



Arnesi, Nora; Borra, Virginia; Boggio, Gabriela

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

MAPEO DE LAS RAZONES DE MORTALIDAD ESTANDARIZADAS DE CÁNCER COLORRECTAL EN ARGENTINA BASADAS EN MODELOS ESPACIALES

Resumen:

El estudio de la distribución geográfica de la mortalidad por causa es útil para los epidemiólogos y tomadores de decisión en el área de la Salud Pública. Una medida frecuentemente utilizada para identificar zonas de riesgo es la denominada razón de mortalidad estandarizada, la cual permite controlar las diferencias en estructura etaria de las regiones a comparar. En este trabajo se abordó en particular el estudio de la mortalidad por cáncer colorrectal en Argentina mediante la estimación de tales medidas a través de modelos estadísticos. En los modelos lineales generalizados mixtos se pueden considerar efectos aleatorios apropiados para tener en cuenta la dependencia espacial entre las unidades, en este caso las provincias argentinas. El mapeo de las razones de mortalidad estandarizadas obtenido bajo el modelo elegido permitió identificar la zona del noroeste argentino como la de menor riesgo, el cual aumenta paulatinamente hacia el sur con la situación de mayor riesgo en la Patagonia.

Palabras claves: modelos lineales generalizados; efectos aleatorios; correlación espacial

Abstract:

The study of the geographical distribution of mortality by cause is useful for epidemiologists and decision makers in the area of Public Health. A measure frequently used to identify areas at risk is the so called standardized mortality ratio, which allows age-adjustment for different regions. In this work, the study of mortality for colorectal cancer in Argentina was particularly addressed through the estimation of such measures through statistical models. In generalized linear mixed models, appropriate random effects can be considered to take into account the spatial dependence among units, which in this case are the Argentine provinces. The mapping of the standardized mortality ratios obtained under the chosen model, allowed identifying the Argentine North-west as the lowest risk area. The risk gradually increases towards the south, with Patagonia being highest risk area.

Keywords: generalized linear models; random effect; spatial correlation

Introducción

En el campo de la Epidemiología es importante poder cuantificar la ocurrencia de enfermedades y los casos que derivan en muerte. Las tasas crudas de mortalidad constituyen una medida resumen de la experiencia de cada población, sin embargo no resultan apropiadas cuando se las utiliza con fines comparativos. En particular, esto sucede cuando las poblaciones difieren en factores tales como edad y sexo. Una alternativa es el uso de las denominadas Razones de Mortalidad Estandarizadas-RME- (Szklo y Nieto, 2000; Preston et al, 2003), las cuales se usan frecuentemente para realizar comparaciones entre diferentes áreas o regiones y en los últimos



años se han utilizado para mostrar la distribución geográfica de la mortalidad debida a enfermedades tales como el cáncer.

El cáncer colorrectal es uno de los más frecuentes en los países occidentales. Es el tumor maligno que se desarrolla en la última porción del tubo digestivo y se produce como consecuencia de una compleja interacción de factores hereditarios y ambientales. El 90% de los cánceres colorrectales se producen en personas mayores de 50 años de edad y tiene una incidencia levemente mayor en hombres que en mujeres.

Este cáncer es el segundo cáncer más frecuente en Argentina, representa el 11,8% del total de casos en ambos sexos, detrás del cáncer de mama (16,8%) y antes del de próstata (9,7%) según estimaciones de la Agencia Internacional de Investigaciones en Cáncer. A pesar de que este tumor es uno de los más factibles de prevenir debido a su lenta progresión y a las altas posibilidades de curación cuando se detecta en una fase temprana, causó 7603 muertes en 2015 en Argentina. Ello representa el 12% del total de muertes por tumores malignos, ubicándose en el segundo lugar luego del cáncer de pulmón (14,9%) y precediendo al cáncer de mama (9%) (INC, 2018).

