
Análisis del número de pacientes fallecidos con cáncer de mama y próstata en la población colombiana por departamento en el periodo (2006 -2015) mediante un modelo Poisson espacio-temporal bayesiano

Analysis of the number of deceased patients with breast and prostate cancer in the Colombian population by department in the period (1998 -2014) using a Bayesian spatio-temporal Poisson model.

María Fernanda Hernández.^a
mariafernandez@usantotomas.edu.co

Dagoberto Bermúdez Rubio.^b
dagobertobermudez@usantotomas.edu.co

Resumen

El cáncer en Colombia, como en el mundo es una problemática latente en toda la población por su gran aumento en los últimos diez años, el término cáncer engloba a un grupo numeroso de enfermedades que se caracteriza por el número de células anormales que se dividen y se crecen sin control en cualquier parte del cuerpo. Comienza, cuando las células crecen descontroladamente sobrepasando a las células normales, lo que dificulta que el cuerpo funcione de la manera adecuada. Según el (INC, 2013) adscrito al Ministerio de Salud, hay más de 100 tipos de cáncer, la mayoría de ellos toman el nombre del órgano, de los tejidos o las células en donde se reproduce. Para el caso puntual de esta investigación, se estudian datos obtenidos del Sistema Nacional de Información del Cáncer en Colombia, donde se tiene el número de pacientes que fallecen por cáncer de mama o de próstata para los años (2006 al 2015) por departamento. Se realizará la estimación de modelos lineales generalizados Poisson bayesianos con enfoque espacio temporal. La estimación de los modelos proporciona resultados importantes en la identificación de las muertes a causa del cáncer de mama y próstata en cada uno de los lugares de estudio, para realizar aportes significativos en las estrategias para la prevención y tratamiento de esta enfermedad y en la implementación de las políticas públicas y de salud del país.

Palabras clave: cáncer de mama, cáncer de próstata, modelo lineal generalizado, modelo espacio temporal, estadística bayesiana.

Abstract

Cancer in Colombia, as in the world is a problem latent throughout the population due to its great increase in the last ten years, the term cancer encompasses a large group of diseases characterized by the number of abnormal cells that divide and They grow without control in any part of the body. It begins when cells grow uncontrollably surpassing normal cells, making it difficult for the body to function properly. According to the (INC, 2013) there are more than 100 types of cancer, most of them take the name of the organ, tissues or cells where it reproduces. For the specific case of this research, data obtained from the National Cancer Information System in Colombia are studied, where the number of patients dying from breast or prostate cancer for the years (2006-2016) by department is available. We will use Bayesian Poisson generalized linear models with temporal space focus. The estimation of the models

^aEstudiante de estadística Universidad Santo Tomás Bogotá

^bEstadístico.Docente Facultad de estadística Universidad Santo Tomás Bogotá

provides important results in the identification of deaths due to breast and prostate cancer in each of the study sites, to make significant contributions in the strategies for the prevention and treatment of this disease and in the implementation of the country's public and health policies.

Keywords: breast cancer, prostate cancer, generalized linear model, spatiotemporal model, Bayesian statistics.

1. Introducción

La presente investigación, trata el tema de la problemática de las muertes de mujeres y hombres por cáncer de mama y próstata, el cáncer es una de las enfermedades por la cual se presentan más muertes en Colombia según el Ministerio de Salud y Protección Social, de las personas que mueren al año por cáncer 16.800 son mujeres y 16.300 son hombres. En Colombia se invierten aproximadamente 9.000 millones de pesos anuales en campañas para prevenir esta enfermedad, el tipo de cáncer de mama y próstata se encuentran entre los cinco que más afectan a los hombre y mujeres colombianos.

Se encuentran diversos análisis acerca del cáncer en la actualidad, en gran parte son exploratorios, o donde solo se miden las enfermedades en el tiempo o el espacio pero en este caso este estudio radica en realizar un análisis espacio temporal por departamento donde se quiere analizar que departamentos son los más afectados por esta enfermedad, teniendo encuentra algunas variables explicativas.

Los modelos espacio temporales Poisson bayesianos en específico el Modelo Besag York Mollié (BYM) estiman el riesgo de las enfermedades, en esta investigación se trabajará con el cáncer de mama y cáncer de próstata que como se ha mencionado en anteriores ocasiones estan dentro de los cinco tipos de cáncer por los que mas mueren personas en el país.

Este trabajo de grado se encuentra organizado de la siguiente manera: en la primera sección la introducción al tema de investigación, en la segunda sección los antecedentes de los trabajos realizados en está problemática, en la tercera sección se encuentra el marco conceptual que consta de las definiciones de modelos espaciales, modelos espacio temporales, cáncer, cáncer de mama y cáncer de prostata; el la cuarta sección se muestra el marco teorico de la investigación el cual esta compuesto por el modelo de regresión Poisson, el modelo de regresión Poisson bayesiano, los modelos condicionales autoregresivos (CAR), los modelos ICAR y el modelo BYM que es el enfoque con el que se va a realizar en la presente investigación el análisis, el modelo Poisson espacial y el modelo Poisson espacio temporal, en la sección cinco se muestran los resultados de los descriptivos de los datos que se van a trabajar, estos descriptivos constan de dos parte la primera es acerca de la visualización de los mapas de las enfermedades y en segunda estancia las pruebas de correlación de la variable respuesta, seguido de la implementación de los modelos y sus respectivos análisis, por ultimo en la sección número seis se encuentran las conclusiones de la investigación y trabajos futuros y al final las referencias bibliograficas.

Con esta investigación, se pretende encontrar los departamentos del país donde se presentan más muertes por cáncer de mama y próstata mediante una metodología estadística como lo son los modelos espacio temporales, con este análisis se pretenden dar indicios de donde se ecuentran más afectaciones en la salud pública y donde se requieren de más recursos de las políticas públicas del país.

2. Antecedentes

Durante la última década en Colombia, las investigaciones sobre el cáncer han sido variadas, en estos estudios se utilizaron diversas metodologías estadísticas y, se encontraron múltiples resultados del cáncer de mama y de próstata. Estos dos son los tipos de cáncer que más atacan a hombres y mujeres en Colombia según el Ministerio de Salud y Protección Social.

En el 2012 en la Universidad Industrial de Santander se realizó un estudio titulado Supervivencia a cinco años de las mujeres con cáncer de mama de Bucaramanga y su área metropolitana, según el estado clínico para ello se usó una cohorte cerrada retrospectiva con 719 casos incidentes del Registro Poblacional de Cáncer en la Área Metropolitana de Bucaramanga (RPC-AMB) entre 2001 y 2004, se estimó la supervivencia a cinco años y su asociación con el estado clínico al momento del diagnóstico haciendo uso de un modelo de regresión de Cox, ajustado por factores pronóstico y estratificado por tipo de régimen al sistema de seguridad social (Osma et al., 2012). Durante este mismo periodo el estudio realizado por (Batista et al., 2012) titulado Análisis espacio-temporal de eventos asociados al cáncer, una herramienta para apoyar estudios epidemiológicos, donde se dio uso de una metodología espacio temporal para mostrar e identificar conglomerados espaciales y/o espacio-temporales que permitan el análisis de la morbilidad de las cinco localizaciones de cáncer en este estudio se encontró la presencia de conglomerados significativos, de las áreas de mayor riesgo de enfermar y morir por cáncer y respecto a esta metodología se demostró ser una buena herramienta para el análisis de la morbilidad del cáncer y permitió la generación de hipótesis sobre posibles determinantes que ayuden a orientar nuevas investigaciones.

En el Instituto Nacional de Cancerología (INC), se realizó una investigación para caracterizar epidemiológicamente el cáncer, así como otros factores que inciden en su letalidad; la investigación Plan Control del cáncer en Colombia 2012-2021. Un análisis formal realizado por (Vergara et al., 2017), con una metodología descriptiva y el modelo para el control del cáncer se encuentra que con un sistema estricto de monitoreo, una evaluación permanente de las actividades del plan y la aplicación de correctivos oportunos, Colombia podría alcanzar las metas en el control del cáncer.

En 2014, (Restrepo et al., 2014) realizaron un estudio titulado Incidencia, mortalidad y supervivencia al cáncer de próstata (CP) en Cali, Colombia, 1962-2011, con base en el Registro Poblacional de Cáncer en Cali (Colombia) y los registros de mortalidad de la Secretaría de Salud Pública Municipal se analizaron las tendencias de las tasas de incidencia, mortalidad y supervivencia por CP en Cali durante el periodo 1962-2011 los métodos usados para esta investigación fueron modelos de regresión joinpoint y se utilizó un modelo de riesgos proporcionales de Cox donde se concluyó que el cáncer de próstata en Cali presenta mejoría en cuanto a la supervivencia y se presenta una disminución en su mortalidad, esto está asociado significativamente a edad, periodo de diagnóstico y estrato socioeconómico. Igualmente, para el 2015, el (INC, 2015) realizó un análisis de la situación del cáncer en Colombia, donde se hizo uso de la base de datos de este Instituto a las cuales se les dio uso para realizar modelos descriptivos para ver el cuidado paliativo, las incidencias, los tratamientos de los pacientes tratados en el INC hasta el año 2015.

En el 2016, en el Análisis de conglomerados espacio-temporales de incidencia del cáncer pediátrico en la provincia de Córdoba, Argentina (2004-2013) realizado (Agost, 2016) haciendo uso de análisis espaciales y temporales exploratorios a partir de la base de datos del Registro de Tumores de la Provincia de Córdoba (años 2004-2013) para determinar agrupamientos (conglomerados) de incidencia de cáncer infantil en la provincia de Córdoba (Argentina), donde se evidenció que se presentaron conglomerados significativos con importantes indicadores asociados en numerosos departamentos de la provincia.

