

Modelos de regresión espacial para describir el comportamiento del cáncer gástrico en Colombia para el periodo 2005-2012

Spatial regression models to describe gastric cancer behavior in Colombia 2005-2012

Francisco Javier Sarmiento Parra¹, Cristhian Eduardo Murcia Galeano², Luis Fernando Santa Guzmán³, Luis Fernando Gómez Rodríguez⁴

“Cómo citar este artículo: Sarmiento Parra, F. J., Murcia Galeano, C. E., Santa Guzmán, L. F. y Gómez Rodríguez, L. F. (2017). Modelos de regresión espacial para describir el comportamiento del cáncer gástrico en Colombia para el periodo 2005-2012. *Análisis Geográficos*, 52, 155-173.

Resumen

El cáncer gástrico encabeza la lista de muertes causadas por enfermedades en Colombia, y por ello su prevención, detección temprana y tratamiento se ha convertido en política de Estado. Teniendo en cuenta lo anterior, este estudio se centró en describir el comportamiento del cáncer gástrico en Colombia a nivel municipal para el periodo 2005-2012, mediante modelos de regresión espacial clásicos que permitieron identificar factores potenciales de riesgo asociados al fenómeno y mapas de la enfermedad. Mediante este análisis se logró determinar patrones y agregaciones espaciales, y clasificar el territorio por zonas, de acuerdo con el nivel de riesgo de prevalencia de la enfermedad.

Palabras clave: modelos lineales generalizados mixtos, regresión binomial negativa, vectores propios de Moran, cáncer gástrico.

Abstract

Gastric cancer heads the list of deaths caused by diseases in Colombia, and therefore, its prevention, early detection, and treatment, has become a State policy. This research describes the behavior of gastric cancer in Colombia at the municipal level for the period 2005-2012, using classic spatial regression models, that allowed to identify potential risk factors associated with the phenomenon, and maps of the disease. Through this analysis was possible to determine spatial patterns and clusters, and to classify the territory by zones according to level of risk of prevalence of the disease.

Keywords: *generalized linear mixed model; negative binomial regression; Moran eigenvectors; gastric cancer.*

¹ Universidad Distrital Francisco José de Caldas, Bogotá, Colombia. Correo: fjsarmientop@gmail.com.

² Universidad Distrital Francisco José de Caldas, Bogotá, Colombia. Correo: cristhianmurcia182@gmail.com.

³ Universidad Distrital Francisco José de Caldas, Bogotá, Colombia. Correo: lfsantag@unal.edu.co.

⁴ Universidad Distrital Francisco José de Caldas, Bogotá, Colombia. Correo: luuruena@yahoo.es.



Introducción

La mayor prevalencia como primera causa de muerte por cáncer en el país, tanto en hombres como en mujeres, es el cáncer de estómago, también llamado cáncer gástrico (Daza, 2012), el cual posee varias características relevantes por las que es pertinente analizarlo, como el hecho de que presenta un notorio patrón de riesgo que muestra relación con la altitud en toda la región andina (lo cual destaca, además, diferencias entre grupos étnicos presentes en las regiones) (Piñeros *et al.*, 2010), y el hecho de que existe una alta asociación entre la adquisición de la bacteria *Helicobacter pylori* (*H. pylori*) en la niñez y el desarrollo de un cáncer de estómago en la adultez (Correa, 2011), identificándose que dicha infección bacteriana es el factor de riesgo más importante para el desarrollo del cáncer gástrico (Strebel *et al.*, 2010).

A partir de diversos estudios epidemiológicos realizados por varios autores, se sabe además que este tipo de cáncer se asocia con distintos factores de riesgo bien identificados, tales como los ingresos, las condiciones de vida, el nivel educativo o el tipo de seguridad social que tiene la población (Daza, 2012), e incluso factores como las condiciones de vida en la niñez, el tamaño de las familias, la higiene o la calidad del agua que se consume han sido identificados como relevantes para la prevalencia del *H. pylori* (Strebel *et al.*, 2010), y, por tanto, corresponderían a factores indirectos importantes para la generación de cáncer gástrico.

Debido a las características y los comportamientos de las enfermedades, las cuales se encuentran en función de las

dimensiones temporal, espacial y poblacional, se hace necesario utilizar herramientas y métodos como los que proporcionan la estadística espacial, cuyo objetivo es el análisis y modelado de estructuras de dependencia espacial, que además son de gran utilidad para comprobar hipótesis etiológicas (Waller & Gotway, 2004). Teniendo en cuenta esto, en el presente documento se describe el comportamiento del cáncer gástrico en Colombia a nivel municipal para los años 2005, 2008 y 2012, mediante modelos de regresión espacial y el mapeo del riesgo de la enfermedad, con lo que se busca explicar las causas que lo originan y dar pautas concisas para una política de salud confiable.

Materiales y métodos

Los softwares utilizados para llevar a cabo el análisis son:

- R v3.1.1: plataforma sobre la que se desarrollan todos los procesos de estadística espacial de este trabajo.
- ArcGIS v10.2 (licencia estudiantil): se hace uso principalmente de las herramientas de unión por atributos y unión espacial, además de ser útil para generar las salidas gráficas presentadas y para la obtención de una variable a partir del modelo digital de elevación ASTER GDEM V2 utilizando el módulo de ArcGIS 3D Analyst.

La siguiente figura muestra un esquema que sintetiza la metodología desarrollada en el proyecto.



