Cuadernos Latinoamericanos de Administración
ISSN: 1900-5016
cuaderlam@unbosque.edu.co
Universidad El Bosque
Colombia

Borda Hernández, Ricardo Alberto; Iral Palomino, Rene; Roy Cabrera, Kenneth
Aplicación de los modelos lineales generalizados mixtos en el modelamiento de datos de conteo georeferenciados por municipios en el departamento de Antioquía.
Cuadernos Latinoamericanos de Administración, vol. VIII, núm. 15, julio-diciembre, 2012, pp. 69-76
Universidad El Bosque
Bogotá, Colombia

Aplicación de los modelos lineales generalizados mixtos en el modelamiento de datos de conteo georeferenciados por municipios en el departamento de Antioquía.[1]

Application of generalized linear mixed models in the modeling of georeferenced count data in the municipalities in the department of Antioquía.

Aplicação dos modelos lineares generalizados mistos na modelação de dados de contagem georeferenciados em municípios no estado de Antioquía

Ricardo Alberto Borda Hernández. [2]
Rene Iral Palomino. [3]
Kenneth Roy Cabrera. [4]

Resumen

El origen de este trabajo se fundamenta en la necesidad de modelar estadísticamente datos de conteo georeferenciados en polígonos irregulares tales como: número de homicidios por barrio, número de habitantes por localidad, enfermos por municipio, entre otros; con el objetivo de encontrar algún tipo de dependencia espacial a partir de la localización geográfica.

El estudio pretendió comparar dos tipos de modelos lineales generalizados mixtos (MLGM), uno cuya estimación de los parámetros del modelo parte de la aplicación de Cadenas de Markov de Monte Carlo (MCMC) y el otro por medio de máxima verosimilitud penalizada; además, se hicieron otras comparaciones con el modelo tradicional auto regresivo simultáneo (SAR) y el modelo auto regresivo condicional (CAR); modelos que parten del supuesto de normalidad, invertibilidad de la matriz de varianzas y covarianzas, y construcción de una matriz de vecindad, supuestos que no necesariamente deben cumplirse con MLGM.

Se encontró, que los MLGM dan indicio de ser una alternativa en el modelamiento de datos de conteo y se comprobó con una aplicación a partir de la georeferenciación por municipio y modelación de los 200 apellidos más frecuentes de Antioquia, en donde igualmente se concluyó que los MLGM muestran el menor error cuadrático medio (ECM).

Palabras clave: apellidos, datos georeferenciados, error cuadrático medio, familia exponencial, modelos lineales generalizados mixtos, normalidad.

Abstract

The origin of this work is based on the need to model statistically georeferenced count data in irregular polygons such as: homicides by area, population density per city, municipality sick people, amongst others, with the aim of finding some kind of spatial dependence from geographical location.

Resumo

A origem deste trabalho se fundamenta na necessidade de modelar estatisticamente dados de contagem georeferenciados em polígonos irregulares tais como: número de homicídios por bairro, número de habitantes por área, doentes por município, entre outros, com o objetivo de encontrar algum tipo de dependência espacial a partir da localização geográfica.

Recibido el 18/09/2012 Aprobado 10/11/2012

1. Artículo de investigación
2. M.Sc. Estadística, Universidad Nacional de Colombia. Especialista Matemática Aplicada, Universidad Sergio Arboleda. Licenciado en Matemática y Estadística., Universidad Pedagógica y Tecnológica de Colombia. Sede Duitama. Docente programa Administración de Empresas, Universidad El Bosque. riabordahe@unal.edu.co , bordaricardo@unbosque.edu.co
3. Magister Universidad Nacional de Colombia - Sede Medellín. Estadística.
M.Sc. Estadística Universidad Nacional de Colombia - Sede Medellín, Pregrado Matemática. Universidad Nacional de Colombia - Sede Medellín. Docente programa Estadística, Universidad Nacional de Colombia - Sede Medellín
4. M.Sc(c) Estadística Pregrado/Universitario Geología. Universidad Nacional de Colombia - Sede Medellín. Universidad Nacional de Colombia - Sede Medellín

The study intended to compare two types of generalized linear mixed models (GLMM), one whose estimate model parameters from the application of Markov Chains Monte Carlo (MCMC) and the other through maximum penalized likelihood, in addition, other comparisons made with the traditional model simultaneous autoregressive (SAR) and the conditional autoregressive model (CAR) models that assume normality, invertibility of the covariance matrix, and construction of a neighborhood matrix, assumptions not necessarily be met with MLGM.