En el 2017, en el INC se realizó un informe titulado, Atlas de mortalidad por cáncer en Colombia, en su cuarta versión donde se hizo uso de un Modelo Lineal Generalizado Ponderado con respuesta Poisson

para analizar la edad de muerte del paciente con cáncer, donde se utilizó la base de datos que maneja el INC, con esta información se desarrollaron estadísticas descriptivas para el análisis de la Situación del Cáncer en Colombia 2015, donde se permite avanzar en la definición de prioridades para la acción en la población a riesgo o que padece esta enfermedad (Piñeros et al., 2017). Al igual el INS realizó una investigación de vigilancia y análisis del riesgo en salud pública protocolo de vigilancia en salud pública cáncer de mama y cuello uterino, donde se le realizó un análisis de correspondencia con los casos reportados en la Cuenta de Alto Costo de forma periódica, donde se encontró que las campañas de prevención van en aumento, los tratamientos son más accesibles, y se cuenta con amplia atención del sistema de salud.

En el 2018, la Universidad del Valle realizó un estudio titulado Estimaciones de incidencia y mortalidad para los principales cinco tipos de cáncer en Colombia, 2007-2011 donde usaron un MLG para analizar la incidencia en cada una de las localizaciones del cáncer y se obtuvo que el número anual de casos esperados (todos los cánceres) fue 62.818 en hombres y en mujeres y se registraron 32.653 muertes. Los principales cánceres incidentes fueron próstata (46.5 por 100.000 años-persona) en hombres y mama (33.8 por 100.000 años-persona) en mujeres. La mortalidad más alta en hombres se presentó en estómago (14.2) y mama en mujeres (9.9)(Pardo et al., 2018). Según (Chou-Chen et al., 2018) en la investigación un modelo jerárquico bayesiano espacio-temporal con variable de conteos: aplicación de VIH/SIDA en Costa Rica los modelos espaciales que suavizan las tasas de mortalidad estandarizada o los riesgos relativos son utilizados ampliamente en el mapeo de enfermedades y se encontró que es un mejor método al usar las variables de conteo las estimaciones son más precisas.

Por lo anterior se identifica que en Colombia, se implementan políticas públicas de prevención, y de pronta atención para que los pacientes de cáncer de mama y cáncer de próstata, tienen una alerta temprana y acceso a un pronto tratamiento, estas políticas implementadas aportan en gran parte a la disminución de las muertes por estos dos tipos de cáncer. Igualmente, diversas metodologías para el análisis del cáncer de mama y cáncer de próstata dentro de estas metodologías analizadas tenemos: estadísticas descriptivas, modelos lineales generalizados, modelos espaciales, modelos de regresión joinpoint, modelos de riesgos proporcionales de Cox, modelos de regresión, análisis de correspondencias. Existen muchas investigaciones y avances en el estudio del cáncer sin embargo, aún quedan muchos aspectos por estudiar, como más causas y riesgos por los que se presenta el cáncer, influencias ambientales en la enfermedad, como también el tipo de seguridad social que se tenga, entre otras problemáticas que puedan afectar al cáncer de mama y cáncer de próstata.

3. Marco Conceptual

Modelos Espaciales

Los datos espaciales se definen como realizaciones de un proceso estocástico indexado por el espacio $Y(s) = \{y(s), s \in D\}$ donde D es un subconjunto (fijo) de $\mathbb{R}$. Los datos reales luego puede representarse mediante una colección de observaciones $y = \{y(s_1), \dots, y(s_n)\}$, donde el conjunto $(s_1, \dots, s_n)$ indica las unidades espaciales en las que se toman las medidas. Dependiendo de que D sea una superficie continua o una contable colección de unidades espaciales d-dimensionales, el problema se puede especificar como un Proceso aleatorio espacialmente discreto o continuo, respectivamente.(Gelfand et al.,2010).

Modelos Espacio-Temporales

Uno de los objetivos primordiales en el diseño de modelos geoestadísticos espacio-temporales es la representación de los diferentes niveles de interacción espacio-temporal que caracterizan la estructura de dependencia de magnitudes definidas en el espacio y tiempo.(Salmerón, 2008). Para hablar de modelos espacio temporales primero se debe definir la realización de un proceso estocastico indexado por el espacio y el tiempo y se observan en n ubicaciones o areas espaciales y en T puntos del tiempo la cual es la

siguiente:

$$Y(s, t) = \{y(s, t), (s, t) \in D \subset \mathbb{R}^2 \times \mathbb{R}\}$$

A continuación, se muestran algunas definiciones importantes para la contextualización de la investigación:

Cáncer Según la OMS el cáncer, es un proceso de crecimiento y diseminación incontrolados de células. Puede aparecer prácticamente en cualquier lugar del cuerpo. El tumor suele invadir el tejido circundante y puede provocar metástasis en puntos distantes del organismo. Muchos tipos de cáncer se podrían prevenir evitando la exposición a factores de riesgo comunes como el humo de tabaco. Además, un porcentaje importante de cánceres pueden curarse mediante cirugía, radioterapia o quimioterapia, especialmente si se detectan en una fase temprana.

Cáncer de Próstata Este cáncer se origina cuando las células sanas de la próstata cambian y proliferan sin control, y forman un tumor. Un tumor puede ser canceroso o benigno. Un tumor canceroso es maligno, lo que significa que puede crecer y diseminarse a otras partes del cuerpo. Un tumor benigno significa que el tumor puede crecer, pero no se diseminará.

Cáncer de Mama El cáncer de mama se presenta con el crecimiento anormal de algunas células mamarias. Estas células se dividen más rápidamente que las células sanas y continúan acumulándose, y de este modo forman un bulto o una masa. Las células pueden esparcirse (hacer metástasis) por toda la mama hasta los ganglios linfáticos u otras partes del cuerpo.

El cáncer de mama más a menudo comienza en las células de los conductos que producen leche (carcinoma de conducto invasivo). Este también puede comenzar en el tejido glandular llamado «lobulillos» (carcinoma lobulillar invasivo), o en otras células o tejidos dentro de la mama.

otro concepto importante para tener en cuenta es la definición del DIC (Deviance Information Criterion)

DIC(Criterio de Información) El DIC es un criterio jerárquico generalizado del AIC y el BIC que está basado en la metodología bayesiana, ya que utiliza las cadenas de Markov- Montecarlo (MCMC). El DIC lo que hace es obtener la devianza, la cuál tiene en cuenta la función completa de verosimilitud. Este mide la complejidad del modelo en términos de sus propiedades de estimación, la complejidad del modelo se refiere al número efectivo de parámetros en el modelo. Desde el punto de vista bayesiano el DIC, consiste en incorporar la información a priori de los parámetros (Spiegelhalter et. al, 2002).

Por tanto la medida de ajuste será el promedio de la devianza

$$\bar{D}(\theta) = -2 \int \log(p(y|\theta)) d\theta$$

Para medir la complejidad del modelo el DIC utiliza también la devianza, estudiando las covarianzas entre los parámetros del modelo.

Al final, el DIC se representa como :

$$DIC = \bar{D}(\theta) + p(D) = D(\bar{\theta}) + 2pD$$

Esto da lugar a que el DIC englobe para su implementación tanto el tamaño muestral como el número de parámetros, así como la relación de covariación entre estos.

4. Marco Teórico

4.1. Modelo de Regresión Poisson

Antes de definir el modelo de regresión Poisson, se presentan algunos resultados que son importantes para la explicación de esta regresión.

Teorema 1

Sea $y \sim \text{Poisson}(\lambda)$, y , pertenece a la familia exponencial uniparamétrica. Donde $p(\lambda)$ es la función de probabilidad de y .

$$\begin{aligned} p(y|\lambda) &= \frac{e^{-\lambda} \lambda^y}{y!} & y = 0, 1, 2, \dots & \quad \lambda > 0 \\ &= e^{y \log(\lambda) - \lambda + \log(\frac{1}{y!})} \end{aligned}$$

reemplazando $\theta = \log(\lambda)$, tenemos:

$$p(y|\theta) = e^{y\theta - e^\theta + \log(\frac{1}{y!})}$$

Donde: $a(\phi) = 1$, $y\theta = y \log(\theta)$, $b(\theta) = e^\theta$, $c(y, \phi) = \log(\frac{1}{y!})$, por lo tanto $y \sim P(\theta)$ pertenece a la familia exponencial uniparamétrica.