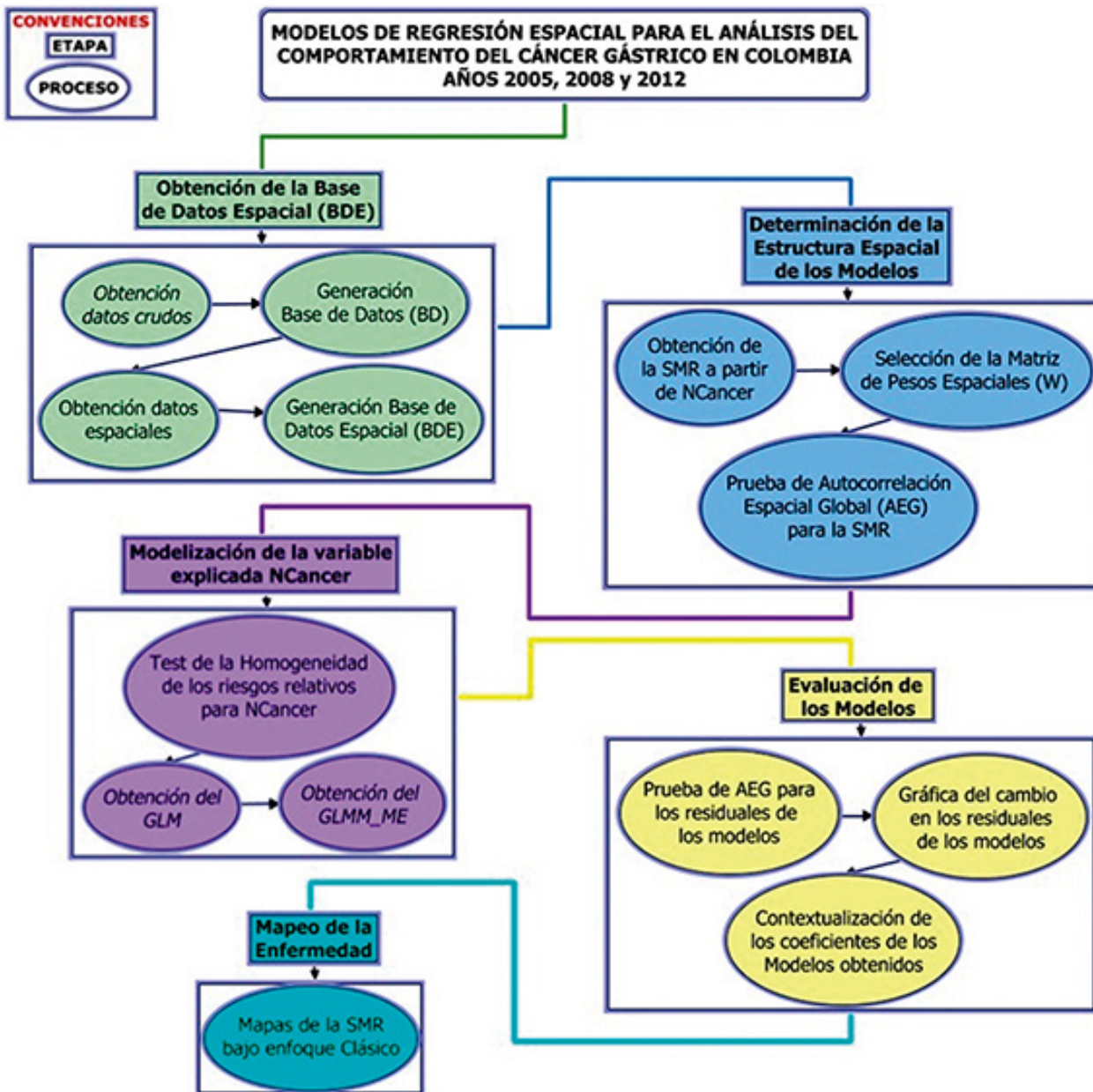


Figura 1. Esquema de la metodología utilizada en el proyecto.

Fuente: elaboración de los autores.



Obtención de la base de datos espacial

El área bajo investigación corresponde a los 1122 municipios de Colombia (figura 2).

Para desarrollar el proyecto, se utilizó una base de datos espacial (BDE) almacenada en un archivo de formas (shapefile). Para crearla, se utilizó información referente al número total de muertes ocurridas por cáncer gástrico durante el periodo 2005-2012 para cada uno de los municipios del país, que fue proporcionada inicialmente (en Excel) por el Departamento de Epidemiología y Demografía del Ministerio de Salud, y se añadió información sobre

otras variables escogidas con base en estudios etiológicos selectos, obteniendo así la base de datos (BD) inicial con la que se formularían los modelos (tabla 1).

Para las variables ubicadas en las celdas de color verde, se escogieron los años 2005, 2008 y 2012, debido a que son representativos durante el periodo de análisis, y principalmente por la disponibilidad de los datos en el Sistema de Información Geográfica para el Ordenamiento Territorial (SIG-OT) y los gestores de bases de datos del DANE y el DNP. Las variables ubicadas en las celdas de color azul no se discriminaron para cada año, porque se calcularon con base en estudios

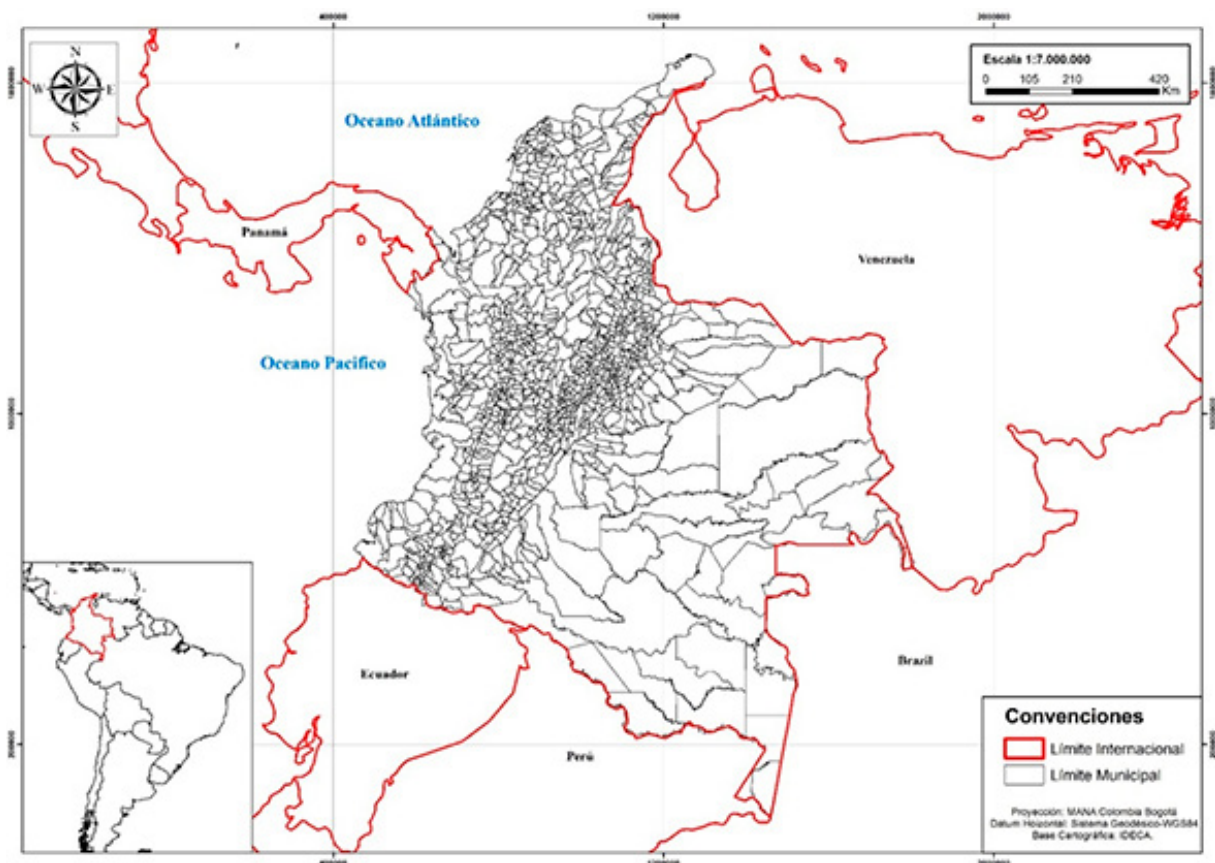


Figura 2. Área de estudio del análisis
Fuente: elaboración de los autores.

