UNIVERSIDAD NACIONAL DE CÓRDOBA

"MODELOS EXTENDIDOS PARA EL ANÁLISIS ESPACIAL EN EPIDEMIOLOGÍA DEL CÁNCER"

Trabajo de Tesis para optar al
Título de Magister en Estadística Aplicada

Contadora Mariana Verónica Gonzalez

CÓRDOBA, REPÚBLICA ARGENTINA
AÑO 2015

COMISIÓN ASESORA DE TESIS

Director

Dra. María del Pilar Díaz

Co-director

Dra. Sonia Alejandra Pou

Miembros

Dra. Margarita Díaz

Dra. Sonia Muñoz

Dr. Alberto Osella

Fecha de aprobación de tesis: 16/12/15

DEDICATORIA

A mis hijos,
Juan Ignacio y Santiago Tomás,
que hacen que todos los esfuerzos valgan la pena.
Los amo profundamente.

AGRADECIMIENTOS

A mi directora, Pilar Díaz y a mi co-directora, Sonia Pou, que me acompañaron en este proceso y compartieron su valiosa experiencia conmigo.

A Santi y Juani que me dieron la fortaleza para avanzar, a pesar de las dificultades.

A Pablo, guerrero de la luz, que a su manera, me apoyó en este trabajo y en muchos otros.

A mis padres, Martha y Julio, por enseñarme la importancia del esfuerzo y de la perseverancia. Gracias a ellos llegué hoy hasta acá.

A mi hermana Natalia y a mi sobrino Tomás, que recorrieron conmigo este camino.

Al Profesor Roberto Giuliodori, que fue el primero que confió en mí y me permitió iniciarme en el maravilloso mundo de la investigación.

A mi amiga, Patricia Caro, que siempre me impulsó a terminar este trabajo.

Al Dr. Fernando Ferrero, por darme la oportunidad de crecer y aprender a su lado.

A la Facultad de Ciencias Económicas, y en especial, al Instituto de Estadística y Demografía, por brindarme el ámbito para desarrollarme profesional y humanamente.

RESUMEN

El monitoreo de la variación geográfica en la distribución de enfermedades y la investigación para comprender las razones subyacentes a dicha variación son, habitualmente, un punto de partida importante en Epidemiología. Se han identificado muchos factores de riesgo significativos como resultado de los hallazgos en el análisis de patrones geográficos de las enfermedades. Mientras el mapeo de enfermedades infecciosas es una práctica bien establecida, la creación de mapas de enfermedades no transmisibles, como el cáncer, está menos desarrollada. Además, los modelos que incorporan la autocorrelación espacial han sido ampliamente estudiados en lo que se refiere a variables continuas pero cuando se trabaja con datos discretos, como sucede en la Epidemiología espacial, las posibilidades de modelación son más reducidas.

El cáncer muestra variaciones espaciales y el conocimiento de su patrón de ocurrencia es esencial para identificar grupos de población vulnerables, así como para desarrollar políticas de salud adecuadas para la prevención, el seguimiento y el control (Díaz *et al*, 2010). En este trabajo se analiza la distribución geográfica de los casos de mortalidad por cáncer de próstata y mama, a nivel de departamentos en la Provincia de Córdoba, incluyendo un análisis exploratorio espacial con herramientas clásicas e índices específicos y diferentes enfoques de modelación para la obtención de inferencias respecto de la comprensión y cuantificación del fenómeno. En el marco de los Modelos generalizados mixtos y para variables latentes (*GLLAMM*) se ajustaron modelos con efectos aleatorios, que permiten considerar la heterogeneidad no observada entre departamentos e incorporan dicha información a la hora de estimar los efectos de covariables de interés. Para la estimación del riesgo relativo se ajustaron Modelos Poisson con intercepto aleatorio en combinación con predicción Bayesiana empírica, como así también Modelos Poisson-Gamma que incorporan una estructura específica de sobredispersión. Por último, se estimaron Modelos Autorregresivos Simultáneos (*SAR*) que incorporan los valores de las otras áreas para modelar la dependencia espacial.