It was found that the indication given GLMM be an alternative in modeling count data and found an application from the georeferencing by town and modeling of the 200 most common surnames of Antioquia, which likewise concluded that GLMM show the least square error (LSE).

Key words: surnames, georeferenced data, least square error, exponential family, generalized linear mixed models, normality

O estudo pretendia comparar dois tipos de modelos lineares generalizados mistos (MLGM), um cuja estimação dos parâmetros do modelo parte da aplicação de Cadeias de Markov de Monte Carlo (MCMC) e o outro por meio de máxima plausibilidade penalizada. Além disso, foram feitas outras comparações com o modelo tradicional auto-regressivo simultâneo (SAR) e o modelo auto-regressivo condicional (CAR), modelos que partem do suposto de normalidade, invertibilidade da matriz de variâncias e covariâncias, e construção de uma matriz de vizinhança, supostos que não necessariamente devem cumprir-se com MLGM.

Encontrou-se que os MLGM dão indício de ser uma alternativa na modelação de dados de contagem e se comprovou com uma aplicação a partir da georeferenciação por município e modelação dos 200 sobrenomes mais frequentes de Antioquia, onde, da mesma forma, conclui-se que os MLGM mostram o menor erro quadrático médio (ECM).

Palavras-chave: sobrenomes, dados georeferenciados, erro quadrático médio, família exponencial, modelos lineares generalizados mistos, normalidade.

Introducción

La estadística espacial es una herramienta que permite analizar información a partir de la ubicación espacial de las observaciones; para esto, cada dato recolectado necesita ser asociado con su coordenada geográfica (Georeferenciación). Áreas del saber como: la geología, la minería, las ciencias ambientales, las ciencias de la salud, las ciencias sociales, la administración, entre otras, hoy en día pueden georeferenciar la información recolectada en sus investigaciones con el objetivo de encontrar tendencias de tipo espacial, que la estadística clásica no alcanza a observar.

Ramas de la estadística espacial como: Geoestadística, Lattices y Patrones Puntuales se pueden convertir en una herramienta que proporciona información consolidada, que aporte pistas sobre las dinámicas ambientales, sociales y culturales de la población en estudio; convirtiéndose en un excelente complemento del trabajo cualitativo para la toma de decisiones. Dentro de las posibles aplicaciones de la estadística espacial, se encuentran los estudios de Isonimia, los cuales buscan caracterizar la distribución de la población por medio del análisis de frecuencia y distribución de apellidos de los pobladores con el fin de establecer relaciones de parentesco y origen. Estos estudios se han limitado por no tener en cuenta el componente espacial.

La investigación "Aplicación de los modelos lineales generalizados mixtos en el modelamiento de datos de conteo georeferenciados por municipios en el departamento de Antioquia" quiso, mediante simulación mostrar que los MLGM son capaces de describir con mayor precisión datos de conteo espaciales, precisión que no se logra con los tradicionales modelos simultáneos autoregresivos (SAR) y los modelos condicionales autoregresivos (CAR). Si el resultado es positivo, en investigaciones futuras se podría aludir los supuestos de normalidad en los datos, invertibilidad de la matriz de varianzas y covarianzas, y construcción de las tediosas matrices de vecindad, supuestos e inconvenientes que a menudo dificultan este tipo de estudios.

