales

como: Marinilla y Rionegro con 57 y 33 personas, respectivamente. También se puede observar, que el apellido Gómez ubicado en el suroeste del departamento de Antioquia presenta un total de 61 habitantes por cada 1000. Se destaca, que el número de personas con el apellido Gómez se presenta en forma constante en la región del Urubá antioqueño con un promedio de 10 personas con este apellido por cada 1000 habitantes.

Visualmente, se puede apreciar presencia de autocorrelación espacial en la variable número de personas con el apellido Gómez; para corroborar la presencia de correlación espacial se calculó el coeficiente de correlación espacial al modelar los datos a partir de los tradicionales modelos SAR y CAR obteniendo una correlación de 0.13 y 0.16 respectivamente; además, se probó la hipótesis nula $Ho : \rho = 0$, tanto en el modelo SAR como en el modelo CAR se rechazo $Ho$ con un valor $p < 0.001$. Lo Anterior indica que la ubicación espacial de las personas de apellido Gómez no es aleatoria dentro del departamento de Antioquia y que su distribución obedece a algún tipo de razón económica, social o histórica.

Se aclara, que los coeficiente de correlación obtenidos bajo el modelo SAR y CAR no son comparables y que el hecho de que el coeficiente de correlación del modelo CAR sea mayor es indicio de que la mayor influencia viene dada principalmente por los municipios vecinos y no ha grandes distancias. El modelamiento de los datos a partir de los modelos SAR y CAR se utilizó la matriz de pesos expuesta en la ecuación 2.10.

### 4.3.2. Resultados por medio del MLGM

Una vez determinada la existencia de correlación por medio de los modelos SAR y CAR se modeló el número de personas con apellido Gómez utilizando máxima verosimilitud penalizada con el objetivo de determinar si variable respuesta depende de las coordenadas $x$ (oriente-occidente) y $y$ (norte-sur). Los resultados de la estimación de los efectos aleatorios y fijos aparecen a continuación:

| Efectos aleatorios | Intercepto | Residual |
|---|---|---|
| Desviación estadar | 0.59 | 0.0024 |

Cuadro 4.1: Efectos aleatorios del números de personas con apellido Gómez por municipio

El valor de 0.59 es la desviación estandar que detecta el modelo si se realizara varias mediciones en cada municipio (medidas repetidas); el modelo detecta un error experimental aproximadamente de 0.0024.

El resultado de los efectos fijos son los siguientes:

| Efectos fijos | value | std.Error | DF | t-value | P-Value |
|---|---|---|---|---|---|
| Intercepto | -0.6482389 | 1.3095640 | 122 | -0.495004 | 0.6215 |
| $x_1$ (oriente-occidente) | 0.0000045 | 0.0000009 | 122 | 4.817199 | 0.0000 |
| $x_2$ (Norte-sur) | -0.0000004 | 0.0000007 | 122 | -0.493283 | 0.6227 |

Cuadro 4.2: Efectos fijos del número de personas con apellido Gómez por municipio

Bajo $Ho : \beta_i = 0$**,** con el estadístico de prueba $t = \frac{\hat{\beta}_i}{\frac{s}{\sqrt{n}}}$, el parámetro $\beta_1$ resultó ser significativo, el modelo detecta que los movimientos migratorios para el número de personas con el apellido Gómez se presentan en el sentido oeste-este en forma creciente. Situación que el modelo no puede detectar en sentido norte-sur en donde la distribución de la variable respuesta parece ser independiente. Al modelar sin $x_2$, los resultados obtenidos son similares a los anteriormente. Para la correcta utilización del estadistico t, se verificó previante la existencia de normalidad en los residuales por medio del test de Shapiro Wilk.

### 4.3.3. Modelación usando el Modelo Log Poisson propuesto por Diggle y Ribeiro

Al ajustar los datos, al modelo Log Poisson por medio de MCMC expuesto en la Sección 2.4.4 se encontraron los siguientes resultados: la estimación del intercepto corresponde a 2.25, retrasformando la variable por medio del exponencial se encontró que la predicción promedio equivale a 9.48. Mediante la ecuación (2.37), se determina una estimación puntual del rango de 281230.8 m; este valor indica que existe una dependencia espacial hasta de 281230.8 metros en promedio y que mas allá de esta distancia, las observaciones georeferenciadas por municipio son independientes. La ventaja que tiene realizar las estimaciones por medio de la metodología expuesta en Diggle y Ribeiro [8] es el poder medir la distancia hasta la cual la variable de estudio tiene influencia.