Los mapas permitieron detectar que las tasas de mortalidad por cáncer de próstata y mama en la Provincia de Córdoba siguen un patrón no aleatorio en su distribución espacial. En relación al cáncer de próstata, existe una concentración de valores elevados del Cociente de Mortalidad Estandarizado (*CME*) hacia el centro-este provincial, mientras que para el cáncer de mama, se pudo identificar un gradiente en el noroeste de la provincia, con *CMEs* en el

quintil superior. Las pruebas de homogeneidad de los riesgos relativos determinaron la existencia de diferencias de significación entre los departamentos de la Provincia en relación a los *CME*. Las distribución espacial de los riesgos relativos estimados con los modelos Poisson con intercepto aleatorio y Poisson- Gamma resultó muy similar para ambas estrategias de modelación, permitiendo detectar una zona de riesgo al este de la Provincia, tanto para cáncer de mama como para el cáncer de próstata. En una segunda etapa se incorporó el efecto de covariables socieconómicas, disponibles para el año 2001 y para toda la provincia, las que no resultaron estadísticamente significativas en el modelo Poisson. En el modelo *SAR*, considerando como variable dependiente la estandarización de los *CME*s, para el cáncer de próstata, resultaron significativas y con coeficiente positivo el porcentaje de la población total sin cobertura de salud y el porcentaje de la población total desocupada.

Las pruebas bondad de ajuste de los modelos revelaron la superioridad del modelo Poisson con intercepto aleatorio respecto del modelo Poisson clásico. El modelo autorregresivo, por su parte, resultó el proporciona un mejor ajuste a los datos, cuando se incorporan covariables en el análisis.

Palabras clave: ***cáncer, distribución espacial, autocorrelación espacial, Poisson, GLLAMM.***

"La primera ley de la Geografía es: cualquier cosa está relacionada con cualquier otra pero las cosas más cercanas están más relacionadas que las más distantes"
(Tobler, 1970).

CAPÍTULO 1: MARCO TEÓRICO

1.1. Conceptos Generales

La Estadística espacial se ocupa de la exploración, la descripción, la visualización y el análisis de los datos considerando sus características de distribución en el espacio, que suelen expresarse a través del uso de coordenadas geográficas. Los campos de aplicación son muy variados, incluyendo el estudio de todos aquellos fenómenos que, pudiendo ser analizados desde el punto de vista de la Estadística, poseen una referencia espacial en sus datos (Sáez y Saurina, 2007).

Los datos espaciales son medidas u observaciones que tienen asociada una localización específica y, de forma general, pueden definirse como datos georeferenciados (López Hernández *et al*, 2000). La localización que tiene asociada cada observación puede ser un punto cualquiera de una determinada superficie, o bien estar asociada a un área, dentro de una superficie, sobre la que se ha realizado una partición. En el primer caso se expresan cada uno de los datos obtenidos junto con los valores de la latitud y la longitud geográfica de la localización, como por ejemplo, las coordenadas de referencia de la localización de un centro contaminante. Los datos referidos a un área son observaciones en las que se obtiene un valor agregado para el conjunto de la superficie considerada (departamento, municipio, fracción censal). Un ejemplo de este tipo de datos puede ser la incidencia (número de casos nuevos de una enfermedad en una población determinada y en un período) o mortalidad de cierta enfermedad cuantificada por localidades o municipios. En función del tipo de dato se determinan los modelos a aplicar, lo que implica utilizar métodos estadísticos diferentes.

Los datos espaciales presentan características particulares que condicionan su tratamiento estadístico. Fundamentalmente, los datos que tienen una localización espacial/geográfica asociada, pueden tener propiedades que se refieren a su localización individual y también a los datos que los rodean. En efecto, en el caso de la estadística espacial la cuestión que se plantea es la pérdida de la condición de independencia de las observaciones tomadas en una determinada área. Las observaciones muestran cierta forma de correlación, en sus diferentes niveles, basada en la localización, por lo que no pueden ser modelados con los métodos estadísticos que suponen independencia.