1. Modelos SAR, CAR y MLGM de interés en el estudio

   Los primeros esfuerzos por modelar datos georeferenciados agregados en polígonos regulares e irregulares, corresponden a los modelos tradicionales simultáneos autoregresivos (SAR), propuestos por Whittle[5] y los modelos condicionales autoregresivos (CAR) propuestos por Besag[6] ; dichos modelos, parten del supuesto de normalidad en los datos, invertibilidad de la matriz de varianzas y covarianzas y la construcción de un matriz de vecindad. Estos modelos se explican a continuación:

5. SWITLLE, P. On stationary processes in the plane. Biometrika Vol. 41, No. 3/4 (Dec. 1954), p. 434-449

6. BESAG, Julian y GLEAVES, Timothy. On the detection of spatial pattern in plan communities. England: University of Newcastle upon Tyne. Biometrics Vol. 32, No. 3 (Sep. 1976), p. 659-667

1.1. Modelos simultáneos auto regresivos (SAR)

El modelo SAR parte de una autoregresión espacial del vector de errores residuales, *e(s)*, en un modelo de regresión lineal con respuesta gausiana de la siguiente forma:

$$Z(s) = X(s) \times \beta + \rho \times W \times e(s) + v$$

donde, *Z(s)* es una variable aleatoria que representa la distribución de todas las posibles realizaciones *z(s)* en el sitio s, ρ corresponde al parámetro de autocorrelación espacial, v corresponde a los errores residuales con media cero y matriz diagonal de varianzas y covarianzas $\Sigma_v = diag\left[\sigma_1^2,...,\sigma_n^2\right]$ y W corresponde a una matriz de vecindad. La matriz de vecindad W que se utilizó en el estudio se describe a continuación:

$$w_{ij} = \begin{cases} 1, & \text{si el sitio i y j son vecidos} \\ 0, & \text{si el sitio i y j no son vecidos} \end{cases}$$

Z(s) se distribuye como una normal multivariada con media:

$$E\left[Z(s)\right] = X(s) \times \beta$$

Asumiendo Σ_v sobre un único parámetro σ_s^2 tal que $\Sigma_v = \sigma_s^2 I$, la matriz de varianzas y covarianzas se define como:

$$VAR\left[Z(s)\right] = \sigma_s^2 \times (I - \rho \times W)^{-1}(I - \rho \times W')^{-1}$$

La forma usual de realizar las estimaciones de los parámetros es por mínimos cuadrados generalizados o por máxima verosimilitud.

1.2. Modelos condicionales autoregresivos (CAR)

En el caso del Modelo CAR, se asume $Z(s_i)$, como una variable aleatoria en el sitio s_i, y que esta depende solo del conjunto de vecinos, i.e, $Z(s_i)$ depende de $Z(s_j)$ solo si la localización de s_i está en el conjunto de vecinos, N_i, de s_i; es decir, la observación que se obtiene a partir de $Z(s_i)$, únicamente depende de los vecinos más cercanos.

Bajo el modelo condicional autoregresivo, se construye el modelo espacial autoregresivo CAR definido por $f\left(Z(s_i)|Z(s_j), s_j \in N_i\right)$, , si cada una de estas distribuciones condicionales es gausiana, entonces estas distribuciones se pueden modelar por medio de:

$$E\left[Z(s)\right] = X(s)\beta$$

y matriz de varianzas y covarianzas:

$$VAR\left[Z(s)\right] = (I - \rho W)^{-1}\Sigma_c$$

donde, $\Sigma_c = diag\left[\sigma_1^2,...,\sigma_n^2\right]$. De forma similar que en el modelo SAR, Σ_c es representado sobre un único parámetro σ_c^2 tal que $\Sigma_c = \sigma_c^2 I$.

La literatura más reciente ofrece como ejemplo las adaptaciones de[7] Schabenberger y Ribeiro[8] a los MLGM, en estas adaptaciones se realiza la estimación de los parámetros por medio de máxima verosimilitud penalizada[9] y Cadenas de Markov de Monte Carlo (MCMC) respectivamente. Estos modelos permiten moverse fuera del supuesto de normalidad, y extenderse a datos que se deriven de distribuciones miembro de la familia exponencial tales como: la distribución binomial, poisson, exponencial, gamma, entre otras. Los MLGM que se emplearon en esta investigación son descritos a continuación:

1.3. MLGM con varianza CAR propuesto en Schabenberger

El autor 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 por medio del MLGM con varianza CAR:

$$E\left[Z(s)\right] = \mu(s)$$

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:

$$g\left(\mu(s)\right) = log\left\{\mu(s)\right\}$$

Esta expresión corresponde al promedio de un proceso gausiano dado por:

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

donde, n_i corresponde al número de habitantes por municipio, β_0, β_1 y β_2 son parámetros del modelo, x_1 y x_2 corres-

7. SCHABENBERGER, Oliver y GOTWAY, Carol A. Statistical Methods for Spatial Data Analysis. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton, 2005, 359-382 p.