A continuación se muestra un resumen de los ECM obtendidos al modelar el número de personas con el apellido Gómez en cada modelo.

| Modelo | ECM |
|---|---|
| SAR | 13363.47 |
| CAR | 12582.43 |
| MLGM | 43.65 |
| Modelo Log Poisson | 10 |

Cuadro 4.3: ECM al usar los modelos estudiados

En la Tabla 4.4 se puede apreciar los ECM obtendido al modelar el número de personas con el apellidos Gómez en el departamento de Antioquia. En este caso, se puede apreciar que el Modelo Log Poisson y el MLGM lograron un excelente ajuste a los datos con respecto a los tradionales modelos SAR y CAR.

## 4.4. Importancia de los estudios de Isonimia a partir de datos georeferenciados y otras aplicaciones de los modelos expuestos

Al incluir técnicas espaciales en estudios de isonimia se busca analizar los procesos de distribución de la población a partir de la georeferenciación de la información. Al modelar la información mediante las técnicas expuestas, se pueden dar cuenta de la estructura espacial de la población y se pueden establecer relaciones geográficas, sociales e históricas de los procesos de migración y poblamiento. A partir del análisis espacial propuesto en este tipo de estudios, se puede formalizar hipótesis y preguntas de investigación en estudios geográficos, históricos y sociales, también se pueden corroborar procesos poblacionales obtenidios con otros medios y, en consecuencia adelantar nuevas propuesta de investigación; la metodología de base estadística sirve como argumento y validación para soportar análisis cualitativos de los fenómenos sociales percibidos en un territorio. Otros argumentos de estudios de isonimia son expuestos por Gómez y Muñeton [4].

Tal y como se extendió el uso de MLGM para describir y modelar fenómenos migratorios, es posible utilizar estas metodologías en campos de aplicación tales como:

En primer lugar, la información territorial siempre es agregada por barrios, municipios, departamentos, paises entre otros. Georeferenciada la información, se puede llegar modelar datos sociales, económicos, ambientales, urbanísticos entre otros, como instrumento para la toma de decisiones.

En segundo lugar, estas metodologías también se podrían utilizar en epidemiología espacial para modelar el comportamiento de enfermedades en donde su localización juega un papel importante.

En tercer lugar, mediante la aplicación de estos modelos, es posible describir fenómenos migratorios tales como el desplazamiento forzado en Colombia.

# Capítulo 5

# Conclusiones y trabajo futuro

Del estudio realizado se obtienen las siguientes conclusiones:

En primer lugar, si se cuenta con la matriz de vecindad, la variable respuesta se puede modelar mediante los modelos tradicionales SAR o CAR y determinar si existe correlación espacial en el área de estudio. Si se tienen indicios de que la correlación espacial es predominante en los vecinos más cercanos, se puede proponer el uso del modelo CAR; pero, hay que tener en cuenta que este modelo parte del supuesto de normalidad en los datos y de la construcción de una distribución conjunta de probabilidad.

En segundo lugar, si el supuesto de normalidad no se cumple; pero, los datos tienen una distribución miembro de la familia exponencial, las simulaciones realizadas en este trabajo, dieron evidencia de que la modelación se podría extender a los MLGM y realizar la estimación por máxima verosimilitud penalizada o por MCMC. En el desarrollo del trabajo, se mostró que al modelar por máxima verosimilitud penalizada un MLGM se lograba el mejor ajuste con inconvenientes de convergencia en algunos conjuntos de datos; situación, que obliga a realizar la modelación de un modelo Log Poisson por medio de MCMC.

Las simulaciones realizadas en este trabajo mostraron, que el modelo Log Poisson, aunque no lograba el mínimo ECM, tampoco excede los ECM obtenidos en los modelos SAR y CAR. Pero, si se realiza el ajuste de los datos por medio del Modelo Log Poisson se debe tener cuidado en el análisis de la información y no realizar interpolaciones en sitios carentes de sentido.

En tercer lugar, a pesar de que los modelos expuestos parten de supuestos distintos y cada uno se ajusta mediante metodologías distintas, estos, puede ser usados de forma complementaria ya que el objetivo en el fondo es modelar la información georeferenciada. Se podría utilizar

el cálculo de la correlación espacial obtenida en los modelos SAR y CAR como indicio de que es viable aplicar las técnicas de datos espaciales. Definida la existencia de correlación espacial, se emplearía máxima verosimilitud penalizada en la estimación de los parámetros del MLGM siempre y cuando no existan problemas de convergencia; esta metodología logra el mejor ajuste y por ende el mínimo ECM. Por último, si se encuentran problemas de convergencia, se modelaria con MCMC y se estimaría el modelo Log Poisson; la ventaja de esta metodología es que permite cuantificar la distancia euclidiana hasta donde puede haber dependencia espacial.

En cuarto lugar, el modelo marginal con varianza CAR presenta problemas de convergencia; el algoritmo que propone Schabenberger[19] funciona unicamente con los valores iniciales que él presenta en el algoritmo y los elementos de la matriz $V$ que se incluye en la expresión 2.20 son introducidos como sus inversos. Problemas similares se presentaron en la simulación conduciendo a problemas de convergencia.