En modelos lineales, se tiene que:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} = \mathbf{x}_i' \boldsymbol{\beta}$$

Donde:

$$\begin{aligned} \mathbf{x}_i' &= (1 \quad x_{i1} \quad x_{i2} \dots x_{ik}) \\ \boldsymbol{\beta} &= \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \end{aligned}$$

Teniendo en cuenta que la variable respuesta y_i , es de conteos, ya que sigue una distribución Poisson, no se admiten reales no negativos. Entonces se propone una función de enlace logarítmica:

$$\begin{aligned} g : (0, \infty) &\rightarrow \mathbf{R} \\ x &\rightarrow \log(x) \end{aligned}$$

Por tanto:

$$\eta_i = \ln(\lambda_i) = \mathbf{x}_i' \boldsymbol{\beta}$$

Donde: $\lambda_i = E(y_i) = \theta_i$ el modelo de Poisson se estructura de la siguiente manera:

$$\begin{aligned} \ln(E(y_i)) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} \\ \lambda_i(y_i) &= E(y_i) = e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}} \end{aligned}$$

4.2. Modelo de Regresión Poisson Bayesiano

Dado que $y_i \sim \text{Poisson}(\theta_i)$, la función de verosimilitud $L(\mathbf{Y}|\boldsymbol{\beta})$, con $\mathbf{Y} = (y_1, y_2, \dots, y_n)$ y $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$ se define como:

$$L(\mathbf{Y}|\boldsymbol{\beta}) = \prod_{i=1}^n \frac{e^{-\theta_i} \theta_i^{y_i}}{y_i!}$$

Tomando la función de enlace y reemplazando se tiene que $\theta_i = e^{x'_i \beta}$, entonces la verosimilitud es:

$$L(\mathbf{Y}|\beta) = \prod_{i=1}^n \frac{e^{-e^{x'_i \beta}} e^{y_i x'_i \beta}}{y_i!}$$

$$= \frac{1}{\prod_{i=1}^n y_i!} e^{\sum_{i=1}^n y_i x'_i \beta - \sum_{i=1}^n e^{x'_i \beta}}$$

Para β , se postula una prior normal multivariada ($k+1$ -variada), con vector de medias $\mu = (\beta_0, \beta_1, \dots, \beta_k)$ y matriz de varianzas y covarianzas $\mathbf{Q}$.

$$\beta \sim N_k(\mu, \mathbf{Q})$$

la distribución a prior para β está dada por:

$$P(\beta) = \frac{1}{\sqrt{2\pi}} |\mathbf{Q}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\beta - \mu)' \mathbf{Q}^{-1}(\beta - \mu)}$$

Entonces la función de densidad posterior queda de la forma:

$$P(\beta|\mathbf{Y}) \propto P(\mathbf{Y}|\beta)P(\beta)$$

$$\propto e^{\left(\sum_{i=1}^n y_i x'_i \beta - \sum_{i=1}^n e^{x'_i \beta}\right)} e^{-\frac{1}{2}(\beta - \mu)' \mathbf{Q}^{-1}(\beta - \mu)}$$

$$\propto e^{\left(\sum_{i=1}^n y_i x'_i \beta - \sum_{i=1}^n e^{x'_i \beta} - \frac{1}{2}(\beta - \mu)' \mathbf{Q}^{-1}(\beta - \mu)\right)}$$

Nótese que no se tiene una distribución multivariada conocida, es por esto que la estimación de los parámetros se hace mediante JAGS, este aplicativo hace uso de un algoritmo numérico para estimar la distribución posterior, el algoritmo que usa el aplicativo es MCMC, finalmente con este se realiza la estimación de los parámetros de interés.

4.3. Modelo Poisson Espacial

Teniendo en cuenta los factores del espacio, se tiene que el modelo espacial se encuentra definido de la siguiente manera:

$$y_i \sim \text{Poisson}(\lambda_i)$$

donde, y_i es el número de muertes por cáncer de mama y próstata y

$$\lambda_i = E_i \rho_i$$

.

donde, $\lambda_{it} = E_{it} \rho_{it}$

λ_i es definida en terminos de una tasa ρ_i y el número esperado de muertes por cáncer de mama y próstata E_i .

para este modelo el predictor lineal está definido en escala logarítmica así:

$$\log(\rho_i) = \eta_i$$

$$\eta_i = b_0 + u_i + v_i$$

donde b_0 es el intercepto, cuantificando la tasa de resultado promedio en toda la región del estudio, mientras que v_i es el efecto específico del área modelado como intercambiable. Además, u_i , otro efecto de área específica, que ahora modelamos como estructurado espacialmente.

4.4. Modelos Condicionales Autorregresivos

Los modelos CAR (Conditional Auto-Regressive models) en los datos de área se fundamenta en una sola medida agregada por cada unidad de área, está medida puede tomar un valor binario, de conteo o continuo, por lo general las unidades de área son volúmenes, en especial con límites bien definidos las unidades de área dividen un volumen D multidimensionalmente en un número finito.

Dado un conjunto de observaciones tomadas en n diferentes unidades de área de una región, interacciones espaciales entre un par de unidades n_i y n_j puede ser modelado condicionalmente como una variable aleatoria espacial u , que es un vector n de longitud $\mathbf{u}_i = (u_1, \dots, u_n)^T$.

Para los modelos CAR, la relación espacial entre las n unidades de área se representan como una matriz de adyacencia $\mathbf{W}$ con dimensiones $n \times n$ donde las entradas w_{ij} y w_{ji} , y son positivos cuando las regiones n_i y n_j son vecinos y cero de lo contrario. La relación de vecinos $i \sim j$, se define en términos de esta matriz, los vecinos de la región i son aquellas regiones que tienen entradas distintas de cero en la fila o columna i . Esta codificación define una estructura reticular sobre n unidades de área.

4.4.1. Modelo Condicional Autorregresivo (CAR)

(Besag, 1974) motiva los modelos CAR para procesos espaciales utilizando los resultados de la física de los sistemas de partículas de celosía y el teorema de Hammersley-Clifford, que proporciona una equivalencia entre una especificación local de la distribución condicional de cada partícula dadas sus partículas vecinas y la especificación global de la distribución conjunta de todas las partículas. Esta especificación de la distribución conjunta a través de la especificación local de las distribuciones condicionales de las variables individuales es una especificación del campo aleatorio de Markov (MRF). La distribución condicional para cada u_i se especifica en términos de un parámetro de media y precisión σ como:

$$u_i | u_{-i} \sim N \left(\mu_i + \sum_{j=1}^n r_{i,j} (u_j - \mu_j), S_i^2 \right), i, j = 1, \dots, n$$

Donde se tiene que :

- μ_i es la media del área i
- $S_i^2 = \frac{\sigma_u^2}{N(i)}$ es la varianza del área que depende del número de vecinos ($N_i = \#(N_i)$)
- r_{ij} es la proximidad espacial y se calcula como $\Phi \times W_{ij}$, donde $W_{ij} = \frac{a_{ij}}{N(i)}$; donde a_{ij} es 1 si las áreas i y j son vecinas, en caso de serlo.

La distribución conjunta correspondiente se puede determinar de manera única a partir del conjunto de distribuciones condicionales completas introduciendo un punto fijo desde el soporte de p

$$u \sim \mathcal{N}_m(\boldsymbol{\mu}, [(\mathbf{I} - \Phi \mathbf{W})]^{-1} \mathbf{S}^2)$$

dónde

- $\boldsymbol{\mu}$ es el vector de medias $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_n\}$
- $\mathbf{S}^2$ es una matriz $n \times n$ de pesos $\mathbf{W}$ donde las entradas i, i son cero y los elementos fuera de la diagonal describen la proximidad espacial de las regiones n_i y n_j
- $\mathbf{I}$ es una matriz identidad de $n \times n$

- $\mathbf{W} = \tau D$ donde D es la diagonal de una matriz $n \times n$

La construcción de la matriz de proximidad espacial $\mathbf{S}^2$ determina la clase de estructura del modelo CAR. En el caso donde $\mathbf{S}^2$ es una matriz definida positiva. Sin embargo, la evaluación de la distribución conjunta requiere calcular la matriz de covarianza descrita por $[(\mathbf{I} - \Phi \mathbf{W})]^{-1} \mathbf{S}^2$. la distribución condicional para $u_i | u_{-i}$

$$u_i | u_{-i} \sim N \left(\mu_i + \Phi \frac{1}{N_i} \sum_{j=1}^n a_{i,j} (u_j - \mu_j), S_i^2 \right)$$

4.4.2. Modelos Condicionales Autoregresivos (ICAR)

Un modelo intrínseco de regresión automática condicional (ICAR) es un modelo CAR, cuando la correlación entre las áreas i y j depende solamente de Φ y $\mathbf{W}$, (Assuncao et al., 2009), esta dado por :

$$\frac{\sqrt{N_i}}{N_j} \frac{(\mathbf{I} - \Phi \mathbf{W})_{ij}^{-1}}{\sqrt{(\mathbf{I} - \Phi \mathbf{W})_{ii}^{-1} (\mathbf{I} - \Phi \mathbf{W})_{jj}^{-1}}}$$

Esta especificación no se usa mucho en los estudios de mapeo de enfermedades dado que el parámetro puede resultar difícil de estimar, una versión simplificada de la formulación anterior se obtuvo de la corrección de $\Phi = 1$, lo que significa que la matriz de covarianza no es positiva definida; esto conduce a la siguiente distribución condicional para u , (Blangiardo et al., 2015).

$$(u_i | \mathbf{u}_{-i}) \sim N \left(\mu_i + \frac{1}{N(i)} \sum_{j=1}^n a_{i,j} (u_j - \mu_j), S_i^2 \right)$$

- $\Phi = 1$
- $\mathbf{S}_i^2 = \frac{\sigma_u^2}{N(i)}$ es la varianza del área que depende del número de vecinos ($N(i) = \#(\mathcal{N}(i))$).
- a_{ij} es 1 si las áreas i y j son vecinas.

Las condiciones anteriores, para el modelo autorregresivo condicional intrínseco (ICAR) producen una distribución inadecuada porque el ajuste $\Phi = 1$ crea una matriz singular $(\mathbf{I} - \mathbf{W})$, (Besag et al., 1995). Además, la distribución conjunta no es identificable; añadiendo cualquier constante a todos los elementos de u deja la distribución conjunta sin cambios. Añadiendo la restricción $\sum_i^n u_i = 0$ se resuelve este problema.

Asumiendo que $\mu_i = 0$ para cada i , la distribución condicional está ahora en la forma presentado típicamente en la literatura sobre mapeo de enfermedades, (Besag et al., 1991), (Best et al., 2005), (Lawson, 2009), (Lee, (2011).

$$(u_i | u_{-i}) \sim N \left(\mu_i + \frac{1}{N(i)} \sum_{j=1}^n a_{i,j} u_j, S_i^2 \right)$$

Si bien este modelo ICAR no se genera directamente en los datos dado que no se puede usar como un modelo para estos, pero es importante usarlo como una parte previa de un modelo jerárquico, que es el papel que desempeña en el modelo BYM.