Tabla 1. Elementos de la BD

<i>Elemento/Variable</i>	<i>Expresión</i>	<i>Unidad/Fuente</i>
<i>NCancer (explicada)</i>	<i>Número de muertes por causa de cáncer gástrico a nivel municipal.</i>	<i>Número/Ministerio de Salud</i>
<i>NPer</i>	<i>Número total de personas que residen por municipio [urbano/rural].</i>	<i>Número/DANE</i>
<i>IDMun</i>	<i>Índice con escala de 0 a 100 del desempeño de los municipios en un año determinado.</i>	<i>Índice/DNP</i>
<i>PEdu</i>	<i>Porcentaje de la cobertura bruta de educación.</i>	<i>Porcentaje/Ministerio de Educación</i>
<i>IGMun</i>	<i>Índice que mide la capacidad administrativa, financiera y sectorial del municipio.</i>	<i>Porcentaje/DNP</i>
<i>PMun</i>	<i>Posición asignada al municipio dentro del departamento.</i>	<i>Número/DNP</i>
<i>PNbi</i>	<i>Porcentaje de personas que tienen insatisfecha alguna de las necesidades definidas como básicas.</i>	<i>Porcentaje/DANE</i>
<i>NperAfro</i>	<i>Porcentaje de personas afrodescendientes para cada municipio con respecto a la población total.</i>	<i>Porcentaje/DANE</i>
<i>AMsnm</i>	<i>Altura media sobre el nivel del mar para cada municipio del país.</i>	<i>Número/ASTER GDEM V2-NASA- Modulo 3D Analyst de ArcGIS 10.2</i>

DNP: Departamento Nacional de Planeación.

DANE: Departamento Administrativo Nacional de Estadística.

MPS: Ministerio de la Protección Social.

Fuente: elaboración de los autores.

realizados por el DANE entre los años 2000-2011, y no se dispone de información desagregada. La variable ubicada en la celda de color rojo no varía en el tiempo, por esta razón su valor es el mismo en el periodo 2005-2012.

Se ejecutó una metodología especial para calcular la variable AMsnm, debido a que

esta información no estaba disponible para todos los municipios del país. El procedimiento se explica de forma detallada en la figura 3. Las otras variables no requirieron intervención adicional porque al ser descargadas de los gestores de bases de datos del SIG-OT, DANE y DNP ya se encontraban estandarizadas y desagregadas a nivel municipal.



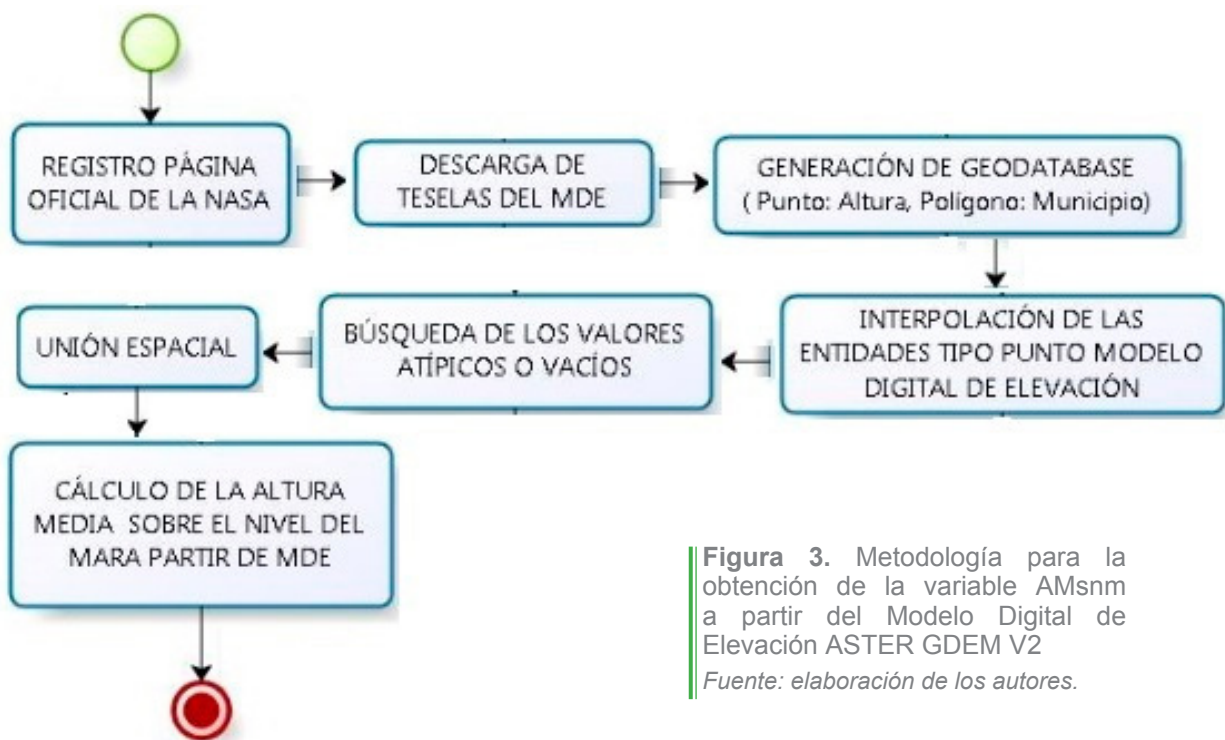


Figura 3. Metodología para la obtención de la variable AMsnm a partir del Modelo Digital de Elevación ASTER GDEM V2
Fuente: elaboración de los autores.

Esa BD se unió a un shapefile de los municipios de Colombia, tomando como atributo identificador el código DANE del municipio, que se encontraba tanto en la BD como en el shapefile, generando así la BDE.

También se utilizaron las coordenadas de los principales centros poblados (en el sistema de proyección local Magna Sirgas, origen Magna Colombia Bogotá) de cada municipio del país. Estas fueron usadas en el cálculo de las matrices de contigüidad para cada año, y a su vez sirvieron para obtener las matrices de pesos espaciales asociadas. Dichas coordenadas fueron obtenidas a partir de información cartográfica digital (shapefile) del Instituto Geográfico Agustín Codazzi (IGAC).

Determinación de la estructura espacial de los modelos

Una vez consolidada la BD, con la

información de NCancer y NPer, se procedió a calcular la tasa de mortalidad estandarizada (SMR, de la sigla en inglés de Standardized Mortality Rate). La SMR es un estadístico muy difundido para representar los patrones espaciales de la distribución de una enfermedad y estandariza los datos, reexpresándolos como la proporción entre el número de casos observados y el número de casos esperados. Se define como (Waller & Gotway, 2004):

$$SMR_i = \frac{NCancer_i}{E_i} \quad (1)$$

Donde: $NCancer_i$ es el número de casos observados en una región i , y E_i el número de casos esperados, i que a su vez se define como:

$$E_i = \frac{\sum NCancer_i NPer_i}{\sum NPer_i} \quad (2)$$

Donde $NPer_i$ es el número de personas con riesgo de contraer la enfermedad en el área i . Este procedimiento se repite para cada año: 2005, 2008 y 2012.

Para definir la mejor matriz de pesos espaciales, fue necesario evaluar distintos tipos de matrices de contigüidad. Se evaluaron en total 10 criterios de contigüidad, a saber: torre (*Rook*), reina (*Queen*), triangulación de Delaunay (*Tri*), esfera de influencia (*Sph*), gráfica de Gabriel (*Gab*), vecinos relativos (*VeR*) y n -vecinos más cercanos con $n=1$ (*knn1*), $n=2$ (*knn2*), $n=3$ (*knn3*) y $n=4$ (*knn4*). Estas se obtienen mediante las funciones *poly2nb*, *tri2nb*, *graph2nb* y *knn2nb*, todas ellas, de la librería *spdep* de R.