El término Epidemiología espacial se emplea para describir estudios sobre los determinantes y la prevención de las enfermedades utilizando diferentes perspectivas de análisis en las que la localización de los eventos es un componente fundamental (Thomas, 1990). Estudia la ocurrencia de eventos de salud-enfermedad (ocurrencia de enfermedades, defunciones) en localizaciones espaciales y sus factores condicionantes. En su forma más simple, se refiere al uso e interpretación de mapas de localización de casos de enfermedad. En este sentido, involucra todas las cuestiones asociadas a la producción de mapas de enfermedad o exposición y al análisis estadístico de estos datos.

Una característica, asociada específicamente a la Epidemiología espacial, es que los datos son frecuentemente discretos. A diferencia de otras áreas del análisis espacial, que trabajan con datos continuos (Geoestadística), los datos encontrados en Epidemiología espacial toman frecuentemente la forma de conteos de casos de enfermedad o muerte dentro de regiones (Lawson, 2001).

El desarrollo de la Epidemiología espacial se remonta a principios del siglo XIX, en el que se realizaron mapas de la ocurrencia de enfermedades en diferentes países con el objetivo de caracterizar la extensión y las posibles causas de brotes de enfermedades infecciosas tales como la fiebre amarilla y el cólera (Walter, 2000).

Con posterioridad, la Epidemiología espacial creció en complejidad, sofisticación y utilidad (Elliott y Wartenberg, 2004), extiendo la amplia tradición de estudios ecológicos, que utilizan las explicaciones de la distribución de enfermedades en diferentes localizaciones geográficas, para comprender mejor la etiología de la enfermedad. Se distingue entre estudios a nivel individual y a nivel ecológico, según cuál sea la unidad de análisis. En los diseños individuales se disponen de observaciones a nivel individual tanto de la variable respuesta (a explicar) como de las posibles variables explicativas. En los diseños ecológicos no se dispone de observaciones individuales, al menos en uno de los dos casos. Por ejemplo, se pueden disponer de datos de ocurrencia de cáncer a nivel individual pero no de posibles factores explicativos de esta ocurrencia a ese nivel (Doll, 1980; Elliott y Wartenberg, 2004).

Los avances en la disponibilidad de datos y en los métodos analíticos para tratarlos han proporcionado nuevas oportunidades para extender las investigaciones epidemiológicas tradicionales, a escala nacional o regional, hasta el estudio de las variaciones en las enfermedades a nivel local, o de área pequeña (Walter, 2000). Sería deseable que tales investigaciones contemplaran además, factores de riesgo para la salud con relevancia a nivel local, tales como exposiciones ambientales, distribución local de condiciones socioeconómicas y los hábitos y estilos de vida (Elliott y Wartenberg, 2004).

Las técnicas de representación geográfica de datos en Epidemiología suelen ser un paso previo para la visualización de la estructura espacial y constituyen una herramienta potente para el análisis exploratorio de los datos antes de proceder a su modelización. Además, en las últimas décadas, el desarrollo de técnicas estadísticas avanzadas y la potencia de computación han promovido el desarrollo de modelos con diferentes enfoques como el frecuencista, bayesiano y multinivel (Lawson *et al.*, 2003).