4.4.3. Modelo BYM

Agregar un término de error estructurado espacialmente CAR a un GLM multinivel proporciona un alisamiento espacial de las estimaciones resultantes. El modelo lognormal Poisson propuesto en (Besag et al., 1991) se usa para datos de recuento en bioestadística y epidemiología. Incluye tanto un componente ICAR para el suavizado espacial como un componente ordinario de efectos aleatorios para la heterogeneidad no espacial, (Morris, 2018).

Las implementaciones de este modelo están disponibles a través de R, BUGS y JAGS, así como INLA (aproximación de Laplace anidada integrada), que es una alternativa rápida a MCMC (INLA intercambia velocidad y escalabilidad por precisión), (Banerjee et al., 2003).

Usando la notación de (Congdon, 2007), El modelo de Besag York Mollié esta dado por:

$$y_i \sim \text{Poisson}(E_i \rho_i)$$

$$\eta_i = \log(\rho_i) = \beta_0 + u_i + v_i$$

donde,

- β_0 cuantifica la tasa promedio de las muertes por cáncer de mama y próstata en los 32 departamentos de Colombia.
- E_i representa los casos esperados de muertes por cáncer de mama y próstata para cada área.
- u_i es el residuo estructurado espacialmente, modelado usando la especificación ICAR explicada en las subsecciones anteriores(4.4.1, 4.4.2, 4.4.3).
- v_i es el modelo residual no estructurado

Dado lo anterior se tiene que:

$$v_i \sim N(0, \sigma_v^2)$$

.

4.5. Modelo de Regresión Poisson Espacio-Temporal

El modelo poisson espacio temporal, está compuesto por unas componentes en en tiempo y en el espacio, el número de muertes por cáncer de mama y próstata se modela como:

$$y_{it} \sim \text{Poisson}(\lambda_{it})$$

donde y_{it} es el número de muertes en el area i en el tiempo t $\lambda_{it} = E_{it} \rho_{it}$

λ_{it} es definida en terminos de una tasa ρ_{it} y el número esperado de muertes por cáncer de mama y próstata E_{it}

para este modelo, el predictor lineal está definido en escala logarítmica asi:

$$\log(\rho_{it}) = \eta_{it}$$

$$\eta_{it} = b_0 + u_i + v_i + \text{Temporal}_t$$

donde b_0 es el intercepto, cuantificando la tasa de resultado promedio en toda la región del estudio, mientras que v_i es el efecto específico del área modelado como intercambiable. Además, u_i , otro efecto de área específica, que ahora modelamos como estructurado espacialmente. $t = 1, \dots, T$ En Temporal_t se puede especificar una estructura paramétrica o no paramétrica.

Para trabajar con datos a nivel de área, la dependencia espacial se tiene en cuenta a través de la estructura del vecindades. Simplificando la notación introducida $(s_1, \dots, s_n)$ se convierta en $(1, \dots, n)$, luego, dada el área i , sus vecinos $\mathcal{N}(i)$ se definen como las áreas que comparten bordes con ella (Vecinos de primer orden) o que comparten bordes con él y con sus vecinos de primer orden (vecinos de segundo orden).

Bajo la propiedad Markoviana, el parámetro θ_i para el área i es independiente de todos los demás parámetros, dado el conjunto de sus vecinos $\mathcal{N}(i)$ (propiedad local de Markov), entonces $\theta_i \perp \theta_{-i} | \theta_{\mathcal{N}(i)}$.

Se pueden especificar varias estructuras sobre u . Una en particular es la que viene dada en (Besag et al., 1995) (modelo CAR), ya mencionada anteriormente.

$$u_i | u_{-i} \sim N \left(\mu_i + \sum_{j=1}^n r_{ij} (u_j - \mu_j), \sigma_i^2 \right)$$

donde μ_i es la media para el área i , $\sigma_i^2 = \sigma_u^2 / \mathcal{N}(i)$ es la varianza para la misma área, que depende de la cantidad de vecinos. Esta estructura de varianza reconoce el hecho de que, en presencia de una fuerte correlación espacial, entre más vecinas a un área tiene más información en los datos sobre el valor de su efecto aleatorio, mientras que el parámetro de varianza σ_u^2 controla la cantidad de variación entre los efectos aleatorios estructurado espacialmente. La cantidad r_{ij} indica la proximidad espacial y puede ser calculada como $r_{ij} = \phi a_{ij} / \mathcal{N}(i)$. Luego $\mathbf{u} \sim N_n(\boldsymbol{\mu}, (\mathbf{I} - \boldsymbol{\phi W})^{-1} \mathbf{S}^2)$.

Tendencia paramétrica del efecto temporal

(Bernardinelli et al., 1995) presentaron una tendencia paramétrica para el componente temporal en la que asume que el predictor lineal puede escribirse como

$$\eta_{it} = b_0 + u_i + v_i + (\beta + \delta_i)t$$

Esta formulación incluye el efecto espacial principal, la tendencia lineal principal β , que representa el efecto del tiempo global, y una tendencia diferencial δ_i que identifica la interacción entre el tiempo y el espacio. Dado que para fines de identificación se impone una restricción de suma a cero sobre el vector δ . Si $\delta_i < 0$, la tendencia específica del área es menos pronunciada que la tendencia media, mientras que $\delta_i > 0$ implica que la tendencia específica del área es más pronunciada que la tendencia media.

5. Resultados y Discusión

En este trabajo de investigación se realiza un estudio del comportamiento de las muertes por cáncer de mama y cáncer de próstata en Colombia durante un periodo de diez años, los años de estudio son del 2006 al 2015, dado que el cáncer de mama y próstata representan en gran cantidad las muertes de los hombres y mujeres colombianos.

5.1. Descriptivos

En los análisis de los datos se realizará una visualización de como se comporta la variable respuesta y a través de tres pruebas de correlación espacial se evalúa si la variable respuesta presenta correlación espacial o aleatoriedad, además se muestra el mapeo de las enfermedades en cada uno de los departamentos del territorio colombiano.

5.1.1. Descripción de los Datos

Los datos utilizados para esta investigación se obtuvieron de el el Sistema Nacional de Información en Cáncer, es un conjunto de datos organizados, procedentes de diferentes fuentes de información, que se originan de los procesos de atención a los pacientes con cáncer, para este estudio contamos con unas variables explicativas y una variable respuesta, presentadas más adelante, los datos fueron obtenidos del Sistema de Información del Cáncer en Colombia y del DANE(Departamento Nacional de Estadísticas), en el periodo 2006-2015 para cada uno de los departamentos del territorio colombiano. La variable respuesta es de tipo conteo, las covariables que se usan en este trabajo son de este mismo tipo y de tipo porcentaje.

Variable respuesta 1: Número de personas fallecidas por cáncer de mama por departamento en Colombia para los años 2006-2015.

Variable respuesta 2: Número de personas fallecidas por cáncer de próstata por departamento en Colombia por departamento para los años 2006-2015.

las covariables son:

- Número de hombres afiliados por departamento al regimen contributivo para los años 2006-2015
- Número de mujeres afiliadas por departamento al regimen contributivo para los años 2006-2015
- Número de hombres afiliados por departamento al regimen subsidiado para los años 2006-2015
- Número de mujeres afiliadas por departamento al regimen subsidiado para los años 2006-2015
- Número de defunciones por departamento para el año 2006-2015
- Población total para cada departamento en los años 2006-2015
- Tasa de desempleo por departamento para los años 2006-2015

5.2. Mapas de Número de Muertes

5.2.1. Mapeo de Enfermedades

(Pascutto et al., 2000), (Lawson, 2009) el mapeo de las enfermedades se usa para observar e identificar el riesgo relativo de alguna enfermedad en particular, ademas de el patron espacial que dicha enfermedad presenta, para este análisis los datos se realizo el conteo en cada una de las zonas. A continuación se presenta la ecuación para realizar el calculo de los riesgos relativos de las enfermedades.

$$r_j = \frac{\sum_{i=1}^n y_{ij}}{\sum_{i=1}^n pop_{ij}}$$

Donde se tiene la relación del número de casos observados y_{ij} y población en estudio pop_{ij} . El número de casos esperados de la enfermedad se calcula con las tasas de referencia estandarizadas r_j y la población en estudio, como se muestra a continuación.

$$E_i = \sum_{j=1}^J pop_{ij} r_j$$

luego de calcular los procedimientos anteriores se procede a realizar el mapeo de la enfermedad en estudio, se cuenta con 32 departamentos y el número de muertes por cáncer de próstata y mama.