8. DIGGLE, Peter y Ribeiro, Paulo. Model based Geostatistics. Brasil: Springer Series in Statistics 2007, 124 p.

9. Statistical Methods for Spatial Data Analysis. Op. Cit., 362-365 p.

ponde 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:

$$VAR[Z(s)] = \sigma_0^2 V_\mu + \sigma_1^2 V_\mu^{\frac{1}{2}} (I - \rho W) V_\mu^{\frac{1}{2}}$$

en donde σ_0^2 mide la sobredispersión asociada a cada sitio s_i, σ_1^2 miden la sobredispersión asociada con la estructura de correlación espacial, V_μ corresponde a una matriz diagonal de orden n con los términos de la varianza $V(\mu(S_i))$ sobre la diagonal, $V_\mu^{\frac{1}{2}}$ es una matriz diagonal de orden n con los términos de la raíz cuadrada de la varianza $\sqrt[2]{V(\mu_i)}$, I corresponde a la matriz identidad, ρ es el coeficiente de correlación espacial y W corresponde a la matriz de vecindad explicada en la sección 1.1. De este modelo, se generaron 1000 simulaciones sobre los 125 municipios de Antioquia, estos datos se ajustaron a los modelos SAR, CAR y MLGM Log poisson propuestos en Ribeiro.

La dificultad en la modelación SAR, CAR y MLGM con varianza CAR radica en que se debe garantizar la invertibilidad de la expresión (I-ρW), expresión que depende directamente de la matriz de vecindad W, pero, en este tipo de modelamiento no se necesita el supuesto de normalidad.

1.4. MLGM Log Poisson

Ribeiro, utiliza un enfoque denominado MLGM 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. En la forma más simple del modelo, los $Z(s_i)$ son conteos condicionales independientes Poisson, con valor esperado condicional μ_i dado por:

$$g\left[(Z(s_i)|X(s_i)\right] = \log(\mu_i) = \beta + S(s_i)$$

donde, $S(.)$ es un proceso gausiano estacionario con media cero que corresponde a los efectos aleatorios del modelo mixto con varianza σ^2, función de correlación $\rho(u)$ y β corresponde a los efectos fijos. La correlación espacial entre $S(s)$ y $S(s')$ se mide a partir de la correlación Matérn:

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

donde, $K_\kappa(u/\phi)$ es la función modificada de Besel de orden κ, con κ>0, el cual corresponde al parámetro de forma. ϕ>0 es el parámetro de escala, el valor de ϕ depende directamente del rango α. De este modelo, se generaron 1000 simulaciones sobre los 125 municipios de Antioquia, estos datos se ajustaron a los modelos SAR, CAR y MLGM con varianza CAR. El MLGM Log Poisson no posee una matriz de vecindad.

2. Materiales y Métodos

Luego de entender los modelos para datos de conteo georeferenciados: SAR, CAR y MLGM, se planteó un ejercicio de simulación que involucró los 125 municipios del departamento de Antioquia; este estudio se dividió en tres etapas: dos escenarios de simulación generados a partir de MLGM con varianza CAR y MLGM Log Poisson y una aplicación a datos reales. La aplicación a datos reales surge de los esfuerzo de Gómez y Muñeton10 en modelar los procesos distribucionales de la población en el departamento de Antioquia luego del análisis de frecuencias y distribución de apellidos de sus pobladores para establecer relaciones de parentesco y origen (estudios de Isonimia). Este tipo de estudios se ha limitado únicamente al estudio de frecuencias y distribución de apellidos de sus pobladores sin tener en cuenta el comportamiento espacial, razón por la cual es de interés adherir el componente espacial a este tipo de investigación.

Dada la aplicación que se pretendía hacer a datos reales, se definió previamente en las simulaciones como variable respuesta al número de personas con el apellido . Esta variable se escaló entre 0 y 2000 para evitar la influencia de municipios sobre poblados, como es el caso de la ciudad de Medellín.