El estudio de la distribución geográfica de enfermedades puede ser abordada desde tres grandes perspectivas (Lawson *et al.*, 2003), que serán desarrolladas en este trabajo desde el punto de vista teórico y de la modelación:

• Estimación del riesgo relativo en áreas pequeñas o *disease mapping.*

• Estudios de asociación geográfica o *ecological analysis.*

• Aglomeraciones de casos o *disease clustering.*

La primera se refiere al uso de modelos para describir la distribución general de la enfermedad en el mapa. Frecuentemente, el objetivo es estimar el verdadero riesgo relativo de la enfermedad de interés en el área geográfica de estudio. Tiene por objetivo proveer un mapa de la incidencia o mortalidad por enfermedad "limpio", habiendo removido todos los efectos aleatorios (*noise*), de manera de proporcionar una estimación precisa de la tasa subyacente en diferentes áreas (Lawson, 2001). Los mapas de enfermedades proporcionan un rápido resumen visual de información geográfica compleja y permiten identificar patrones en los datos que de otro modo podrían pasar inadvertidos (Elliott y Wartenberg, 2004). De hecho, los mapas se utilizan con propósitos descriptivos, con el objetivo de generar hipótesis etiológicas; para la vigilancia epidemiológica, a fin de detectar áreas con un aparente mayor riesgo; y como ayuda en la definición de políticas de salud y de asignación de recursos. Como ejemplo de este abordaje, se menciona el (ya) clásico ejemplo del análisis de cáncer de labio en 56 condados escoses, para el período 1975-1980, a fin de identificar qué condados tienen mayor riesgo (Rabe-Hesketh y Skrondal, 2012), el cual lo retoman otros autores como Clayton y Kaldor (1987), Breslow y Clayton (1993) y Leyland (2001). Como se explicará más adelante, los Cocientes de Mortalidad Estandarizados (*CME*) crudos se calculan como el cociente entre el número observado de casos para una unidad de análisis (condado en el ejemplo citado) y el número de casos esperados. Cuando el *CME* supera el valor 100, la unidad (condado) tiene más casos que los esperados, dada su distribución por edad.

Los estudios de asociación geográfica, por su parte, son de gran importancia dentro de la investigación epidemiológica, enfocados en el análisis de la distribución geográfica de enfermedades en relación con variables explicativas. Su objetivo es examinar, desde un punto de vista geográfico (y ecológico) la relación entre variables de respuesta (incidencia, mortalidad) y factores medioambientales de tipo ecológico (contaminación atmosférica, del agua y del suelo), factores de tipo ocupacional, condiciones socioeconómicas y demográficas, o variables relacionadas con estilos de vida (tales como hábito de consumo de tabaco, alcohol y dieta). Usualmente se llevan a cabo a un nivel espacial agregado, y se ocupan de la incidencia regional comparada con mediciones de los factores explicativos a nivel de región u otro nivel de agregación (Greenberg *et al*. ,1996). Estos estudios son considerados como "generadores de hipótesis" (Elliott y Wartenberg, 2004), por cuanto la unidad de observación la constituyen individuos agrupados geográficamente. Cuando las asociaciones observadas a este nivel (ecológico) no necesariamente se mantienen a nivel individual surge el problema conocido como "falacia ecológica", debida fundamentalmente a que cada grupo no es completamente homogéneo con respecto a las covariables. Lawson *et al.* (2003) presentan varios ejemplos de este tipo de estudios. Uno de ellos corresponde a las muertes por cáncer (neoplasias malignas) en Carolina del Sur en 1999, empleando como covariable la densidad de población, obtenida a partir de información censal. En otro trabajo, los mismos autores, analizan la mortalidad por cáncer de labio en Alemania, para el período 1980-1989, incorporando la variable explicativa porcentaje de población que trabaja en la agricultura, la pesca o la silvicultura, como *proxi* de la exposición al sol.

Los estudios de aglomeraciones de casos son particularmente importantes en la vigilancia de salud pública, donde interesa localizar *clusters* específicos de enfermedades, así como identificar focos contaminantes. En efecto, se postula que la incidencia de muchas enfermedades respiratorias, de la piel y degenerativas puede estar relacionada con la contaminación del medio ambiente y, por lo tanto, cualquier fuente localizada de dicha contaminación podría dar lugar a cambios en la incidencia de estas enfermedades en las comunidades contiguas. En este tipo de análisis, denominado *focussed clustering*, el objetivo es detectar patrones espaciales en los casos cercanos o expuestos al foco, más que realizar una modelización espacial general. Por ejemplo, Lawson (2001) analiza el número de muertes por cáncer respiratorio en el período 1978-1983 en 26 regiones censales de Falkirk, Escocia central, mostrando además, el foco de riesgo, un taller de fundición que puede haber contaminado el aire antes del período de estudio (Figura 1.1). El autor indica que, debido al efecto de la latencia, es posible que haya tenido un impacto en los casos de cáncer respiratorio de los que viven en las zonas adyacentes a la fuente.