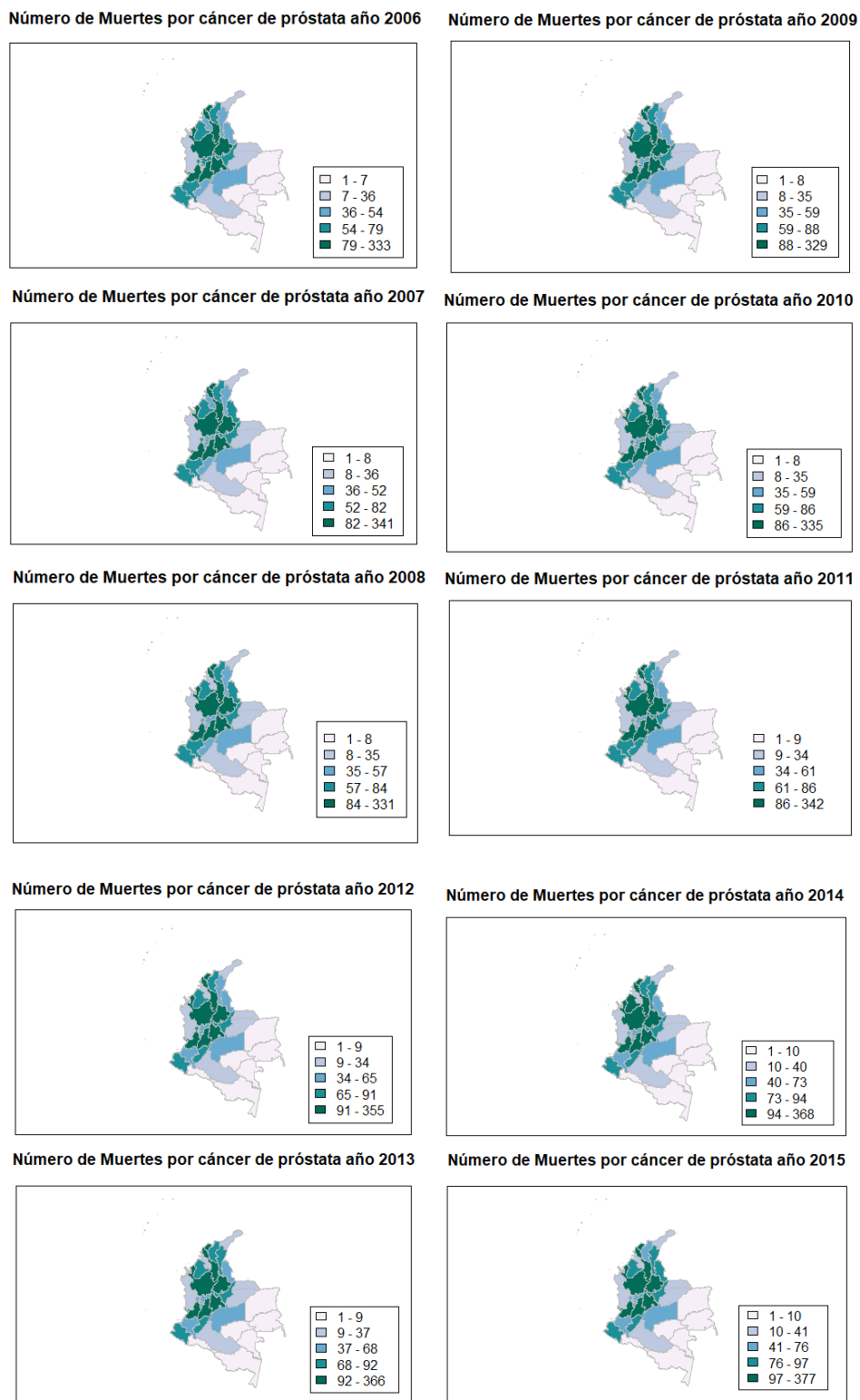


Figura 1: Muertes a causa de cáncer de próstata por departamento del año 2006 al 2015 .

Fuente: Elaboración propia a partir del software RStudio.

En la figura 1, se puede observar que durante los diez años de estudio las muertes por cáncer de próstata, se presentan con una gran proporción en la región Andina, en la región Pacífica y Caribe se presentan departamentos con gran cantidad de casos, pero en menos cantidad de áreas, la región Orinoquia y Amazónica son las regiones donde menor cantidad de muertes por cáncer de próstata se presentan. Los departamentos donde se presentaron mayor cantidad de muertes durante los años de estudio fueron: Cundinamarca, Antioquia, Valle del cauca, Santander, Bolívar y Atlántico. Esta tendencia se puede presentar debido a que son las áreas más pobladas y donde se encuentra la mayor cantidad de centros de salud especializados para los tratamientos del cáncer.

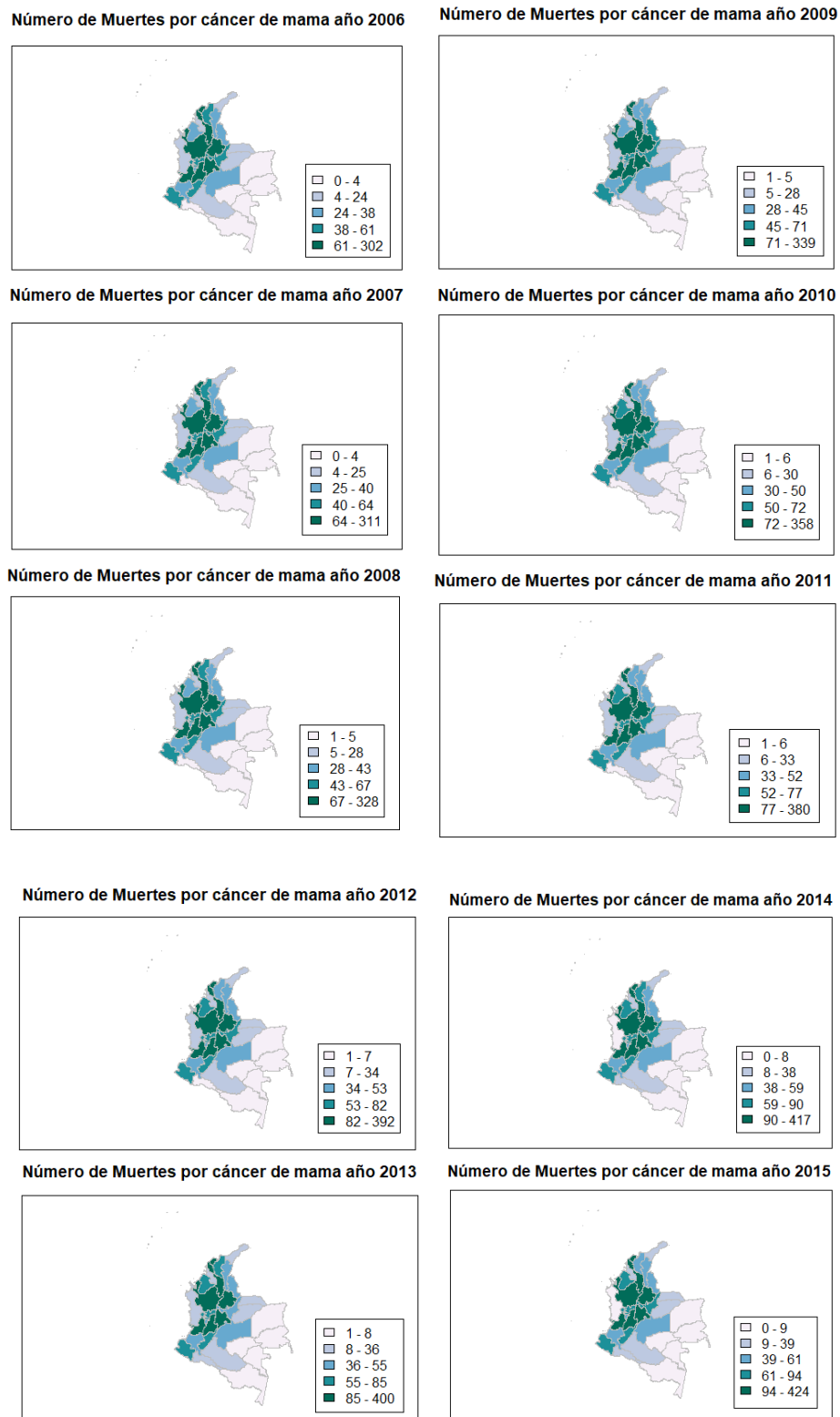


Figura 2: Muertes a causa de cáncer de mama por departamento del año 2006 al 2015 .

Fuente: Elaboración propia a partir del software RStudio

En la figura 2, se puede identificar que durante los diez años de estudio las muertes por cáncer de mama, se presentan en gran cantidad en la región Andina, así como también lo fueron las muertes a causa de cáncer de próstata en la región Pacífica y Caribe se presentan departamentos con gran cantidad de casos, pero en menor número de departamentos, la región Orinoquia y Amazónica son las regiones donde menor cantidad de muertos por cáncer de próstata se presentan. Los departamentos donde se presentaron más muertes durante los años de estudio fueron los departamentos de Cundinamarca, Antioquia, Valle del Cauca, Santander, Bolívar, Atlántico. Esta tendencia se puede presentar debido a que son las áreas más pobladas y donde se encuentran la mayor cantidad de centros de salud especializados para los tratamientos del cáncer.

Se identifica que las muertes por los dos tipos de cáncer se comportan de igual manera presentándose su mayor mortalidad en los mismos departamentos, aunque el de mama con un aumento más significativo en el número de casos, esto puede ser debido a que la mayoría de los lugares son principales ciudades del país y en estas se ubican la mayor cantidad de centros asistenciales para el tratamiento de la enfermedad del cáncer.

5.3. Pruebas de Correlación de la Variable Respuesta

En el modelo a utilizar debe existir una correlación espacial en la variable respuesta a través del tiempo de estudio, donde dicha condición se valida a través de las pruebas de Mantel, Moran y Geary. Estas pruebas se basan en la noción de autocorrelación espacial de la variable de respuesta a través de los años, debido a que asocia la idea que los valores observados en áreas geográficas adyacentes serán más similares que los esperados bajo el supuesto de independencia espacial (Cressie, 1993).

Si $Z(s)$ es el atributo Z observado en el sitio s , entonces el término auto-correlación espacial se refiere a la correlación entre $Z(s_i)$ y $Z(s_j)$. Es una correlación entre el mismo atributo en dos sitios diferentes. Si hay auto-correlación espacial positiva, la proximidad en el espacio debe emparejarse con la similitud en los valores de los atributos de interés.