2.1. Primer escenario de simulación: datos obtenidos del MLGM con varianza CAR propuesto en Shabenberger[11].

Un primer escenario de simulación es obtenido de los MLGM con varianza CAR que se muestra en la sección 1.3, de este modelo se obtienen 1000 realizaciones; luego, de cada realización se estimaron los parámetros de los modelos SAR, CAR y MLGM

10. GOMEZ, Santiago; HINESTROSA, Paula y MUÑETON, Guberney. Procesos poblacionales en Antioquia, Colombia, a partir de relaciones de parentesco intermunicipales. En: Papeles poblacionales, Julio-septiembre, número 057, Universidad Autónoma del Estado de Mexico, Toluco, p 257-274, 2008

11. Statistical Methods for Spatial Data Analysis. Op. Cit., 359-382 p

Log Poisson expuestos en la sección 1.1, 1.2 y 1.4 con el fin de determinar cuál de estas metodologías lograba ajustar mejor los datos y por ende mostraba un ECM menor. El objetivo en esta etapa fue evaluar la capacidad que tienen las MCMC para encontrar los parámetros de un MLGM Log Poisson expuesto en Diggle y Ribeiro en comparación a la estimación por máxima verosimilitud utilizada en los modelos tradicionales SAR y CAR.

2.2. Segundo escenario de simulación: datos obtenidos del MLGM propuesto en Ribeiro[12].

Un segundo escenario de simulación es obtenido del MLGM Log Poisson propuesto por Diggle y Ribeiro, mostrado en la sección 1.4; nuevamente se simularon 1000 realizaciones provenientes de este modelo, la variable respuesta , también, corresponde al número de personas con el apellido en una escala entre 0 y 2000. Cada simulación fue estimada a partir de los modelos SAR, CAR y MLGM con varianza CAR. El objetivo fue evaluar la capacidad que tiene la metodología de máxima verosimilitud penalizada para encontrar los parámetros de un MLGM con varianza CAR en comparación a la estimación por máxima verosimilitud utilizada en los modelos tradicionales SAR y CAR.

2.3. Aplicación a datos reales

La tercera parte del estudio consistió en seleccionar los 200 apellidos más frecuentes en el departamento de Antioquia. Cada apellido se georeferenció en el centroide de cada municipio y fue ajustado a los modelos: SAR, CAR, MLGM con varianza CAR y MLGM Log Poisson propuesto por Diggle y Ribeiro.

El objetivo, fue extender y corroborar los resultados obtenidos en las simulaciones por medio de datos reales, y mostrar que la estadística espacial es una excelente herramienta que proporciona información consolidada en estudios de Isonimia. Para esta etapa se contó con las bases de datos del SISBEN del año 2005 desagregadas por municipio y apellido del departamento de Antioquia.

Las simulaciones y estimaciones de los modelos fueron realizadas en el software estadístico R, versión 2.14.0.

3. Resultados y discusión

3.1. Resultados obtenidos del ajuste de los Modelos SAR, CAR y MLGM Log Poisson.

Para la evaluación de las realizaciones generadas del MLGM con varianza CAR, se ajustaron las 1000 realizaciones a los modelos SAR, CAR y MLGM Log Poisson y se calculó el ECM. Finalmente, el modelo que tenga los menores ECMs habrá realizado el mejor ajuste.

Los resultados encontrados muestran que los modelos tradicionales SAR y CAR generan los ECMs más altos; situación contraria, se puede apreciar en los ECMs que se obtuvieron del MLGM Log Poisson. Los logaritmos de los ECMs se pueden apreciar en la Figura 1, resultados que fueron llevados a la escala logarítmica para mejorar su visualización.

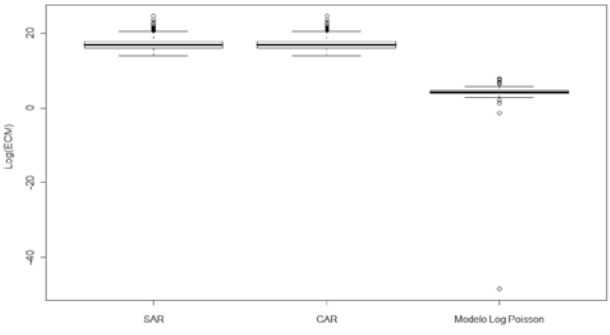

Figura 1. Logaritmos de los ECMs del ajuste de datos de un MLGM con varianza CAR a partir de los modelos: SAR, CAR y MLGM Log Poisson.