Por último, la variación de $\kappa$ en la función de correlación Matérn no influye en el ajuste del modelo y en la predicción. Los ECM y estimacioens de los parámetros del modelo Log Poisson no influyen significativamente al variar el valor de $\kappa$.

Una propuesta a futuros estudios es utilizar la estimación obtenida del modelo Log Poisson obtendio por MCMC y tomarla como valores iniciales en la estimación de los parámetros de los MLGM realizada por máxima verosimilitud penalizada y verificar si se reducen los problemas de convergencia que se encontraron en este trabajo.

Tambien se propone en un futuro, crear otras estructuras a partir de una matriz de varianzas y covarianzas pueden ser un camino para afrontar los diversos problemas que tienen los modelos SAR y CAR.

En futuros estudios, utilizar un modelo espacial no paramétrico para simular los datos, y luego sobre estos datos simulados comparar los ajustes de las diferentes metodologías propuestas en este trabajo.

Considerar el fenómeno de sobre dispersión que podría estar involucrado en los ejercicios empíricos, y que posiblemente no se capturen con un modelo Poisson.

En un futuro, extender esta aplicación a estudios de isonimia en todo Colombia y a otras variables de conteo georeferenciadas.

# ANEXOS

## Anexo 1. Algoritmo utilizado para la estimación de los parámetros del MLGM

```
#Modelo lineal genralizado mixto (Gaussina)
library(MASS)
pdf("MLGMGauss.pdf")
parametros6<-function(y){
# attach(as(antioquia, "data.frame"))
antioquia$y1 <- coordinates(antioquia)[, 1]/1000
x1<-coordinates(antioquia)[, 1]/1000
x2<-coordinates(antioquia)[, 2]/1000
antioquia$y2 <- coordinates(antioquia)[, 2] /1000
sp1ape <- corSpatial(1, form = ~y1 + y2, type = "gaussian")
scorape <- Initialize(sp1ape, as(antioquia, "data.frame")[, c("y1", "y2")], nugget = FALSE)
antGLMMp <- glmmPQL(y~ x1 + x2, data = antioquia, family = poisson, random = ~1 | codigo, correlation = scorape)
glmm<-summary(antGLMMp)
ecm<-sum((antGLMMp$residuals)^2)
b0<-glmm$tTable[1,5]
b1<-glmm$tTable[2,5]
b2<-glmm$tTable[3,5]
return(c("ecm"=ecm,"b0"=b0,"b1"=b1,"b2"=b2))
mlgmagaus<-apply(prueba3,2,parametros6)
```

```
mlgmgau<-data.frame(t(mlgmagaus))
dev.off()
```

# Anexo 2. Algoritmo utilizado para la estimación de los parámetros del modelo Log Poisson

```
library(spdep)
library(rgdal)
require(SparseM)
require(quantreg)
require(emplik)
require(geoR)
require(geoRglm)
# Lectura de base de datos
ant<- readOGR(".", "Antioquia")
cord = coordinates(ant)
x1<-cord[, 1]
x2<-cord[, 2]
mil<-read.csv2("amil.csv", dec=".") # base general
y<-mil[,1] # Valores iniciales
sigma=0
kapp=2.5
phi=281560/(2*sqrt(kapp)) # formula del rango
# Lectura de datos como un Geodata
parametros<-function(y){
resp<-cbind(cord[,1] ,cord[,2], y)
co=c(sigma,phi) # Vector de parámetros
dat<-as.geodata(resp) # Deja los datos bajo estructura geoespacial
dat$cov.pars=co # asignar los parámetros
```

```
dat$kappa=kapp # asigna valor de kappa

set.seed(371) MCc <- mcmc.control(S.scale = 500, phi.sc = 100, n.iter = 100000, burn.in = 10000, thin = 100, phi.start = phi)

PGC <- prior.glm.control(phi.prior = "exponential", phi = phi,phi.discrete = seq(0,281560, by = 1000), tausq.rel = 0)

OC <- output.glm.control(sim.pred = T, quantile=c(0.025,0.5,0.975)) # cuartiles

locs <- cbind(x1,x2) #ubicaciones a predecir

pkb <- pois.krige.bayes(dat, loc = locs, prior = PGC,mcmc = MCc, out = OC) # Lista que contiene los resultados de la simulación

pe<-pkb$posterior$phi$mean

be<-pkb$posterior$beta$mean

se<-pkb$posterior$sigmasq$mean

pred<-pkb$predictive$median

ec<-sum(y-pred)^2

return(c("pe"=pe,"be"=be,"se"=se, "ec"=ec))

}

sim<-apply(mil,2,parametros)

simul<-data.frame(t(sim))

write.table(simul, file = "res_kappa_cero", append = FALSE, quote = TRUE, sep = "; ", eol = "\n", na = "NA", dec = ".", row.names = TRUE, col.names = TRUE, qmethod = c("escape", "double"))

# Agrupación de la información

d<-as.data.frame(simul)

Phi<-d$pe

Betta<-d$be.beta

Sigma<-d$se

Error_cuadrático<-d$ec

# Intervalos de confianza

iPhi<-quantile(Phi,c(0.025,0.5,0.975))

ibe<-quantile(Betta,c(0.025,0.5,0.975))
```