Con el objeto de suavizar los patrones de comportamiento sobre la mortalidad por cáncer colorrectal se recurre a la estimación de las RME mediante modelos. Los modelos lineales generalizados mixtos (MLGM) pueden considerar efectos aleatorios apropiados para analizar los casos de mortalidad por enfermedades correlacionadas espacialmente y obtener de esta manera inferencias válidas (Rasmussen, 2004; Stroup, 2013; Agresti, 2015).

En este trabajo se aborda, entonces, el estudio de la mortalidad por cáncer colorrectal en Argentina en el año 2016. Se propone el ajuste de modelos para datos de conteo que tengan en cuenta la correlación espacial, para finalmente lograr una mejor visualización de las variaciones provinciales en las RME mediante el respectivo georreferenciamiento.

En las secciones siguientes se presenta la información disponible, se describen los procedimientos de estimación de las RME y finalmente se presentan los resultados alcanzados.

Los datos

A partir de la información provista por la Dirección de Estadísticas e Información en Salud (DEIS) sobre la mortalidad en Argentina según causa de muerte, se obtiene el número de muertes por cáncer colorrectal, de ambos sexos, registrados en el año 2016.

La información correspondiente al total de la población por grupos quinquenales de edad se obtiene a partir de las estimaciones al 1° de julio del año 2016, provista por el Instituto Nacional de Estadísticas y Censos (INDEC). Estas estimaciones fueron proyecciones para el 2016 basadas en datos del censo de población, hogares y vivienda de 2010. Se cuenta además con información censal sobre algunas características sociodemográficas de las provincias argentinas, a saber: densidad poblacional, porcentaje de hogares con necesidades básicas insatisfechas (NBI), porcentaje de población rural y de población analfabeta (INDEC, 2018).

Los mapas de Argentina por provincia se obtienen del Instituto Geográfico Nacional de la República Argentina (IGN, 2018).

Estimación de las razones de mortalidad estandarizadas

Las RME se definen como el cociente entre el número de muertes observadas y el número de muertes esperadas en cada área en estudio, en este caso las provincias argentinas. Para la obtención de los valores esperados se tiene en cuenta el total poblacional en cada grupo etario



de las provincias a comparar y las tasas específicas por grupo de edad de Argentina, tomada como población estándar. Dichas tasas se calcularon a partir del número total de casos observados y la población argentina estimada al 1° de julio de 2016 para cada grupo quinquenal de edad.

Así, para cada provincia se obtiene una RME de la forma:

$$R_i = \frac{Y_i}{E_i} \quad i = 1, \dots, 24$$

donde Y_i y E_i representan el número de muertes observadas y esperadas por cáncer colorrectal en la i -ésima provincia, respectivamente.

Con el objeto de controlar la heterogeneidad del tamaño poblacional de las provincias de Argentina se obtiene una estimación suavizada de las RME a través del ajuste de diferentes modelos para datos de conteos.

En general, el conteo de enfermedades en áreas que se encuentran próximas geográficamente muestra dependencia espacial, por lo que se recurre a los MLGM, que pueden acomodar los efectos aleatorios espaciales y proporcionar un medio flexible que permite analizar casos de enfermedades correlacionadas espacialmente (Breslow y Clayton, 1993).

Antes de especificar los MLGM considerados se presentan los métodos utilizados para describir la correlación espacial.

Con el fin de verificar si las unidades se encuentran distribuidas aleatoriamente en el espacio según los valores de RME, se recurre al índice de asociación global de Moran. Este índice resume la intensidad y la dirección de la dependencia espacial y varía entre -1 y 1. Un coeficiente mayor a su valor esperado, indica autocorrelación espacial positiva entre las unidades, mientras que un coeficiente menor que su valor esperado manifiesta autocorrelación espacial negativa. Un valor cercano a su esperanza, representa ausencia de autocorrelación (Moran, 1950).