El 99.6% de los ECMs al ajustar los datos al MLGM Log Poisson fueron inferiores a 1000 personas, mientras que el 100% de los ECMs obtenidos por medio de los modelos SAR y CAR registraron ECMs superiores a 1000 personas. Estos datos dan evidencias para concluir que la estimación que utiliza MCMC para encontrar los parámetros de un MLGM Log Poisson pueden ser de mayor eficiencia que la estimación por medio de máxima verosimilitud utilizada en los modelo SAR y CAR.

El MLGM Log Poisson no requiere de la construcción de la matriz de vecindad y por ende no tendrá los problemas de invertibilidad que se encuentran en los modelos SAR, CAR y MLGM con varianza CAR, situación favorable, ya que

12. Model based Geostatistics. *Op. Cit., 79-83 p*

su construcción es demorada y la invertibilidad de la matriz de varianzas y covarianzas puede ser incierta. Esta investigación concluye que los datos de conteo provenientes de una distribución Poisson se pueden modelar por medio de MLGM Log poisson, situación limitante en los modelos SAR y CAR en donde los datos deben derivarse de la distribución normal.

3.2. Resultados obtenidos del ajuste de un Modelo SAR, CAR y MLGM con varianza CAR.

Para la evaluación de las realizaciones generadas de un MLGM Log Poisson, se ajustaron las 1000 realizaciones al modelo SAR, CAR y MLGM con varianza CAR y se calculó el ECM. De forma similar a la sección 3.1, el modelo que tenga los menores ECMs realizó el mejor ajuste.

Los resultados encontrados muestran que los modelos tradicionales SAR y CAR generaron los ECMs más altos, situación contraria, se puede apreciar en los ECMs que se obtuvieron del MLGM con varianza CAR. El logaritmo de los ECMs se puede apreciar en la Figura 2, resultados que fueron llevados a la escala logarítmica para mejorar su visualización.

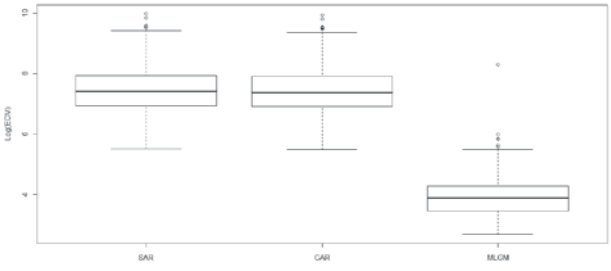

Figura 2. Logaritmo del ECM al ajustar datos MLGM Log Poisson por medio de los modelos SAR, CAR y MLGM con varianza CAR.

El 91.3% de los ECMs obtenidos al ajustar datos al MLGM con varianza CAR son inferiores a 100 personas. Situación muy distinta al ajustar las 1000 realizaciones a los modelos SAR, CAR e incluso al MLGM Log Poisson de la sección 3.1, en donde se encontró que el 62.4% de los ECMs para el MLGM Log Poisson fue inferior a 100 personas. Aunque, el 91.3% y 62.4% se obtuvieron de realizaciones distintas, estos resultados dan indicios para concluir que máxima verosimilitud penalizada logra obtener un mejor ajuste que MCMC para llegar a MLGM con varianza CAR y MLGM Log Poisson respectivamente.