Figura 1.1. Mapa de casos de cáncer respiratorio en Falkirk y foco supuesto de peligro (*).

Fuente: Lawson A. (2001). Statistical Methods in Spatial Epidemiology.

Las técnicas de representación geográfica de datos en Epidemiología suelen ser un paso previo para la visualización de la estructura espacial y constituyen una herramienta potente para el análisis exploratorio de los datos, antes de proceder a su modelización. Sin embargo, cuando se dispone de una mayor cantidad de información previa o existen hipótesis más sustantivas en relación con el problema, puede ser ventajoso considerar la construcción de un modelo (Lawson *et al.*, 2003). La representación gráfica de datos de muerte o incidencia varía desde representaciones de conteos dentro de áreas, hasta mapas con estimaciones provenientes de complejos modelos que pretenden describir la estructura de los eventos.

Los mapas de enfermedades proporcionan un rápido resumen visual de información geográfica compleja y permiten identificar patrones en los datos que de otro modo podrían pasar inadvertidos en las presentaciones tabulares (Elliott y Wartenberg, 2004). Las observaciones correspondientes a los casos brutos (datos crudos, tal y como se observan) pueden ser representadas directamente sobre la estructura geográfica de una zona obteniendo un mapa, por ejemplo, de los casos de muerte por una determinada enfermedad. Sin embargo, esta sencilla representación gráfica no siempre proporciona información de interés. Cualquier interpretación de la estructura de los casos incidentes está limitada por la falta de información sobre la distribución espacial de la población que podría estar "en riesgo" de padecer la enfermedad y que, consecuentemente, ha generado esos casos incidentes. A la representación de los casos brutos, se preferirá, en general, la representación de razones que permiten incorporar el efecto de la población a riesgo.

Si bien los mapas proporcionan información visual importante sobre la distribución espacial, deben interpretarse en conjunto con la información estadística (Lawson, 2001). En efecto, es necesario complementar el análisis con la cuantificación de la variabilidad espacial y estimaciones de los riesgos relativos para áreas particulares, mediante diferentes estrategias de modelación.

Desde el campo de la Epidemiología espacial, se desarrolla entonces el trabajo de tesis que se presenta a continuación. Se estudiará la distribución espacial de la mortalidad por cáncer de próstata (en hombres) y mama (en mujeres) en la Provincia de Córdoba, a partir de la información obtenida de la Dirección de Estadísticas e Información de Salud de la Nación Argentina y de la Dirección de Estadísticas y Censos de la Provincia de Córdoba.

A continuación, se expone el marco teórico de referencia de esta investigación, presentando primeramente los aspectos generales referidos al cálculo de tasas y al estudio de la autocorrelación espacial. Luego se exponen las bases conceptuales de los modelos estadísticos que permiten abordar el estudio de la distribución geográfica de enfermedades, destacando su importancia en la investigación epidemiológica. Finalmente, se presenta una breve introducción a la epidemiología del cáncer, con especial referencia a los cánceres de interés en este trabajo.