La matriz de contigüidad que presentó el menor valor para el coeficiente del criterio de información de Akaike (AIC, de la sigla en inglés de Akaike Information Criterion) fue seleccionada, pues es la que representa mejor la dependencia espacial presente en la variable endógena. Para evaluar el índice AIC en R se hizo uso de la función *test.W* de la librería *spacemakeR*. A continuación se define la matriz de pesos espaciales (W), la cual tiene la siguiente forma (Bivand, Pebesma & Gómez-Rubio, 2008):

$$W = \begin{bmatrix} 0 & \dots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \dots & 0 \end{bmatrix} \quad (3)$$

Donde w_{ij} refleja la intensidad de la dependencia espacial entre pares de regiones; así, las que están más cerca tendrán un peso mayor en el cálculo con respecto a las que están más lejos. La obtención de la matriz de pesos espaciales a través de

la matriz de contigüidad seleccionada es directa mediante la función *nb2listw* de la librería *spdep* de R.

Enseguida, se evaluaron las pruebas de autocorrelación espacial global (AEG) de la I de Moran, y la C de Geary. Para su cálculo se utilizó la librería *spdep* del software R, con las funciones *moran.test* y *geary.test*.

La I de Moran es un coeficiente que mide la similitud de una variable con respecto a áreas que están espacialmente relacionadas. Se denota como (Waller & Gotway, 2004):

$$I = \frac{n \sum_i \sum_j w_{ij} [SMR_i - \overline{SMR}] [SMR_j - \overline{SMR}]}{[\sum_i \sum_j w_{ij}] \sum_k (SMR_{ji} - \overline{SMR}n)^2} \quad (4)$$

Donde:

\overline{SMR} es la media de la SMR en todas las unidades espaciales,

w_{ij} refleja la intensidad de la dependencia espacial entre las regiones i y j , y

$n=1122$ corresponde al total de unidades espaciales.

Una I de Moran de 0 indica que se acepta la hipótesis nula de no agregación, es decir que los datos se distribuyen de forma aleatoria; un valor positivo indica una agregación de áreas cuyos atributos presentan valores similares, y un valor negativo indica que regiones vecinas presentan valores disimilares de un atributo.

La C de Geary es un coeficiente similar a la I de Moran, pero no considera la similitud entre regiones vecinas, sino entre



pares de regiones. Su rango oscila entre 0 y 2, donde cero indica perfecta autocorrelación espacial positiva, y dos, perfecta autocorrelación espacial negativa para cualquier par de regiones. Este coeficiente se representa como (Waller & Gotway, 2004):

$$C = \frac{(n-1) \sum_i \sum_j W_{ij} (SMR_i - SMR_j)^2}{2 \sum_j W_{ij} (SMR_i - \overline{SMR})^2 (\sum_i \sum_j W_{ij})} \quad (5)$$

Modelización de la variable explicada

El primer paso en esta etapa fue evaluar la homogeneidad de los riesgos relativos para NCancer en cada año de estudio, lo cual permite identificar si la mejor distribución estadística para modelar ese fenómeno es la distribución de Poisson o la distribución binomial negativa. La prueba se define como (Bivand, Pebesma & Gómez-Rubio, 2008):

$$x^2 = \sum_{i=1}^n \frac{[NCancer_i - \theta E_i]^2}{\theta E_i} \quad (6)$$

Donde θ es la SMR global que asintóticamente sigue una distribución chi cuadrado con n grados de libertad.

La prueba considera como hipótesis alternativa que no todos los riesgos relativos son iguales, y en ese caso el mejor modelo de regresión espacial es uno de tipo binomial negativo. En R, se hace uso de la función `achisq.test` de la librería `DCluster` para evaluar esa prueba.

Enseguida se estimaron dos tipos de modelos de regresión. En el primero se evaluó un modelo lineal generalizado (GLM, de la sigla en inglés de Generalized Linear Model) sin estructura espacial, el cual se define como (Dunteman & Ho, 2006):

$$g(\mu_i) = \beta_0 + \beta_{x_{1i}} + \beta_{x_{2i}} + \dots + \beta_{x_{mi}} \quad (7)$$

Donde:

el término μ_i es la variable dependiente bajo análisis,

x representa las covariables explicativas almacenadas en una matriz de tamaño ($m \times i = 7 \times 1122$) y

g(.) es la función de enlace que transforma la variable respuesta en un predictor lineal.

El término μ_i sigue una distribución de Poisson cuando representa el número de eventos que ocurren en un intervalo temporal o espacial de tamaño dado, tal como NCancer. Esto se denota como (Waller & Gotway, 2004):

$$p[NCancer; \mu] = \begin{cases} \frac{e^{-\mu} \mu^y}{NCancer!}, & \text{para } NCancer = 0, 1, 2, \dots; \mu > 0 \\ 0, & \text{en cualquier otro caso} \end{cases} \quad (8)$$

Donde $\mu = E\theta$, corresponde a la media de casos en cada región para el conteo de eventos regionales $NCancer_1, \dots, NCancer_i$ independientes, e idénticamente distribuidos como variables aleatorias de Poisson, con media y varianza igual $E[NCancer_i]$. El valor esperado se puede modelar con una regresión de Poisson en función de covariables regionales como (Waller & Gotway, 2004):



$$\log[E(NCancer_i)] = NCancer_i \beta_i, E(NCancer_i) = \exp(NCancer_i \beta_i) \quad (9)$$

Cuando $NCancer_i$ presenta sobredispersión, de modo que la variación de los datos sea mayor a su media, no se recomienda utilizar una distribución de Poisson en las variables, y el modelo de regresión lineal debe ser ampliado para que se permita una mayor varianza. Esto se logra asumiendo que los datos sigan una distribución binomial negativa, que puede ser considerada como un modelo mixto que involucra un efecto aleatorio θ , que sigue una distribución gamma. Esta distribución se conoce con el nombre de Poisson-Gamma (PG), y se estructura en dos niveles (Bivand, Pebesma & Gómez-Rubio, 2008):

$$NCancer_i | \theta E_i \sim Poisson(NCancer_i E_i) \quad (10)$$

$$\theta \sim Gamma(v, \alpha) \quad (11)$$

El riesgo relativo θ corresponde a una variable aleatoria que se extrae de una distribución gamma con media v/α y varianza v/α^2 . Nótese que en la ecuación 10 la distribución $NCancer_i$ está condicionada por el valor de θ . La distribución condicionada de $NCancer_i$ es una binomial negativa con probabilidad $\frac{\alpha}{\alpha+E}$ y tamaño v .

Es necesario evaluar un segundo tipo de modelo debido a que los GLM tan solo son adecuados si se asumen los supuestos de que los datos observados son independientes, y de que la variación espacial observada en los resultados del modelo es explicada por las covariables (Dunteman & Ho, 2006), algo que no ocurre en los fenómenos reales, dada su complejidad.