Para la caracterización espacial de la información se asocia el valor R_i con el centroide s_i de cada provincia. La estructura de la correlación espacial se modela considerando al valor observado como una realización espacial de una variable aleatoria. En general, se asume que el comportamiento de la variable responde a un proceso estocástico estacionario de segundo orden. Cuando la covariancia depende sólo de la distancia que separa las unidades y no de la orientación se dice que el proceso es isotrópico y resulta que $Cov(R_i, R_{i'}) = C(h)$ siendo $h = \text{dist}(i, i')$.

Otra forma de expresar la correlación entre las unidades es a través del semivariograma, el cual se define como:

$$\gamma(R_i, R_{i'}) = \frac{1}{2} \text{Var}(R_i - R_{i'}).$$

Cuando el proceso es estacionario de segundo orden e isotrópico, la expresión del semivariograma es:

$$\gamma(R_i, R_{i'}) = \gamma(h) = \text{Var}(R_i) - \text{Cov}(R_i, R_{i'}).$$

En estos modelos de semivariograma se distinguen los parámetros de meseta (valor límite máximo constante que alcanza el semivariograma), rango (valor de la distancia a partir de la cual se alcanza la meseta) y efecto pepita (discontinuidad que presenta el semivariograma en el origen), los cuales permiten describir la correlación espacial.



Una primera aproximación al modelo de semivariograma es su estimación por el método de momentos, conocido como semivariograma empírico. Esta representación es útil para seleccionar un modelo de semivariograma teórico, entre los que se distinguen el exponencial, esférico y gaussiano. Los parámetros de estos modelos se estiman por el método de mínimos cuadrados ponderados y para la elección del más adecuado se utiliza el Criterio de Información de Akaike (AIC).

El caso particular del modelo de semivariograma gaussiano (sin efecto pepita) tiene la forma:

$$\gamma(R_i, R_{i'}) = \sigma^2 \left[1 - e^{-\left(\frac{h}{a_0}\right)^2} \right] \quad \forall h > 0,$$

donde σ^2 es el parámetro meseta, a_0 es el rango y $h = \text{dist}(i, i')$.

La meseta se alcanza de forma asintótica, por lo que desde un punto vista estricto no tiene rango. El llamado rango efectivo se define como la distancia para la cual el valor del semivariograma alcanza el 95% del valor de la meseta y resulta igual a $\sqrt{3}a_0$ (Cressie, 1993; Ambrosio, 2000).

Modelización estadística

Bajo el supuesto de que Y_i : número de muertes observadas en la provincia i -ésima, se distribuye Poisson de parámetro μ_i , el MLGM más simple se formaliza:

$$\ln(\mu_i) = \ln(E_i) + \beta_0 + u_i \quad i = 1, \dots, 24$$

donde u_i es el efecto aleatorio asociado a cada provincia. Este modelo supone efectos aleatorios independientes con distribución $N(0; \sigma^2)$ (Agresti, 2015).

Un MLGM que tiene en cuenta la dependencia espacial entre las unidades se obtiene a través de la especificación de una estructura de correlación para los efectos aleatorios (Waller y Gotway, 2004; Rasmussen, 2004; Stroup, 2013). En el caso particular de que esa estructura responda a un modelo de semivariograma gaussiano, los elementos de la matriz de variancias y covariancias de los efectos aleatorios están dados por:

- $\text{var}[u_i] = \sigma^2$
- $\text{cov}[u_i; u_{i'}] = \sigma^2 e^{-\left(\frac{h}{a_0}\right)^2} \quad \forall h > 0.$

Los modelos presentados pueden además incluir efectos fijos asociados a variables explicativas que reflejen características sociodemográficas de las provincias.

La elección del modelo más adecuado se basa nuevamente en el valor del AIC.

La metodología descripta se implementa utilizando el EPIDAT, el entorno computacional R y los procedimientos VARIOGRAM y GLIMMIX de SAS.