1.2. Cálculo de Tasas

1.2.1. Tasas crudas

Considerar el total de casos de una enfermedad o el número de muertes es útil para determinar la magnitud de un problema en salud pública, pero no resulta aplicable al problema de comparar grupos de población, ni tampoco para comparar tendencias. Si se parte del mismo riesgo de desarrollar un evento, un grupo de población más numeroso desarrollará más eventos que uno pequeño, solo por su tamaño. Por esta razón, para comparar diferencias relativas entre grupos a lo largo del tiempo y, suponiendo que el riesgo es constante en todo el período, se debe hablar del riesgo de la población. Las tasas crudas de mortalidad *(TCMs*) son la medida de riesgo más simple y surgen de dividir el número total de casos observados, en una localidad determinada *i,* por el total de la población en riesgo en dicha localidad. Sin embargo, el uso directo de las tasas crudas no permite la comparación entre distintas localidades, ya que las diferencias observadas entre ellas pueden ser debidas a factores que no hayan sido tenidos en cuenta.

1.2.2. Tasas de mortalidad estandarizadas

En Epidemiología, la mayoría de las tasas (incidencia, prevalencia y mortalidad) son fuertemente dependientes del grupo etario, ya que, en algunos casos, el riesgo de un evento (enfermar o morir) aumenta con la edad y, en otras situaciones, disminuye. La comparación de tasas crudas, a través del tiempo o entre poblaciones, puede ser engañosa si la composición por edades subyacente difiere entre las poblaciones que se comparan. Para superar este inconveniente es posible aplicar un proceso de normalización o ajuste, comúnmente denominado estandarización.

Existen varias técnicas para ajustar las tasas específicas por edad. Entre ellas se encuentran la estandarización directa e indirecta (Wolfenden, 1923), la media geométrica (Schöen, 1970), las tasas de mortalidad promedio equivalentes (Hill, 1977), las tasas de la tabla de vida, el índice de Yerushalmy (Yerushalmy, 1951), las tasas de mortalidad acumulativas (Breslow y Day, 1981), las probabilidades absolutas de muerte y el índice de mortalidad comparativa (Peto *et al*., 1994; Breslow y Day, 1987; Esteve *et al*., 1994). Sin embargo, con la creciente disponibilidad de tasas específicas por edad, empleando la población estándar mundial como referencia (*World Standard Population)*, el uso de la normalización directa se ha convertido en la técnica predominante en la mayoría de las aplicaciones epidemiológicas (Ahmad *et al.,* 2001). Éste método produce una tasa de mortalidad ajustada, que resulta de un promedio ponderado de las tasas específicas por edad, para cada una de las poblaciones que se desean comparar, en relación a una población estándar. La distribución, por grupos edad, de la población de las localidades geográficas que se quieren comparar se hace corresponder con la de un área estándar o de referencia (Sáez y Saurina Canals, 2007).

La tasa de mortalidad estandarizada por el método directo, para una localidad *i-ésima* (TME_{di}), está dada por:

$$TME_{di} = \sum_{j=1}^{k} TCM_{ij} \left(\frac{n_{js}}{\sum_{j=1}^{k} n_{js}} \right),$$

donde n_{js} es la población en el grupo de edad *j* de la población estándar y TCM_{ij} es la tasa cruda de mortalidad en la localidad *i* y para el grupo de edad *j* (*j* = 1, 2,…,*k*). Las expresiones obtenidas en cada una de las localidades no son más que una media ponderada de las razones crudas para cada grupo de edad, utilizando las mismas ponderaciones $w_{js} = \left(\frac{n_{js}}{\sum_{j=1}^{k} n_{js}} \right)$ en todas las localidades.

Es evidente que en este tipo de estandarización, la elección del área de referencia es crucial y que los resultados dependen claramente de la opción realizada. En este trabajo se utilizó como población estándar la propuesta por la Organización Mundial de la Salud (OMS) en el año 2001, especialmente definida para reflejar la estructura promedio de edad de la población mundial y proyectada hasta el año 2025, en base a censos de población y otras fuentes demográficas (Ahmad *et al*., 2001).