La forma de cumplir con los supuestos es utilizando un modelo lineal generalizado mixto (GLMM, de la sigla en inglés de Generalized Linear Mixed Model) (Kaiser, Cressie & Lee, 2002), en el que se establece que la distribución de cada resultado depende de procesos espaciales no observados, que explican los patrones espaciales entre eventos.

Acá entran en juego los vectores propios de Moran (ME, de la sigla en inglés de Moran Eigenvector), que se obtienen a partir de las matrices W , y cuyo objetivo es ser añadidos a un GLM y, de esa manera, provocar que la dependencia espacial presente en los residuales del modelo sin estructura espacial pueda ser movida y tenida en cuenta dentro de dicho modelo, obteniendo así un GLMM_ME, lo que significa que estos vectores pueden ser usados como variables representativas de agentes demográficos y causales, para así lidiar con la autocorrelación y la multicolinealidad (Voutilainen *et al.*, 2014). Este método utiliza fuerza bruta para buscar el conjunto de ME de la matriz W que se define como (Dray, Legendre & Peres-Neto, 2006):

$$ME = MWM \quad (12)$$

Donde:

M es una matriz proyección simétrica e idempotente (es decir que la matriz es igual a su cuadrado) y W son los pesos espaciales. Además M se define como:



$$M = I - X[X^T X]^{-1} X^T \quad (13)$$

Donde:

I: es la matriz identidad, y

X: es un vector de unos de tamaño n, que funciona solo como un intercepto.

Los ME que son incluidos se escogen calculando los valores empíricos de la I de Moran del modelo de regresión inicial más cada uno de los ME de las matrices de ponderaciones espaciales simétricas doblemente centradas. El primer ME se escoge como aquel que posee el menor valor de la I de Moran. El procedimiento se repite hasta que el menor valor de la I de Moran remanente tiene un valor de probabilidad basado en permutaciones, arriba de un valor α establecido. Al final se escoge el subconjunto de n ME que reducen el residual de autocorrelación espacial en el error del modelo con covariables (Dray, Legendre & Peres-Neto, 2006).

Para esta etapa, las principales funciones a utilizar son `glm` o `glm.nb` de la librería MASS y ME de la librería `spdep`, todas ellas, del software R.

Evaluación de los modelos

En esta etapa se utilizaron principalmente los coeficientes AIC de cada modelo obtenido en la etapa anterior, los gráficos de diferencia entre los residuales para cada uno de los años, el análisis de los residuales mediante pruebas de autocorrelación espacial para analizar cómo fue tenida en cuenta la dependencia espacial en cada tipo de modelo y año, y se interpretan los coeficientes de los modelos obtenidos para evaluar si sus valores concuerdan con la información empírica que se posee de la

enfermedad, prestando especial atención al signo de los coeficientes, y si los valores son similares en cada uno de los tres años de estudio. La principal función de R utilizada en esta etapa fue `residuals` de la librería estándar.

Mapeo de la enfermedad

Por último, fueron creados mapas de la enfermedad obtenidos a partir de los valores calculados en los pasos anteriores. Estos mapas son de gran importancia, pues representan de forma visual la distribución de la enfermedad y permiten identificar patrones y agregaciones espaciales. Además, son un insumo importante para estratificar la zona objeto de estudio de acuerdo con el nivel de riesgo de prevalencia de la enfermedad.

Se utilizó un mapa de la SMR bajo enfoque clásico, el cual se usó cuando se rechazó el supuesto de que los datos seguían una distribución de Poisson. Para probar esto se estimó la dispersión de la variable del conteo de los casos observados almacenados en la variable `NCancer`, por el ajuste del GLM, incluyendo solo el intercepto y la población en riesgo almacenada en la variable `NPer`. En los casos en que se presentó sobredispersión en los datos se rechazó la hipótesis de que los datos seguían una distribución de Poisson y, por lo tanto, se asumió una distribución binomial negativa que permitía tratar varianzas grandes.

Este mapa se grafica utilizando la función `empbaysmooth`, de la librería `DCluster` del software R.



Resultados y discusión

Determinación de la estructura espacial de los modelos

Tabla 2. Estadísticas básicas de las variables SMR y NCancer para cada año de estudio

Año	SMR						NCancer					
	Mín.	1st Qu.	Median	Mean	3rd Qu.	Máx.	Mín.	1st Qu.	Median	Mean	3rd Qu.	Máx.
2005	0	0	0,6781	1,108	1,677	14,05	0	0	1	4,03	2,75	746
2008	0	0	0,6804	1,047	1,580	17,60	0	0	1	4,03	3,00	793
2012	0	0	0,6013	1,029	1,512	16,94	0	0	1	4,14	2,00	854

Fuente: elaboración de los autores.

La importancia de utilizar la SMR para analizar la estructura de dependencia espacial de NCancer para cada año, frente a utilizar sus datos crudos, se observa de forma clara en la tabla 3, donde la variabilidad de los datos se reduce drásticamente después de realizar una adecuada estandarización, lo que permite analizar sin ruido características determinantes en la incidencia de la enfermedad.

En la tabla 4 se presentan los diez criterios de contigüidad utilizados, y se resaltan los que presentaron el menor valor de índice AIC para cada año.

Se destaca el hecho de que la matriz de contigüidad seleccionada para el año 2005 no sea la misma para los años 2008 y 2012. Analizando con detalle los datos de

la BDE, se deduce que eso se debe a que para el 2005 existían cuatro municipios que hacían parte, como corregimientos, de otros municipios de mayor extensión y, por lo tanto, los registros de población y casos de muerte por cáncer gástrico en esos municipios tenían valor de cero.

Con las matrices de contigüidad seleccionadas, a continuación se pasó a definir la matriz de pesos espaciales asociada, en estilo binario, para cada año, que fue la utilizada en los cálculos posteriores.

Se procedió entonces a evaluar si los casos de cáncer gástrico de regiones vecinas eran similares, lo que podría significar la influencia de regiones vecinas entre sí. Esto se efectuó mediante los estadísticos y sus p-valores que se muestran en la tabla 4.

Tabla 3. Valor coeficiente AIC para cada criterio de contigüidad

Año	Rook	Queen	Tri	Sph	Gab	VeR	knn1	knn2	knn3	knn4
2005	---	---	-75.15	-659.45	-229.67	-172.05	159.98	-631.37	-615.89	-460.25
2008	---	---	-232.09	-573.07	-235.09	-322.59	-88.20	-705.67	-611.69	-428.99
2012	---	---	-243.89	-514.65	-242.61	-173.77	136.34	-847.26	-481.25	-454.71

Fuente: elaboración de los autores.



Tabla 4. Pruebas de AEG para la SMR

Año	I de Moran		C de Geary	
	Estadístico	p-valor	Estadístico	p-valor
2005	0,2487	2,20E-16	0,7379	1,38E-05
2008	0,2209	2,463e-16	0,7899	2,48E-02
2012	0,1964	3,52E-10	0,8182	6,90E-02

Fuente: elaboración de los autores.