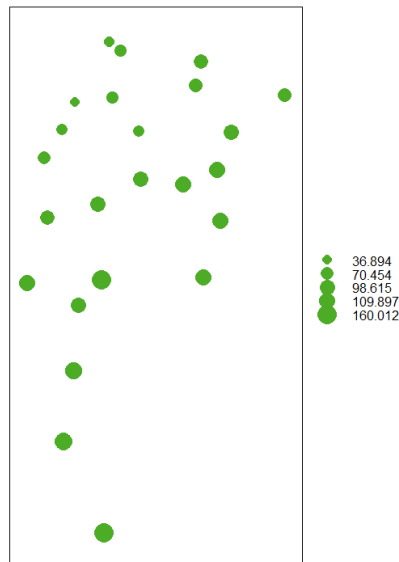
Resultados

En base a la información provista por la DEIS y el INDEC se pudieron construir las RME para las provincias de Argentina, utilizando como estándar las tasas de mortalidad específicas por grupo etario del país. La menor RME observada igual a 36,89% corresponde a la provincia de Catamarca, indicando que las muertes observadas son inferiores a las esperadas si sobre ella



operaran las tasas de mortalidad de Argentina. Por el contrario, la más alta, corresponde a La Pampa (160,01%) mostrando que en esta provincia el riesgo es 60% mayor del que se esperaría si tuviese la experiencia de mortalidad imperante en el país. Estas RME se pueden observar en la Figura 1, donde están representadas con círculos de menor o mayor superficie según su valor. Vale mencionar que la representación de las RME de todas las provincias en este tipo de gráfica reproduce aproximadamente la forma de la región a la que pertenecen. Por último, se observa que círculos de superficie pequeña se encuentran cercanos a otros de iguales características y lo mismo sucede con los círculos de superficies mayores, sugiriendo correlación espacial.

Figura 1: Mapa de las RME observadas

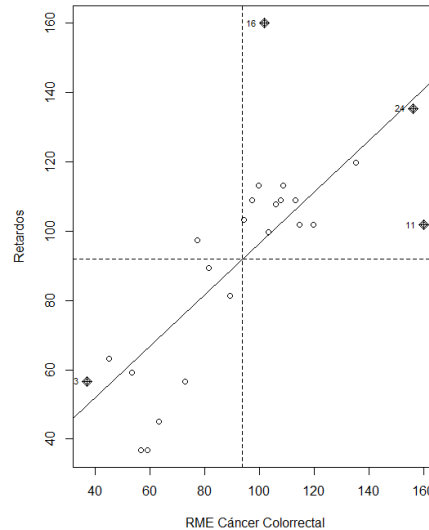


Una forma de verificar dicha correlación es mediante el índice de asociación global de Moran, cuyo valor resulta igual a 0,74, resultando estadísticamente significativo ($p=0,0004$); por lo que se puede asumir que las RME observadas no ocurren aleatoriamente en la Argentina.

El gráfico de dispersión de Moran muestra que la mayoría de los puntos se ubican en los cuadrantes I y III, corroborando la correlación espacial positiva medida a través del índice. Es decir, las provincias que presentan RME grandes se encuentran rodeadas de provincias con similares características y lo mismo sucede con provincias con RME pequeñas (Figura 2).



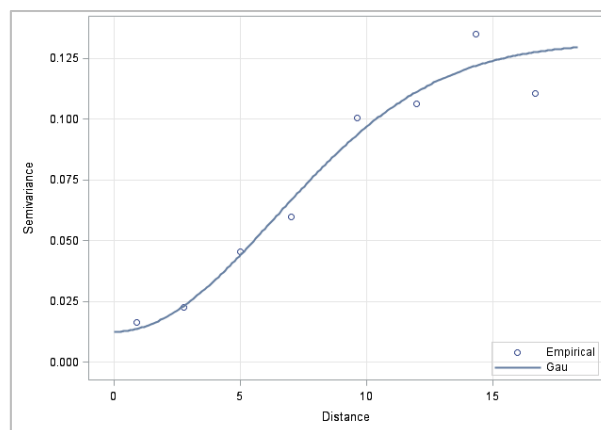
Figura 2: Diagrama de dispersión de Moran



Para el mapeo de las RME se recurre al uso de estimaciones basadas en diferentes MLGM; el más simple con sólo interceptos aleatorios independientes, seguido por el que incorpora la correlación espacial y por último los que tienen en cuenta la información complementaria proveniente de registros oficiales referida a la densidad poblacional de las provincias, el porcentaje de hogares con NBI, el porcentaje de población rural y de población analfabeta de las provincias.