El sistema de hipótesis para estas tres pruebas se plantea de la siguiente manera:

Prueba de Mantel

Permite comprobar estadísticamente si las observaciones que provienen de un proceso estocástico donde las variables se encuentran correlacionadas espacialmente. El estadístico de prueba está dada por:

$$M = \sum_{i=1}^n \sum_{j=1}^n W_{ij} U_{ij}$$

donde W_{ij} se denota como una medida de proximidad espacial entre (s_i) y (s_j) y U_{ij} una medida de proximidad de los atributos. Para el valor del estadístico de prueba se tiene que entre mayor sea el valor de M se tiene mayor dependencia lineal

Prueba de Moran La prueba de autocorrelación de Moran, permite evaluar la información de los vecinos más cercanos con el siguiente estadístico de prueba:

$$I = \frac{n * \sum_{i=1}^n \sum_{j=1}^n W_{ij} (Z(s_i) - \bar{Z})(Z(s_j) - \bar{Z})}{\sum_{i=1}^n \sum_{j=1}^n W_{ij} (\sum_{i=1}^n W_{ij} (Z(s_i) - \bar{Z}))^2}$$

donde n es el número de unidades espaciales indexados por i y j , tenemos que Z es la variable de interés; $\bar{Z}$ es la media de Z ; y W_{ij} se denota como una medida de proximidad espacial entre (s_i) y (s_j) .

Se tiene que los valores positivos (entre 0 y 1) indican autocorrelación directa (similitud entre valores cercanos) y valores negativos (entre -1 y 0) indican autocorrelación inversa (disimilitud entre las áreas cercanas). Valores del coeficiente cercanos a cero apoyan la hipótesis de aleatoriedad espacial.

Prueba de Geary Permite el cálculo de estacionariedad de primer y segundo orden (Media y varianza constante.), el estadístico de prueba se define como :

$$C = \frac{(n-1)}{2S_0} \left(\frac{n \sum_{i \neq j} W_{ij} (z(s_i) - z(s_j))^2}{\sum_{i \neq j} (z(s_i) - \bar{z})^2} \right)$$

dónde n es el número de unidades espaciales indexadas por i y j , Z es la variable de interés, $\bar{Z}$ es la media de Z w_{ij} es una matriz de pesos espaciales con ceros en la diagonal y s_0 es la suma de todos los w_{ij} .

Su interpretación es opuesta a la de la I de Moran, ya que los valores de c que tienden a 0 ($c < 1$) indican autocorrelación positiva, mientras que cuando c tiende a 2 ($c > 1$) corresponde a autocorrelación negativa y, en ausencia de autocorrelación espacial significativa, $c = 1$. Es decir que altos valores de c equivalen a bajos valores de I y viceversa.

Para las dos variables respuesta de la investigación los resultados del p-valor de estas pruebas se presentan a continuación en las tablas siguientes:

Tabla de p-valores de las pruebas de correlación para el cáncer de mama

Año	Prueba de Mantel	Prueba de Moran	Prueba de Geary
2006	0.625	0.1825	0.6662
2007	0.705	0.1681	0.6473
2008	0.693	0.1688	0.6537
2009	0.691	0.152	0.6483
2010	0.7173	0.1196	0.6301
2011	0.679	0.1156	0.6312
2012	0.689	0.1094	0.6063
2013	0.665	0.1115	0.5931
2014	0.719	0.109	0.5646
2015	0.775	0.1237	0.5692

Tabla 1: Resultados p-valor Pruebas cáncer de mama

Fuente: Elaboración propia a partir de los resultados obtenidos de las pruebas realizadas en el software estadístico RStudio.

En la tabla 1, se encuentra la prueba de autocorrelación espacial para la variables respuesta tasa estandarizada de muertes por cáncer de mama tiene como resultado un p-valor no significativo estadísticamente para ninguno de los años, con un nivel de significancia del 5 % no existe suficiente evidencia estadística para decir que existe correlación estadística, por lo tanto la variable respuesta presenta aleatoriedad.

Tabla de p-valores de las pruebas de correlación para el cáncer de próstata

Año	Prueba de Mantel	Prueba de Moran	Prueba de Geary
2006	0.435	0.3695	0.8185
2007	0.463	0.281	0.8942
2008	0.477	0.359	0.8829
2009	0.441	0.341	0.8799
2010	0.5763	0.347	0.8739
2011	0.5825	0.345	0.8678
2012	0.5866	0.343	0.8646
2013	0.5842	0.361	0.8679
2014	0.5727	0.371	0.8717
2015	0.5724	0.395	0.8734

Tabla 2: Resultados p-valor Pruebas cáncer de próstata

Fuente: Elaboración propia a partir de los resultados obtenidos de las pruebas realizadas en el software estadístico RStudio.

Las pruebas de autocorrelación espacial para la variable respuesta número de muertes por cáncer de próstata tiene como resultado un p-valor no significativo estadísticamente para ninguno de los años, con un nivel de significancia del 5 % no existe suficiente evidencia estadística para decir que existe correlación estadística, por lo tanto la variable respuesta presenta aleatoriedad.

Los resultados anteriores implican que el modelo espacio temporal podría no llegar a ser óptimo para analizar estos datos, dado que no se presenta correlación espacial en la variable respuesta.

5.4. Modelos Espacio-temporales

Antes de estimar el modelo, se procede a ver la matriz de vecindades de los 32 departamentos, los departamentos estan enumerados del 1 al 32 en orden alfabético en los dos casos de estudio. en la figura 3, se encuentra la matriz de adyacencia para las muertes por cáncer de mama y en la figura 4, la matriz de adyacencia para las muertes por cáncer de próstata.

Matriz de Adyacencia Una matriz de adyacencia se define de la siguiente manera: Sea G una gráfica con vértices n que se supone que están ordenados de v_1 a v_n . La matriz $n \times n$ A , en la que

$a_{ij} = 1$ si existe una relación entre i y j .

$a_{ij} = 0$ de lo contrario

por ejemplo las filas y columnas identificar áreas; los cuadrados identifican los vecinos. (Julio, 2014)

Matriz de adyacencia de cáncer de mama de los 32 departamentos

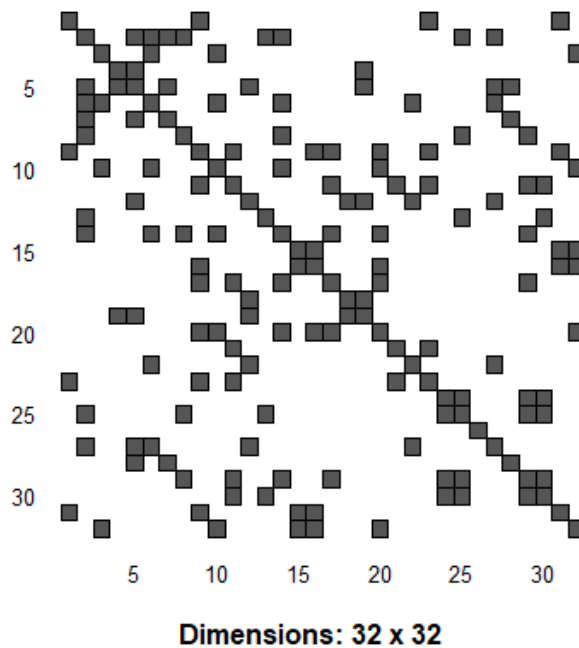


Figura 3: Matriz de adyacencia para las muertes por cáncer de mama en los departamentos de Colombia. ejemplo: filas y columnas identificar áreas; los cuadrados identifican los vecinos.

Fuente: Elaboración propia a partir del software RStudio.

Se puede identificar que el departamento que más vecinos tiene es el departamento numero cinco el cual es Bolívar, pero en general se presenta bastante aleatoriedad, los departamentos no presentan en su mayoría vecinos.

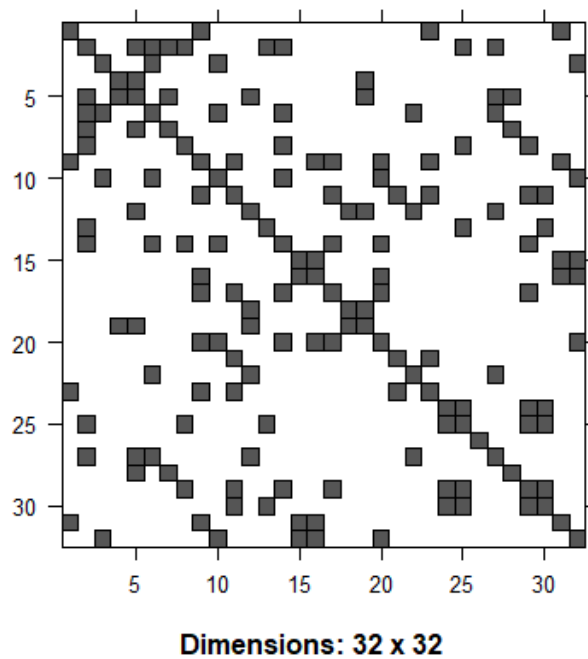
Matriz de adyacencia de cáncer de próstata de los 32 departamentos

Figura 4: Matriz de adyacencia para las muertes por cáncer de próstata en los departamentos de Colombia. ejemplo: filas y columnas identificar áreas; los cuadrados identifican los vecinos.

Fuente: Elaboración propia a partir del software RStudio.

Luego de observar las matrices de adyacencia se procede a realizar la estimación de los modelos espacio temporales, estos modelos se llevan a cabo con las variables anteriormente mencionadas.

5.4.1. Cáncer de Mama

En esta subsección, se presentan los resultados acerca de los modelos para la variable respuesta de número de muertes por cáncer de mama, las variables fueron renombradas, x1 es el número de mujeres afiliadas al regimen subsidiado, x2 es el porcentaje del PIB y x3 es el número de mujeres afiliadas al regimen contributivo. En la siguiente tabla encontraremos la media la desviacion y los intervalos de credibilidad, esta estimación fue realizada con INLA.