3.3. Resultados obtenidos a partir de la aplicación a datos reales.

Al comparar los ECMs obtenidos en los modelos SAR y CAR de la sección 3.1 y 3.2, se observa que los ECMs de la segunda sección son menores a los obtenidos en la primera sección; mientras que los ECMs obtenidos para los MLGM Log Poisson son aproximadamente los mismos que los ECMs obtenidos en el MLGM con varianza CAR. En esta sección, se modeló el número de personas de los 200 apellidos más frecuentes del departamento de Antioquia por medio de las cuatro metodologías, como criterio adicional a los resultados obtenidos en las secciones 3.1 y 3.2 y como aplicación a datos reales. Los resultados se muestran a continuación:

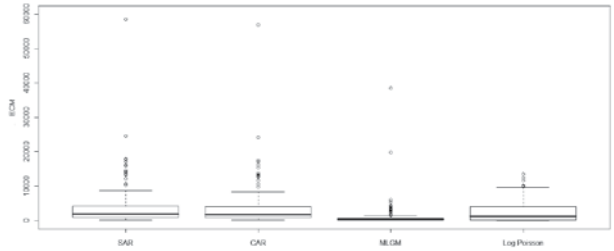

Figura 3. ECMs de los modelos SAR, CAR, MLGM con varianza CAR y MLGM Log Poisson al ajustar los 200 apellidos más frecuentes de Antioquia.

Los ECMs de los modelos SAR, CAR y MLGM Log Poisson son los más altos, presentando algunas sobrestimaciones en los modelos SAR y CAR. Los ECMs más bajos se presentaron en la estimación por máxima verosimilitud penalizada al ajustar los datos al MLGM con varianza CAR.

A pesar del excelente ajuste que mostró máxima verosimilitud penalizada, se encontraron apellidos como Cuesta y Palacio que al ser modelados mostraron ECMs superiores a 80.000 personas, la modelación de los demás apellidos presentó ECMs muy bajos.

Finalmente, se muestra la distribución de las personas que se apellidan Gómez en el departamento de Antioquia.

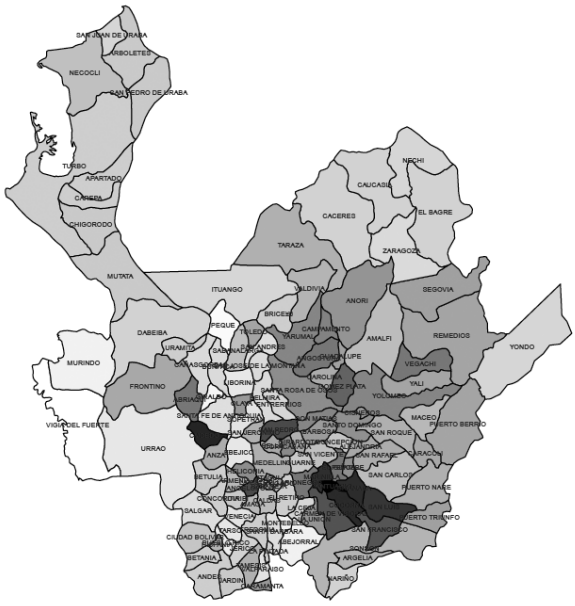

Figura 4. Número de personas con el apellido Gómez en el departamento de Antioquia en una escala entre 0 y 2000 habitantes.

En la Figura 4. se puede apreciar que las personas apellidadas Gómez se presentaron con mayor frecuencia en la subregión de Oriente, Nordeste y Magdalena Medio principalmente. 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 164 personas por cada 2000 habitantes; también se destaca que los municipios ubicados en el sur oriente de este municipio toman valores muy similares al Santuario; estos municipios son: Cocorná, Granada y San Luis con 118, 116 y 116 personas por cada 2000 habitantes respectivamente. De forma similar, sucede en municipios ubicados en el nor-occidente de este municipio tales como: Marinilla y Rionegro con 114 y 66 personas, respectivamente. También se puede observar, que el apellido Gómez ubicado en el suroeste del departamento de Antioquia presenta un total de 122 habitantes por cada 2000. Se destaca, que el número de personas con el apellido Gómez se presentaron en forma constante en la región del Urabá antioqueño con un promedio de 20 personas con este apellido por cada 2000 habitantes[13].

13. BORDA, Ricardo; IRAL, Rene y CABRERA, Kenneth. 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. Medellín: Universidad Nacional de Colombia, 2011

4. Conclusiones y recomendaciones

Al ajustar las realizaciones al modelo SAR, CAR y MLGM Log Poisson se encontró que este último modelo generó los menores ECMs, situación favorable, ya que estos modelos no requieren de la construcción de la matriz de vecindad W, no se tendrían posibles inconvenientes en la invertibilidad de la matriz de varianzas y covarianzas y se podría extender su aplicación a datos que conteo que se deriven de una distribución poisson.