En lo que se refiere al modelo con efectos aleatorios con dependencia espacial, el semivariograma que mejor ajusta los datos es el correspondiente al modelo gaussiano (Figura 3).

Figura 3: Semivariogramas empírico y gaussiano para las RME



En cuanto a las características sociodemográficas consideradas, la covariable densidad poblacional no resultó significativa en ninguno de los modelos planteados y el porcentaje de población analfabeta en general tampoco resultó significativo. En cambio el porcentaje de hogares con NBI y el porcentaje de población rural mostraron efecto significativo.



En la Tabla 1 se aprecia la notoria disminución en el valor del AIC en los modelos que consideran la dependencia geográfica entre las unidades y más aún cuando se considera también el efecto de factores fijos que caracterizan a las provincias.

Tabla 1: Medidas de AIC para diferentes modelos

Modelo	AIC
1. Efectos aleatorios independientes	255,78
2. Efectos aleatorios con correlación espacial	237,39
3. Efectos aleatorios independientes + efecto fijo NBI + efecto fijo Rural	245,16
4. Efectos aleatorios con correlación espacial + efecto fijo NBI + efecto fijo Rural	231,23
5. Efectos aleatorios con correlación espacial + efecto fijo NBI	231,71

Los coeficientes de correlación entre las RME observadas y las estimadas por los dos últimos modelos resultan muy altas: 0,958 para el modelo 4 y 0,954 para el modelo 5. En base a estos resultados hallados se elige el modelo más simple, el modelo 5.

En la Tabla 2 se puede apreciar la significación estadística de los parámetros de covariancia asociados al modelo de semivariograma gaussiano elegido. Y el coeficiente negativo asociado al porcentaje de hogares con NBI muestra que se esperan RME menores en aquellas provincias con más hogares con NBI.

Tabla 2: Estimaciones del modelo elegido

	Coeficiente (error estándar)	Probabilidad asociada
Efecto fijo NBI	-0,03 (0,01)	0,0049
Parámetros del modelo gaussiano		<0,001
Meseta	0,07 (0,05)	
Rango	8,36 (2,01)	

En la Figura 4 se presenta en diagrama de caja de las RME estimadas bajo modelo elegido y el correspondiente a las observadas. Como era de esperar, se puede observar una mayor dispersión en las RME observadas que en las estimadas. El mínimo valor aumenta a 54,32% el máximo disminuye a 137,25%.

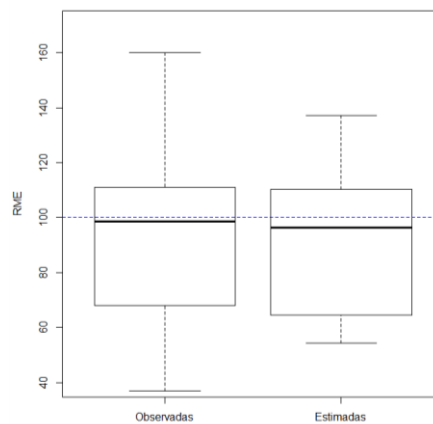
Se destaca el valor 100% ya que las RME definidas como el cociente entre el número de muertes observadas en cada provincia y el número de muertes esperadas, expresadas en porcenta-



je, se utiliza para identificar aquéllas provincias en situación de riesgo respecto a la mortalidad por cáncer colorrectal (con valores de las RME mayores a 100 %).