Resultados del modelo número uno

	media	sd	0.025	0.5	0.975
(Intercepto)	3.27	0.04	3.20	3.27	3.34
x1	0.00	0.00	-0.00	0.00	0.00
x2	0.19	0.00	0.19	0.19	0.20
x3	0.00	0.00	0.00	0.00	0.00

Tabla 3: Resultados del Modelo1 para el cáncer de mama

Fuente: Elaboración propia a partir de los resultados obtenidos del modelo realizado en el software estadístico RStudio.

En este modelo completo (Modelo1) se puede identificar que existen dos variables que no le aportan información al modelo, que no brindan información para la explicación de este, por tanto se procede a quitarlas y realizar nuevamente otro modelo para ver los resultados y al final comparar el DIC de cada modelo y analizar que modelo explicaría mejor estos datos. Las variables que se proceden a quitar del modelo son:

- Número de mujeres afiliadas al regimen contributivo
- Número de mujeres afiliadas al regimen subsidiado

Los resultados del modelo número 2 (Modelo2) quitando las variables anteriormente mencionadas queda de la siguiente manera.

Resultados del modelo número dos

	media	sd	0.025	0.5	0.975
(Intercepto)	3.26	0.04	3.19	3.26	3.33
x2	0.22	0.00	0.22	0.22	0.22

Tabla 4: Resultados del Modelo2 para el cáncer de mama

Fuente: Elaboración propia a partir de los resultados obtenidos del modelo realizado en el software estadístico RStudio.

La variable $x2$ que el porcentaje de participación del PIB, aporta al modelo, procedemos a ver la tendencia temporal lineal global para la mortalidad por cáncer de mama, la línea continua identifica la media posterior para β_t , mientras que las líneas discontinuas son el 95 % intervalos de credibilidad.

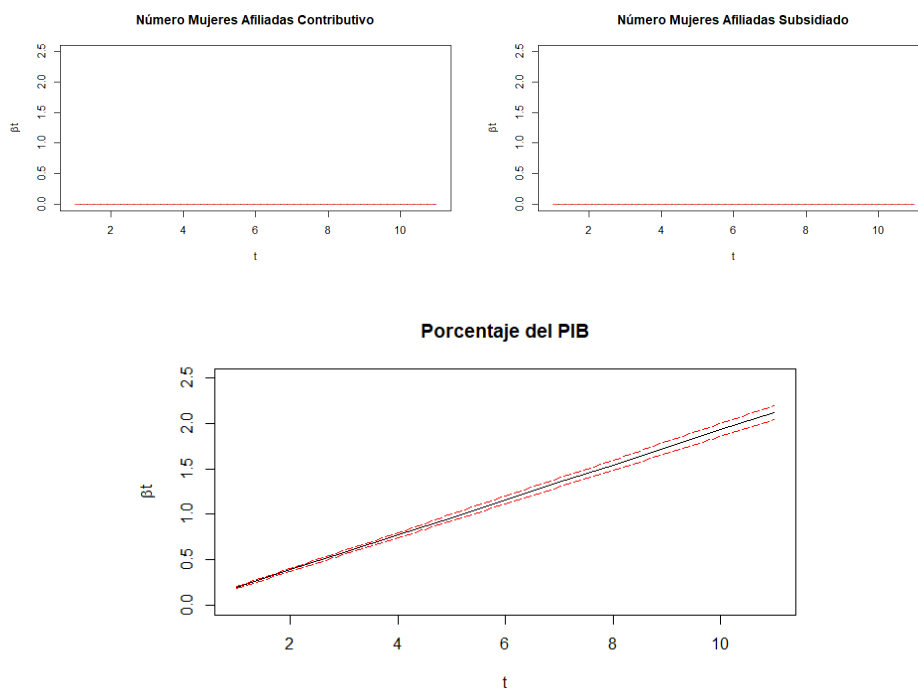


Figura 5: Tendencia de las covariables

Fuente: Elaboración propia realizado en el software estadístico RStudio.

la gráfica de la figura 5, respresenta el efecto de tendencia que tienen las covariables en el periodo evaluado, el porcentaje de participación de PIB muestra un gran crecimiento a través del tiempo, se identifica un positivo incremento en el riesgo en la mortalidad de cáncer de mama para cada año. Estas gráficas dan una mejor aproximación del comportamiento real de los efectos fijos, mostrando el verdadero efecto sobre las muertes por cáncer de mama, esto puede ser explicado porque cada año el porcentaje del PIB aumenta en la mayoría de los departamentos, pero también estos son los que presentan una mayor población. Se identifica que las covariables del número de mujeres afiliadas al régimen subsidiado y contributivo no aportan información y no tienen ninguna tendencia en el tiempo.

Comparación a través del DIC de los Dos Modelos

Tabla de comparación de los modelos mediante el DIC

Modelo	DIC	Efectividad del número de parámetros
Modelo1	13200.58	33.11
Modelo2	13317.38	30.42

Tabla 5: Comparación de los modelos

Fuente: Elaboración propia realizado en el software estadístico RStudio.

Dentro de los modelos presentados no se explica de manera óptima las muertes por cáncer de mama, de las covariables elegidas para este estudio no son significativas por lo que se deben analizar otras covariables que den un mayor nivel de explicación al modelo para trabajos futuros.

5.4.2. Cáncer de próstata

En esta subsección se presentan los resultados acerca de los modelos para la variable respuesta de número de muertes por cáncer de próstata, las variables fueron renombradas, x1 es el número de hombres afiliados al regimen subsidiado, x2 es el porcentaje del PIB y x3 es el número de hombres afiliados al regimen contributivo. En la siguiente tabla encontraremos la media la desviacion y los intervalos de credibilidad, esta estimación fue realizada con INLA.

Resultados del modelo uno

	media	sd	0.025	0.5	0.975
(Intercepto)	3.53	0.03	3.48	3.53	3.58
x1	0.00	0.00	0.00	0.00	0.00
x2	0.16	0.00	0.16	0.16	0.17
x3	0.00	0.00	0.00	0.00	0.00

Tabla 6: Resultados del modelo uno para el cáncer de prostata

Fuente: Elaboración propia realizado en el software estadístico RStudio.

En este modelo completo (Modelo1) para el cáncer de próstata se puede identificar que existen dos variables que no le aportan información al modelo, que no nos ayudan a la explicación de este, por tanto se procede a quitarlas y realizar nuevamnete otro modelo para ver los resultados y al final veremos el DIC de cada modelo y se identificara que modelo explicaria mejor estos datos.las variables que se quitaron fueron:

- Número de hombres afiliados al regimen contributivo
- Número de hombres afiliados al regimen subsidiado

Los resultados del modelo número 2 (Modelo2) quitando las variables anteriormente mencionadas queda de la siguiente manera.

Resultados del modelo dos

	media	sd	0.025	0.5	0.975
(Intercept)	3.54	0.02	3.49	3.54	3.58
x2	0.19	0.00	0.19	0.19	0.19

Tabla 7: Resultados del modelo dos para el cáncer de prostata

Fuente: Elaboración propia realizado en el software estadístico RStudio.

La variable $x2$ porcentaje de participación del PIB, aporta al modelo, procedemos a ver la tendencia temporal lineal global para la mortalidad por cáncer de próstata, la línea continua identifica la media posterior para β_t , mientras que las líneas discontinuas son el 95 % intervalos de credibilidad

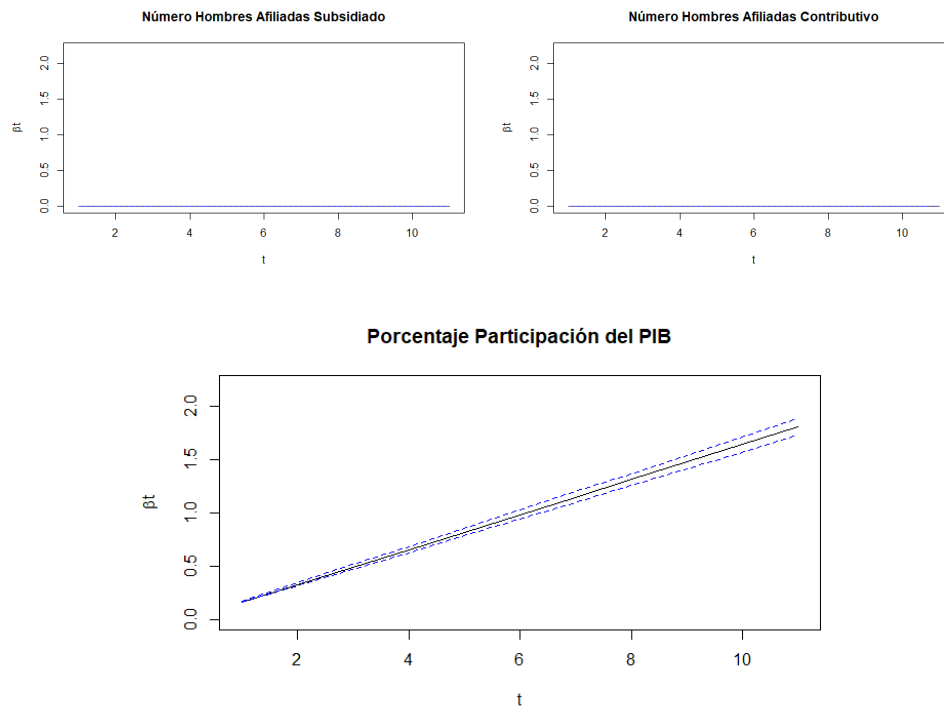


Figura 6: Tendencia de las covariables

Fuente: Elaboración propia realizado en el software estadístico RStudio.

la gráfica de la figura 6, respresenta el efecto de tendencia que tienen las covariables en el periodo evaluado, el porcentaje de participación de PIB muestra un gran crecimiento a través del tiempo, se identifica un positivo incremento en el riesgo en la mortalidad de cáncer de prostata para cada año. Estas gráficas dan una mejor aproximación del comportamiento real de los efectos fijos, mostrando el verdadero efecto sobre las muertes por cáncer de prostata, esto puede ser explicado porque cada año el porcentaje del PIB aumenta en la mayoría de los departamentos, pero también estos son los que mayor población presentan. Se identifica que las covariables del número de hombres afiliados al régimen subsidiado y contributivo no aportan información y no tienen ninguna tendencia en el tiempo.