ise<-quantile(Sigma,c(0.025,0.5,0.975))

iec<-quantile(Error_cuadrático,c(0.025,0.5,0.975))

int<-rbind(iPhi,ibe,ise,iec)

media<-rbind(mean(Phi), mean(Betta), mean(Sigma), mean(Error_cuadrático))

estimacion<-cbind(int,media) write.table(estimacion, file = "est_kappa_cero", append = FALSE, quote = TRUE, sep = "; ", eol = "\n", na = "NA", dec = ".", row.names = TRUE, col.names = TRUE, qmethod = c("escape", "double"))

pdf("ribeirocero.pdf")

par(mfrow=c(4,1))

hist(Phi, xlab=expression(phi), ylab="Frecuencia", main="Histograma de frecuencias para las estimaciones de los parámetros y error cuadrático") hist(Betta,xlab=expression(beta), ylab="Frecuencia", main="")

hist(Sigma,xlab=expression(sigma), ylab="Frecuencia", main="")

hist(Error_cuadrático, xlab="Error cuadrático", ylab="Frecuencia", main="") dev.off()

# Bibliografía


[1] Besag, J. Y Gleaves, J.T. (1973), On the detection of spatial pattern in plan communities.

[2] Bresslow NE, Clayton DC (1993), Appoximate inference in generalized linear mixed models. JASA 88:9-9

[3] Bivand, R. y Pebesma, E (2008), Applied Spacial Data analysis with R.

[4] Chasco,C. (2003), Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales. Madrid, España.

[5] Cliff, A y y Ord, J (1973), Spatial autocorrelation. London: Pion.

[6] Cressie, N. (1993), Statistics for Spatial data Revised Edition, New York: John Wiley & Sons, Inc. 900 p

[7] Diggle, P y Ribeiro, P . (2007), Model-based Geostatistics.

[8] Diggle. P y Ribeiro P. (2007), Manual the geoR Package, http://www.leg.ufpr.br/geoR.

[9] Giraldo, Ramón (2009), Estadística espacial

[10] Gómez, S., Hinestrosa, P. y Muñeton, G., (2009), Analisis de los procesos poblacionales en el departamento de Antioquia, en la república de Colombia, a partir de las relaciones de parentesco existentes entre las poblaciones municipales.

[11] Haining, R.P.(2003), Spatial data analysis: Theory and practice. Cambridge University Press, Cambridge.

[12] Journel, AG y Huijbregts CJ (1978), Geoestadística Minería, Londres, Academic Press

[13] Liang, K y Zeger, S (1986), Longitudinal data analysis using generalized lineal Models.Biometrika, 73:13-22

[14] McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models, Second Edition. Chapman and Hall, New York.

[15] Nelder y Wedderburn (1772), Generalized Lineal Models

[16] Ord, J. (1975), “Estimation Methods for Models of Spatial Interaction", Journal of the American Statistical Association, núm. 70, pp. 120-126.

[17] Pinheiro, J.C. y Bates, D.M.(2000), Mixwd-Effects Models in S and S-Plus, Springer, New York.

[18] Schabenberger,O (2002) y Pierce, Contemporary Statistical Models. Crc. Press

[19] Schabenberger,O (2005) y Gotmway. C, Statistical Methods for Spatial Data Analysis. Chapman y Hall/Crc.

[20] Whittle, P. (1954), On stationary processes in the plane, Biometrica, 41:434-449

[21] Wolfinger, R.D. y O’Connell, M. (1993) Generalized linear mixed models: a pseudo-likelihood approach. Journal of Statistical Computing and Simulation, 48:233–243.

[22] Zhao, Y. y M. Wall, M (2004) Investigating the Use of the Variogram for Lattice Data, Journal of Computational and Graphical Statistics, Volumen 13, Number 3, Pages 719-738

# CIBERGRAFIA

[1] Lista de R, la discusión se titula: “false convergence” (2005). Sitio web, http://r.789695.n4.nabble.com/unexpected-quot-false-convergence-quot-td790625.html

[2] Página CRC Press Company. Sitio web, http://www.crcpress.com