Para el test de I de Moran la prueba es significativa en cada uno de los 3 años, comprobándose la existencia de una estructura espacial en la distribución de NCancer, además de que, al ser positivo el estadístico, sugiere que en la distribución espacial de casos los municipios con valores altos de SMR estaban rodeados de municipios que presentaban ese mismo comportamiento, y los municipios con tasas muy bajas de SMR tenían como vecinos municipios con valores bajos.

El test de disimilitud de Geary, igualmente sugiere que, para los tres años, los pares de valores vecinos son similares, y de hecho lo son bastante, si se tiene en cuenta el alto valor obtenido por ese estadístico para cada año.

Modelización de la variable explicada

En seguida, se evaluó si habían diferencias entre los diferentes riesgos, lo cual se

hizo mediante las pruebas mostradas en la tabla 5.

Se observa que en la columna de evaluación se obtuvieron valores negativos (lo que implica que el valor del estadístico de contraste es mayor al valor de la distribución de prueba para cada caso), por lo que se concluye que no existe homogeneidad de los riesgos relativos para los valores de NCancer en cada año. Eso implica la existencia de una sobredispersión en los datos, de tal manera que un modelo con la distribución Poisson deja de ser apropiado para explicar a NCancer.

Teniendo en cuenta lo anterior, desde acá se empieza a analizar el fenómeno trabajando con un GLM sin estructura espacial (NCancerNB_GLM) y un GLMM_ME con estructura espacial al incluirle los vectores propios de Moran (NCancerNB_GLMM), cada uno siguiendo una distribución binomial negativa para cada año (2005, 2008 y 2012), tal como se muestra en las tablas 6, 7 y 8.

Tabla 5. Pruebas de la homogeneidad de los riesgos relativos

Año	Prueba	Estadístico de contraste	Evaluación estadístico-prueba
	Chi cuadrado	chtest	Chi cuadrado - chtest
2005	1044,270065	2736,025	-1691,754935
2008	1044,270065	2450,255	-1405,984935
2012	1044,270065	2602,564	-1558,293935

Fuente: elaboración de los autores.



Tabla 6. Coeficientes y p-valor de las covariables, modelos año 2005

Modelo	NCancer05NB_GLM		NCancer05NB_GLMM	
AIC	3683.7		3606.9	
Variables	p-valor	Estimado	p-valor	Estimado
Intercept	2e-16	-9.641e+00	2e-16	-9.705e+00
NperAfro05	4.97e-05	-8.315e-03	1.32e-06	-9.663e-03
PNbi05	2.34e-08	-1.064e-02	1.09e-11	-1.219e-02
PMun05	4.59e-06	5.048e-03	4.05e-09	6.337e-03
AMsnm	8.96e-15	2.759e-04	3.04e-10	2.186e-04
ME40	7.07e-05	2.792e+00
ME5	1.04e-06	5.456e+00
ME3	0.001224	4.963e+00
ME485	0.003344	2.832e+00
ME4	0.000655	-4.160e+00
ME31	6.09e-06	3.818e+00
ME19	0.000281	3.673e+00

Fuente: elaboración de los autores.

Tabla 7. Coeficientes y p-valor de las covariables, modelos año 2008

Modelo	NCancer08NB_GLM		NCancer08NB_GLMM	
AIC	3573.9		3511.1	
Variables	p-valor	Estimado	p-valor	Estimado
Intercept	2e-16	-1.030e+01	2e-16	-1.033e+01
NperAfro08	0.000105	-7.290e-03	4.41e-05	-7.429e-03
PNbi08	0.000549	-5.830e-03	1.14e-05	-7.029e-03
IDMun08	2.76e-05	1.149e-02	5.42e-06	1.161e-02
PEdu08	0.011670	3.676e-03	0.011719	3.516e-03
PMun08	0.036173	2.047e-03	0.017873	2.189e-03
AMsnm	2e-16	2.774e-04	2e-16	2.745e-04
ME242	1.708e+00	0.030533
ME9	-1.712e+00	0.001530
ME53	2.222e+00	0.000363
ME33	3.190e+00	3.97e-05
ME13	2.010e+00	0.000798
ME311	1.725e+00	0.022017
ME14	3.338e+00	2.31e-05
ME97	-3.153e+00	0.005093

Fuente: elaboración de los autores.



Tabla 8. Coeficientes y p-valor de las covariables, modelos año 2012

Modelo	NCancer12NB_GLM		NCancer12NB_GLMM	
AIC	3566.2		3539.8	
Variables	p-valor	Estimado	p-valor	Estimado
Intercept	2e-16	-1.047e+01	3.324e-01	-1.061e+01
NperAfro12	0.001970	-5.860e-03	1.820e-03	-5.913e-03
PNbi12	5.61e-07	-9.158e-03	1.770e-03	-9.412e-03
IDMun12	0.000104	1.024e-02	2.544e-03	1.146e-02
PEdu12	0.000193	5.282e-03	1.380e-03	5.417e-03
PMun12	0.027756	2.451e-03	No significativa	No significativa
AMsnm	1.42e-14	2.711e-04	3.435e-05	2.833e-04
ME242	0.016803	1.954e+00
ME9	0.000759	-2.082e+00
ME53	6.69e-05	2.753e+00
ME32	0.002200	-2.247e+00
ME748	0.019114	2.055e+00

Fuente: elaboración de los autores.

Las variables resaltadas en azul fueron significativas tanto en el modelo sin estructura espacial como en el que la tiene; sin embargo, no coinciden en los tres años. Aquellas variables resaltadas en verde fueron significativas tanto en el modelo sin estructura espacial como en el que la tiene, y además aparecen en los modelos para los tres años analizados. Se infiere que son las que más influyen en el comportamiento de la variable endógena NCancer.

Evaluación de los modelos

Se aprecia que el valor del coeficiente AIC disminuyó en los modelos que presentaban estructura espacial, lo que implica que la inclusión del componente espacial es fundamental para explicar el fenómeno del cáncer gástrico en el país.

Se observa además que para cada año resultó un número diferente de vectores propios, a pesar de que en el 2008 y el 2012 se utilizó la misma matriz de pesos espaciales. Es importante analizar con detalle esos ME, y una manera fácil y rápida de evaluar sus efectos en la absorción de la dependencia espacial. En la tabla 9 y la figura 4 se muestra la diferencia entre los residuales del modelo con estructura espacial con respecto a los residuales del modelo sin estructura espacial.

Tabla 9. Prueba de AEG de los residuales de los modelos analizados

Año	NCancerNB_GLM		NCancerNB_GLMM	
	Estadístico I de Moran	p-valor	Estadístico I de Moran	p-valor
2005	0,1625	2,20E-16	0,0895	4,89E-05
2008	0,2068	3,19E-14	0,1516	2,50E-08
2012	0,1605	3,69E-09	0,1343	7,74E-07

Fuente: elaboración de los autores.