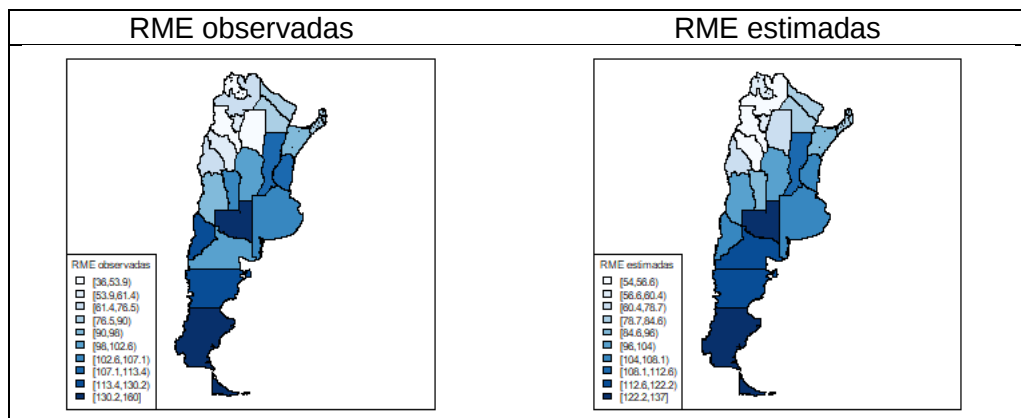
Ambos diagramas de caja muestran aspectos semejantes en lo que se refiere a: medianas con valores aproximadamente iguales a 98% y valores más concentrados en el 25% de las RME más pequeñas que en el 25% de las más elevadas. Sin embargo, es notorio que dentro del 50% de los valores centrales, la concentración es mucho mayor para el 25% de las RME con valores más altos y mucho más dispersos en el 25% de los valores más bajos.

Figura 4: Diagrama de caja para las RME observadas y estimadas



Por último, se georreferencian las RME observadas y las estimadas mostrando en diferentes tonalidades los respectivos deciles donde los colores más intensos corresponden a los valores mayores de RME (Figura 5).

Figura 5: Mapa de percentiles de las RME observadas y estimadas



Las RME estimadas a partir del modelo elegido suavizan los respectivos valores observados presentando un patrón espacial más homogéneo por regiones geográficas. Claramente se observa que la zona del noroeste es la de menor riesgo, aumentando paulatinamente hacia sur con la situación de mayor riesgo en la Patagonia. Cabe mencionar que la alta RME estimada



para La Pampa interrumpe de algún modo el patrón geográfico descripto.

Discusión

El estudio de las variaciones regionales de la mortalidad de una enfermedad, tal como el cáncer colorrectal en la Argentina, se abordó a través del uso de modelos estadísticos de manera de tener en cuenta ciertos factores de reconocida importancia: la diferencia de estructura por edad de las poblaciones de cada área, la heterogeneidad del tamaño poblacional de dichas áreas y la dependencia espacial (Breslow y Clayton, 1993; Clayton y Kaldor, 1987).

La consideración en el modelo de la posible semejanza de las RME entre provincias cercanas mediante la correspondiente modelización espacial resultó adecuada, como así también la inclusión de características sociodemográficas.

El posterior mapeo de las RME permitió una rápida visualización de las variaciones regionales identificando al sur del país como el área de mayor riesgo (con RME mayores a 100%) en relación a la mortalidad por este tipo de cáncer.

Este tipo de estudios sobre la distribución geográfica de la RME por causa se considera importante ya que proporciona información valiosa para los epidemiólogos y tomadores de decisiones en el área de Salud Pública, otorgando herramientas objetivas para las intervenciones en salud. Para una mejor comprensión de esta problemática sería apropiado complementar los estudios con los correspondientes a la incidencia de las enfermedades y a la accesibilidad a los sistemas de salud.