Comparación a través del DIC de los dos modelos

Tabla de comparación de los modelos mediante el DIC

Modelo	DIC	Efectividad del número de parametros
Modelo1	9422.12	33.88
Modelo2	9493	31.93

Tabla 8: Comparación de los modelos

Fuente: Elaboración propia realizado en el software estadístico RStudio.

Dentro de los modelos presentados se identifica, que ninguno explica de manera óptima las muertes por cáncer de mama y prostata, de las covariables elegidas estas no son significativas por lo que se deben

análizar otras covariables que den un mayor nivel de explicación al modelo para trabajos futuros.

6. Conclusiones

- El análisis exploratorio del número de muertes a causa de cáncer de mama y cáncer de próstata, permite establecer que existen departamentos de Colombia con una alta incidencia de mortalidad, las zonas más afectadas son los departamentos que presentan mayor densidad poblacional, donde se encuentran las grandes ciudades, como por ejemplo Cundinamarca, Bolívar, Santander, Antioquia, Valle del Cauca.
- El análisis realizado para los diez años evidencia que las muertes por cáncer sí han tenido un aumento pero muy pequeño, en esto se evidencia que las políticas públicas, no se pueden ver reflejadas, se harían recomendaciones para que en las grandes ciudades, se presenten mayores recursos focalizados a los departamentos donde más se presenta esta problemática de salud pública.
- Los resultados de las estimaciones aunque no son lo suficientemente buenos, la covariable del porcentaje del PIB que afecta las muertes por cáncer de mama y próstata aporta información acerca de que los departamentos donde se presenta mayor participación del PIB, también se presentan mayores muertes por estos tipos de cáncer.
- La metodología utilizada para analizar los datos de número de muertes por cáncer de mama y próstata, es una dentro de muchas, en trabajos futuros se pueden analizar desde la misma metodología con covariables que brinden mayor efectividad del modelo, como también analizar la incidencia por cáncer de mama y próstata.

7. Trabajos Futuros

Para analizar de forma óptima el cáncer de mama y próstata se planea obtener mayor cantidad de datos, con otras covariables que permitan identificar con mayor calidad un modelo que explique a cabalidad esta problemática que se presenta en el país, además de analizar esta problemática por grupos de edad, por incidencia, como también es importante hacer este análisis para cada uno del municipio del territorio colombiano. Para esto se podrían presentar inconvenientes para la obtención de la información dado que no se encuentra completa y de libre acceso.

Referencias

- Agost, L. (2016). Análisis de conglomerados espacio-temporales de incidencia del cáncer pediátrico en la provincia de Córdoba, Argentina (2004-2013). Archivos argentinos de pediatría, 114(6), 534-542.
- Assuncao, R. and Krainski, E. (2009). Neighborhood dependence in Bayesian spatial models. Biometrical Journal, 5(51), 851-869.
- Banerjee, S., Carlin, B. P. y Gelfand, A. E. (eds.) (2004). Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall, Boca Raton
- Banerjee, S., Carlin, B., and Gelfand, A. (2003). Hierarchical Modeling and Analysis for Spatial Data. Monographs on Statistics and Applied Probability. Chapman & Hall, New York.
- Blangiardo, M., & Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons.

- Batista Hernández, Norma, Antón Fleites, Oscar, & Alegret Rodríguez, Milagros. (2012). Análisis espacio- temporal de eventos asociados al cáncer: una herramienta para apoyar estudios epidemiológicos. *MediSur*, 10(2), 171-181. Recuperado en 28 de febrero de 2019, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1727-897X201200020006&lng=es&tlng=es.
- Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. y Songini, M. 1995. Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine* 14: 2433- 2443
- Besag, Julian, and Charles Kooperberg. "On conditional and intrinsic autoregression." *Biometrika* (1995): 733-746.
- Besag, J., York, J. C. y Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* 43:1-59.
- Besag, Julian. "Spatial interaction and the statistical analysis of lattice systems." *Journal of the Royal Statistical Society. Series B (Methodological)* (1974): 192-236.
- Best, N., Richardson, S., and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 1(14), 35?59.
- Cressie, N. (1993). *Statistics for Spatial Data*. Wiley
- Chou, S. W., & Alvarado, R. (2018). Un Modelo jerárquico Bayesiano espacio-temporal con variable de conteos: aplicación de VIH/SIDA en Costa Rica. *Comunicaciones en Estadística*, 11(1), 9-35.
- Congdon, P. (2007). *Bayesian Statistical Modelling*. John Wiley and Sons, Ltd
- Instituto Nacional de Cancerología (2015). Análisis de Situación del Cáncer en Colombia. Bogotá: Giana María Henríquez Mendoza.
- Julio, Y. R. (2014). Bases de datos geográficas y autocorrelación espacial para identificar patrones de distribuciones espaciales. *Investigación e Innovación en Ingenierías*, 2(1).
- Gelfand, A., Diggle, P., Fuentes, M., and Guttorp, P., editors (2010). *Handbook of Spatial Statistics*. Chapman & Hall.
- González-Robledo, L. M., González-Robledo, M. C., Nigenda, G., & López-Carrillo, L. (2010). Acciones gubernamentales para la detección temprana del cáncer de mama en América Latina. Retos a futuro. *Salud pública de México*, 52(6), 533-543.
- Lawson, A. (2009). *Bayesian Disease Mapping. Hierarchical Modeling in Spatial Epidemiology*. CRC Press.
- Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-Temporal Epidemiology*, 2(2), 79?89.
- Minsalud.gov.co. (2013). [online] Available at: <https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/GCFI/guia-ross-cancer.pdf> [Accessed 21 May 2019].
- Morris, M. (2018). *Spatial Models in Stan: Intrinsic Auto-Regressive Models for Areal Data*. [online] Mc-stan.org. Available at: <https://mc-stan.org/users/documentation/case-studies/icarstan.html> [Accessed 23 May 2019].
- Osma , Martinez, S. (2012). Supervivencia a cinco años de las mujeres con cáncer de mama de Bucaramanga y su área metropolitana según el estado clínico. Magister. Universidad Industrial de Santander.

- Pascutto, C., Wakefield, J., Best, N., Richardson, S., Bernardinelli, L., Staines, A., and Elliott, P. (2000). Statistical issues in the analysis of disease mapping data. *Statistics in Medicine*, 19(17-18), 2493-2519
- Pardo, C., & Cendales, R. (2018). Estimaciones de incidencia y mortalidad para los principales cinco tipos de cáncer en Colombia, 2007-2011. *Colombia Medica*, 49(1).
- Pardo C, Cendales R. Incidencia estimada, mortalidad y prevalencia por cáncer en Colombia 2007-2011. Bogotá, D.C: Instituto Nacional de Cancerología; 2015.
- Pardo Ramos, C., & Cendales Duarte, R. (2010). Incidencia estimada y mortalidad por cáncer en Colombia 2002-2006. In *Incidencia estimada y mortalidad por cáncer en Colombia 2002-2006*.
- Piñeros Petersen, M., Pardo Ramos, C., Gamboa Garay, Ó., & Hernández Suárez, G. (2017). Atlas de mortalidad por cáncer en Colombia . In *Atlas de mortalidad por cáncer en Colombia* .
- Restrepo JA, Bravo LE, García-Perdomo HA, García LS, Collazos P, Carbonell J. Incidencia, mortalidad y supervivencia al cáncer de próstata en Cali, Colombia, 1962-2011. *Salud Pública Mex* 2014;56:440-447.
- Salmeron, R. & Ruiz -Medina,M.D. (2008). Multispectral decomposition of FAR(p models. *Stochastic and Environmental Research and Risk Assessment*, DOI: 10.1007/s00477-008-0213)
- Spiegelhalter D., Nicola G., Bradley P., Vander A., (2002): Bayesian Measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B* 64:583-639.
- Vergara-Dagobeth, E., Suárez-Causado, A., & Gómez-Arias, R. D. (2017). Plan Control del cáncer en Colombia 2012-2021. Un análisis formal. *Gerencia y Políticas de Salud*, 16(33), 6-18.
- Zambrano, S. E. O., & Uribe, C. (2014). Tasas de incidencia y mortalidad por cáncer de mama en el área metropolitana de Bucaramanga en el período 2001-2005. *MedUNAB*, 16(3), 101-111.

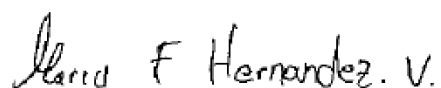
Bogotá D.C, 06 de junio de 2019

Señores

Comité de opción de grado
Facultad de Estadística
Universidad Santo Tomas

Cordial saludo,

Los datos para este estudio fueron obtenidos a través del Sistema de Información del Cáncer en Colombia, dicha información es de carácter público y publicado en la página oficial de la entidad por lo cual se da la autorización para su divulgación.



María Fernanda Hernández Villaizan
CC. 1074136295
mariafhernandez@usantotomas.edu.co