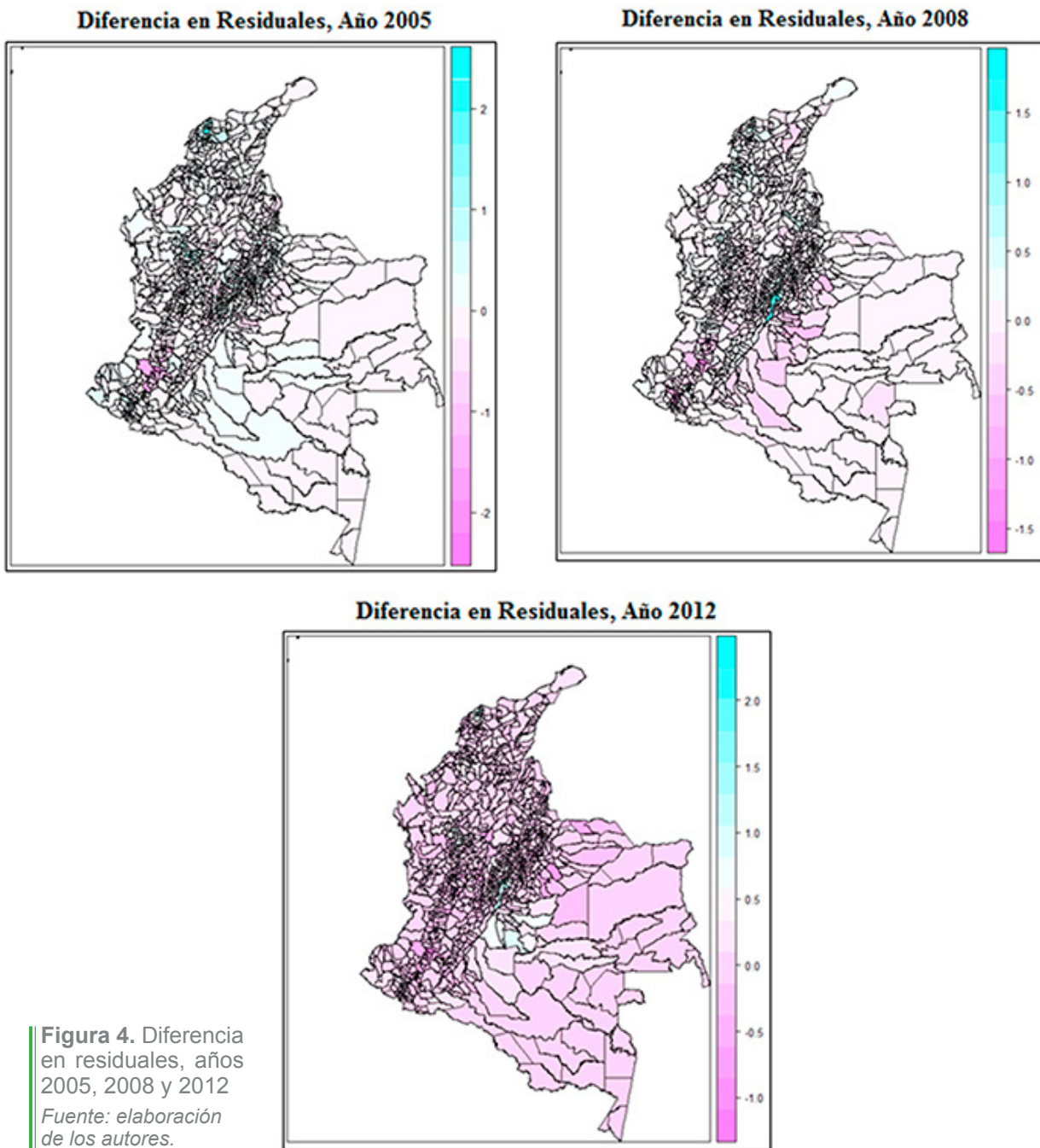


Figura 4. Diferencia en residuales, años 2005, 2008 y 2012
 Fuente: elaboración de los autores.

De las gráficas de diferencia de residuales entre modelos (sin y con estructura espacial) para cada año analizado, se destaca la reiterada aparición de municipios del departamento del Cauca y Tolima en los casos de agrupamiento de disminución de residuales, y la aparición de municipios de

Antioquia y Cundinamarca en los casos de agrupamiento de aumento. Todos ellos destacan por el hecho de que los vectores propios de Moran, utilizados como variables representativas, poseían allí valores sobresalientes que provocaban una estructura de dependencia espacial.



Se aprecia que las variables que fueron significativas para los tres años analizados poseen valores similares en sus coeficientes, lo cual ayuda a confirmar que los signos y valores de dichas variables son los apropiados.

Es así como se puede afirmar que la variable PNbi tiene una relación inversa con NCancer para los tres años, algo que fue un poco fuera de lo común si se tiene en cuenta cuál es la cantidad que cuantifica el NBI, habiéndose esperado un valor de coeficiente positivo.

Otra variable importante fue la AMsnm, que además obtuvo un coeficiente positivo, tal como se esperaba, de acuerdo con los patrones de distribución del cáncer en el país.

Finalmente, analizando la variable NperAfro, se deduce la importancia de las variaciones genéticas en la población en cuanto a la explicación del cáncer gástrico, pues, al ser negativo su coeficiente, indica que existen menos probabilidades de morir por cáncer gástrico si se pertenece al grupo étnico de afrodescendientes.

Mapeo de enfermedades

En la figura 5 se presentan los mapas de la SMR desde el enfoque clásico para los años 2005, 2008 y 2012. De forma global, estos mapas muestran un incremento en la SMR para el periodo 2005-2012, con tasas más altas, y la inclusión de unidades espaciales en los tonos rojizos, que indican alta prevalencia de la enfermedad. Para el 2012, se observa un incremento significativo de riesgo en los municipios de Paz de Ariporo, Cravo Norte, Fortul y

Chiscas, ubicados en el departamento de Arauca y en el norte de Casanare, y además en Potosí, Ipiales y Cumbal, ubicados en el departamento de Nariño,

Para los tres años, se observa un patrón muy diferenciado de alto riesgo de mortalidad por cáncer gástrico en la región andina, especialmente en los municipios aledaños a centros urbanos como Bogotá, Medellín, Cali, Cúcuta e Ibagué. De otro lado, las regiones de la Costa Atlántica, Caribe y Amazonía presentaron bajo riesgo de mortalidad.

Conclusiones

El comportamiento espacial de una enfermedad como el cáncer gástrico no es un fenómeno fácil de estudiar debido a las múltiples limitaciones que tiene, dentro de las que se cuentan los insuficientes registros de casos de ese tipo de enfermedad, los escasos registros de las variables que pueden llegar a explicar ese fenómeno o la dificultad de estudiar esa enfermedad de forma indirecta, como fue el caso del presente estudio.

Sin embargo, los resultados obtenidos con los modelos de regresión se consideran satisfactorios, ya que el valor de dos de los coeficientes de las variables que fueron significativas están en consonancia con las suposiciones previamente formuladas respecto a las posibles causas de la enfermedad. Con ellos es posible respaldar las afirmaciones de que la incidencia del cáncer gástrico está fuertemente relacionada con factores genéticos de la población y con factores físicos del entorno natural, como la altura sobre el nivel del mar.