Al ajustar las realizaciones al modelo SAR, CAR y MLGM marginal con varianza se obtiene de forma similar que el MLGM generó los menores ECMs, situación igualmente favorable, ya que estos modelos se podrían extender a datos de conteo que se deriven de una distribución Poisson.

Al realizar la aplicación a datos reales, se encontró que los ECMs de los modelos SAR, CAR y MLGM Log Poisson son los más altos, presentando algunas sobrestimaciones en los modelos SAR y CAR. Los ECMs más bajos se presentaron en la estimación por máxima verosimilitud penalizada al ajustar los datos al MLGM con varianza CAR con algunos problemas de convergencia en algunos apellidos.

Al encontrar problemas de convergencia en el ajuste de datos al MLGM con varianza CAR por medio de máxima verosimilitud penalizada, se podrían realizar las estimaciones por medio de MLGM Log Poisson. La ventaja de este tipo de modelación es que permite cuantificar la distancia euclidiana hasta donde puede haber dependencia espacial.

Este estudio, se extendió a la distribución Poisson, en posteriores estudios se podría aplicar este tipo de simulaciones a otras distribuciones miembro de la familia exponencial y poder generalizar los resultados a otras distribuciones.

Se propone en futuros estudios, utilizar un modelo espacial no paramétrico para simular los datos, y luego comparar los ajustes de las diferentes metodologías propuestas en este trabajo. También se recomienda, considerar el fenómeno de sobredispersión que podría estar involucrado en los ejercicios empíricos, y que posiblemente no se capturen con un modelo Poisson.

Bibliografía

BESAG, Julian y GLEAVES, Timothy. On the detection of spatial pattern in plan communities. England: University of Newcastle upon Tyne. *Biometrics* Vol. 32, No. 3 (Sep. 1976), p. 659-667.

BIVAND, Roger y PEBESMA, Edzer. Applied Spatial Data Analysis with R, Springer, 2011

BORDA, Ricardo; IRAL, Rene y CABRERA, Kenneth. 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. Medellín: Universidad Nacional de Colombia, 2011

CHASCO,Coro. Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales. Madrid: Consejería de Economía e Innovación Tecnológica, 2003.

CRESSIE, Noel. Statistics for Spatial data Revised Edition, New York: John Wiley & Sons, 1993.

DIGGLE, Peter y Ribeiro, Paulo. Model based Geostatistics. Brasil: Springer Series in Statistics, 2007.

GIRALDO, Ramón. Estadística espacial. Bogotá: Universidad Nacional de Colombia, 2009.

GOMEZ, Santiago; HINESTROSA, Paula y MUÑETON, Guberney. Procesos poblacionales en Antioquia, Colombia, a partir de relaciones de parentesco intermunicipales. En: Papeles poblacionales, Julio-septiembre, número 057, Universidad Autónoma del Estado de México, Toluco, p 257-274, 2008.

HAINING, Robert. Spatial data analysis: Theory and practice. Cambridge: University Press, 2003.

MCCULLAGH, P. Y NELDER, J.A. Generalized Linear Models, Second Edition. New York: Chapman and Hall, 1989

ORD, Keith. Spatial autocorrelation. London: Pion, 1973.

PINHEIRO, Jose. y BATES, Douglas. Mixwd-Effects Models in S and S-Plus. New York: Springer, 2000.

SCHABENBERGER, Oliver y GOTWAY, Carol. Contemporary Statistical Models. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton. Press, 2002.

SCHABENBERGER, Oliver y GOTWAY, Carol A. Statistical Methods for Spatial Data Analysis. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton, 2005.

WITLLE, P. On stationary processes in the plane. Biometrika Vol. 41, No. ¾, P. 434-449, 1954.

WOLFINGER, R.D. y O'CONNELL. M.Generalized linear mixed models: a pseudo likelihood approach. Journal of Statistical Computing and Simulation, 48:233–243, 1993.