Por último, en lo que se refiere a metodología estadística, se está trabajando en la puesta a prueba de otros modelos que contemplen dependencia espacial (Biggeri et al, 2000; Skrondal A., Rabe-Hesketh S, 2004; Mohebbi et al, 2011a; Mohebbi et al, 2011b) y en un nivel de desagregación de la información a nivel departamento con el objeto de detectar áreas de riesgo que no estén sujetas a las grandes áreas definidas por la división política en provincias. Esta desagregación trae aparejado el problema de la existencia de conteos nulos o bajos, situación en la cual será conveniente considerar un enfoque de estimación bayesiano (Ugarte et al, 2009; Lawson, 2013).

REFERENCIAS BIBLIOGRÁFICAS

- Ambrosio, L. (2000). *Estadística Espacial*. E.T.S.I.A. Madrid.
- Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Biggeri A., Marchi M., Lagazio C., Boehning D., Martuzzi M. (2000) Non-parametric Maximum Likelihood Estimators for Disease Mapping, *Statistics in Medicine*, 19, 2539–2554
- Breslow, N.; Clayton, D. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88 (421): 9-25.
- Clayton, D.; Kaldor, J. (1987) Empirical Bayes estimates of age standardized relative risks for use in disease mapping. *Biometrics*, 43: 671-681.
- Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, N. Y. 990p.
- EPIDAT Versión 4.2. (2016). Dirección Xeral de Saúde Pública (Xunta de Galicia).



IGN. Instituto Geográfico Nacional de la República Argentina. <http://www.ign.gob.ar/NuestrasActividades/InformacionGeoespacial/CapasSIG> . Fecha de consulta: 09/2018.

INDEC. Censo Nacional de Población, Hogares y Viviendas del año 2010. www.indec.mecon.gov.ar. Fecha de consulta: 11/2018.

INC. Instituto Nacional del Cáncer. Cáncer Colorrectal. Disponible en <http://www.msal.gov.ar/inc/acerca-del-cancer/cancer-colorrectal/>. Fecha de consulta: 11/2018.

Lawson, A. B. (2013). *Bayesian disease mapping. Hierarchical modeling in spatial epidemiology*. 2º ed. Chapman & Hall.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37:17-33.

Mohebbi, M.; Wolfe, R.; Jolley, D. (2011a) A Poisson regression approach for modelling spatial autocorrelation between geographically referenced observations. *BMC Medical Research Methodology*, 11:133. <http://www.biomedcentral.com/1471-2288/11/133>.

Mohebbi, M.; Wolfe, R.; Jolley, D.; Forbes, A. B.; Mahmoodi, M.; Burton, R. C. (2011b). The spatial distribution of esophageal and gastric cancer in Caspian region of Iran: An ecological analysis of diet and socio-economic influences. *International Journal of Health Geographics*, 10:13.

Preston, S; Heuveline, P; Guillot, M. (2003). *Demography. Measuring and modeling population processes*. Blackwell Publishers.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Recuperado de <https://www.R-project.org/>

Rasmussen, S. (2004). Modelling of discrete spatial variation in epidemiology with SAS using GLIMMIX. *Elsevier. Computer Methods and Programs in Biomedicine*. 76: 83-89.

SAS/STAT 9.4. (2017). Institute Inc. Cary, N.C, USA.

Skrondal A., Rabe-Hesketh S. (2004). *Generalized latent variable modeling: multilevel, longitudinal and structural equation models*. Boca Raton (USA): Chapman & Hall/CRC.

Stroup, W. W. (2013). *Generalized Linear Mixed Models. Modern concepts, methods and applications*. Chapman & Hall.

Szklo, M.; Nieto, J. (2000). *Epidemiology beyond the basics*. Geithersburg (MD). Aspen Publishers.

Ugarte, M. D.; Goicoa, T.; Militino, A. F. (2010). Evaluating the performance of spatio-temporal Bayesian models in disease mapping. *Environmetrics*. 21: 270-289.

Waller, L., Gotway, C. (2004). *Applied spatial statistics for public health data*. Wiley; Hoboken.

Agradecimientos

Se agradece al Lic. Carlos Guevel, Director de la DEIS, Ministerio de Salud y Desarrollo Social de la Nación, por el aporte de información utilizada en este trabajo.