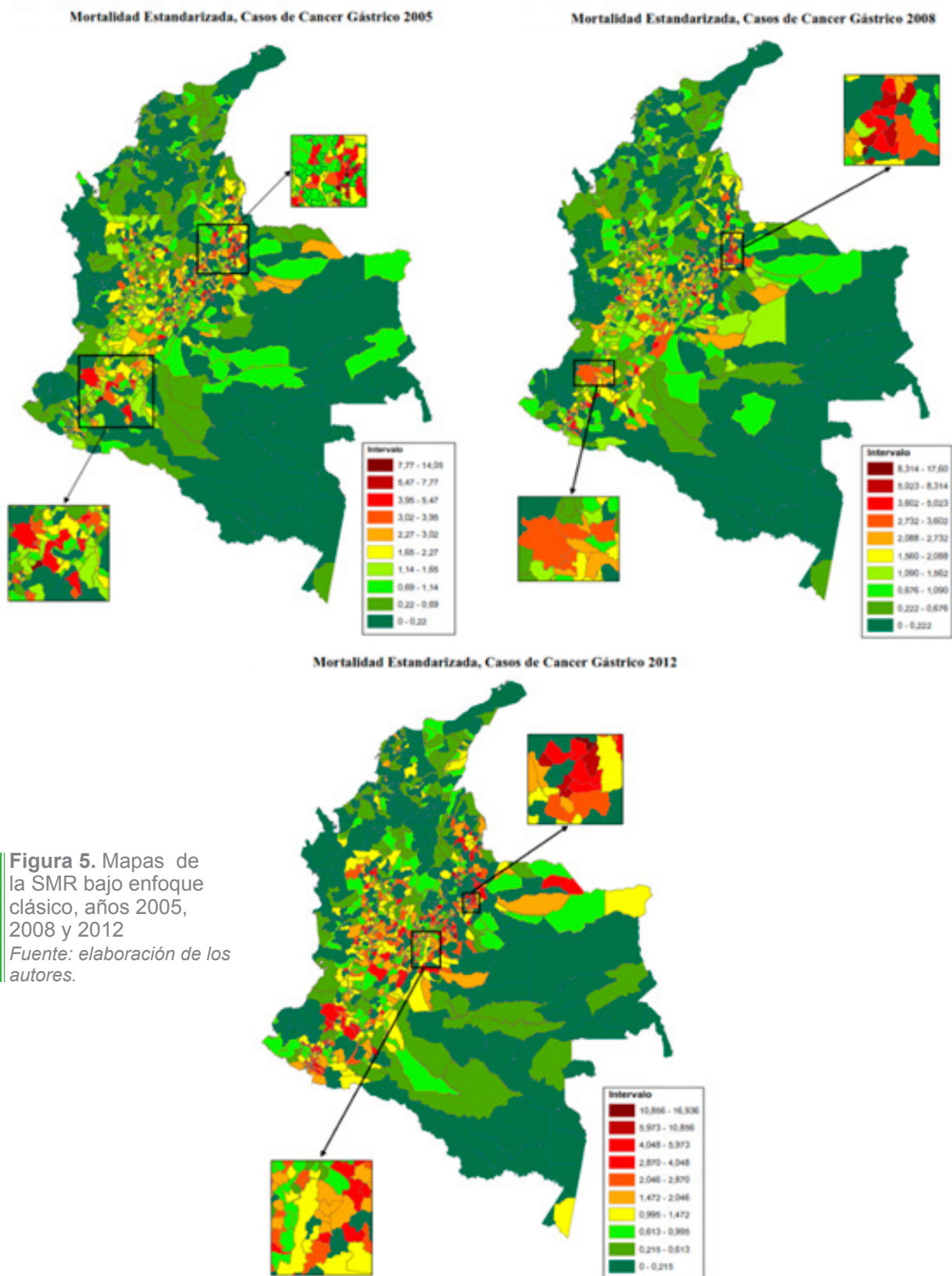


Figura 5. Mapas de la SMR bajo enfoque clásico, años 2005, 2008 y 2012
 Fuente: elaboración de los autores.



Es de especial importancia el hallazgo de la relación positiva entre la altura y un mayor número de casos de cáncer gástrico en el país para los tres años de estudio, pues el hecho de que sea un factor que no había sido tenido en cuenta de forma medible por ningún autor en estudios previos permite dar nuevas pistas y variables medibles para cuantificar qué tanto afecta la ubicación geográfica de una persona para la prevalencia de una enfermedad crónica como el cáncer gástrico.

Ahondando en el análisis de los modelos estimados, se observó también que el uso de vectores propios de Moran puede actuar como un método complementario adecuado para el modelado de causas etiológicas, pues, tal como se evidenció, proporcionó un medio para mejorar

la comprensión de la estructura espacial de la variable NCancer y las variables explicativas. De esa manera, también fue útil para evaluar la importancia espacial de factores explicativos no considerados, ya fuese porque eran difíciles de medir o aún se desconocen, sobre todo en aquellos municipios con agrupamientos de disminución (color magenta) o aumento (color cian) de residuales de los modelos.

Definitivamente, la conveniencia del análisis se evidenció mejor a través de la visualización espacial de los datos y resultados por medio de los programas estadísticos y los sistemas de información geográfica, pues permitieron identificar con claridad patrones y agregaciones espaciales como los que se observan en la figura 5.



Bibliografía

- Bivand, R., Pebesma, E. & Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. New York: Springer.
- Correa, P. (2011). Gastric cancer: an infectious disease. *Revista Colombiana de Cirugía*, 26(2), 111-117.
- Daza, D. (2012). *Cáncer gástrico en Colombia entre 2000 y 2009*. Bogotá: Universidad del Rosario. Recuperado de <http://repository.urosario.edu.co/handle/10336/4004>.
- Dray, S., Legendre, P. & Peres-Neto, P. (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling*, 196(3-4), 483-493. <https://doi.org/10.1016/j.ecolmodel.2006.02.015>.
- Dunteman, G. & Ho, M. (2006). *An Introduction to Generalized Linear Models*. London: SAGE Publications.
- Kaiser, M., Cressie, N. & Lee, J. (2002). Spatial mixture models based on exponential family conditional distributions. *Statistica Sinica*, 12(2), 449-474. Recuperado de <http://ro.uow.edu.au/cgi/viewcontent.cgi?article=9826&context=in-fopapers>.
- Piñeros, M., Pardo, C., Gamboa, O. & Hernández, G. (2010). *Atlas de mortalidad por cáncer en Colombia*. Bogotá: Instituto Nacional de Cancerología; Instituto Geográfico Agustín Codazzi.
- Strebel, K., Rolle-Kampczyk, U., Richter, M., Kindler, A., Richter, T. & Schlink, U. (2010). A rigorous small area modelling-study for the Helicobacter pylori epidemiology. *Science of the Total Environment*, 408(18), 3931-3942. DOI: 10.1016/j.scitotenv.2010.03.045.
- Voutilainen, A., Tolppanen, A., Vehviläinen-Julkunen, K. & Sherwood, P. (2014). From spatial ecology to spatial epidemiology: modeling spatial distributions of different cancer types with principal coordinates of neighbor matrices. *Emerging Themes in Epidemiology*, 11(11), 1-10. doi: 10.1186/1742-7622-11-11.
- Waller, L. & Gotway, C. (2004). *Applied Spatial Statistics for Public Health Data*. New Jersey: John Wiley & Sons.