La idea de una norma verdaderamente internacional fue sugerida por Ogle en 1892, basada en la experiencia de siete países europeos (Ogle, 1892). Sin embargo, no hay evidencia de su posterior adopción para comparaciones internacionales. Se han propuesto diversos estándares desde entonces, pero ninguno adoptado ampliamente. El debate se ha centrado, en gran medida, en torno a la cuestión de si un estándar es más adecuado que otros. Esta cuestión fue debatida en el año 1965 en una reunión de la Unión Internacional de lucha contra el Cáncer (UICC) realizada en Londres. Se propusieron tres poblaciones estándar, cada una apropiada para determinados tipos de población. La primera, con una alta proporción de jóvenes, considerada adecuada para hacer comparaciones con poblaciones de África (Knowelden et al., 1962). Una segunda ("europea") basada en la experiencia de poblaciones escandinavas, con una proporción relativamente alta de personas mayores y juzgada particularmente adecuada para la comparación dentro de Europa Occidental (Doll y Cook,1967). La tercera fue propuesta por Segi (1960) como un estándar intermedio para el "mundo", basado en la experiencia de 46 países. Esta población estándar fue posteriormente adoptada por la OMS para su uso en el cálculo de las tasas de mortalidad estandarizadas por edad (Tabla 1.1).

La desviación estándar para la TME_{di}, denotada en este trabajo por DS (TME_{di}) es (Breslow y Day, 1987):

$$DS(TME_{di}) = \sqrt{\frac{\sum_{j=1}^{k} w_{js}^2 . TCM_{ij}}{n_{ij}^2}},$$

donde w_{js} son las ponderaciones para el grupo de edad j (j = 1, 2,…,k) de la población estándar, TCM_{ij} es la tasa cruda de mortalidad en la localidad i y para el grupo de edad j y n_{ij} es la población de la localidad i en el grupo de edad j. El cálculo anterior supone la independencia de las TCM para los diferentes grupos etarios, lo que resulta razonable para realizar inferencias (Chiang, 1961; Keyfitz, 1996).

Tabla 1.1. Distribución de la población estándar (en porcentajes).

Grupo de edad	Segi	Europea	OMS (*)
0-4	12,00	8,00	8,86
5-9	10,00	7,00	8,69
10-14	9,00	7,00	8,60
15-19	9,00	7,00	8,47
20-24	8,00	7,00	8,22
25-29	8,00	7,00	7,93
30-34	6,00	7,00	7,61
35-39	6,00	7,00	7,15
40-44	6,00	7,00	6,59
45-49	6,00	7,00	6,04
50-54	5,00	7,00	5,37
55-59	4,00	6,00	4,55
60-64	4,00	5,00	3,72
65-69	3,00	4,00	2,96
70-74	2,00	3,00	2,21
75-79	1,00	2,00	1,52
80-84	0,50	1,00	0,91
85 o más	0,50	1,00	0,63

(*)Para permitir la comparación, el grupo de edad estándar de la OMS de 85 o más es un agregado de los grupos de edad 85-89, 90-94, 95-99 y 100 o más.

Fuente: Age Standardization of Rates: a New WHO Standard (2001).

En general, el intervalo de (100(1 -α)) % de confianza para la TME_{di}, con desviación estándar DS (TME_{di}) puede expresarse como:

$$TME_{di} \pm z_{\alpha/2} \, . \, DS(TME_{di}),$$

donde $z_{\alpha/2}$ es el percentil de la distribución normal estándar tal que $P(z < z_{\alpha/2}) = 1-\alpha/2$ (Armitage and Berry, 1987).

Cuando los conteos específicos por estrato de edad son pocos, como suele suceder en las localidades pequeñas, las estimaciones de las tasas específicas por estrato a través del método directo pueden estar fuertemente influenciadas por variaciones aleatorias. En su lugar, se recomienda aplicar el procedimiento de estandarización indirecta que permite obtener tasas específicas por estrato, a partir de una población estándar de tamaño y relevancia suficiente. Estas tasas se promedian utilizando como ponderaciones los tamaños de cada estrato en población estudiada. Simbólicamente, la tasa de mortalidad estandarizada por el método indirecto, para la localidad *i* (TME_{ii}), estará dada por la siguiente ecuación:

$$TME_{ii} = \sum_{j=1}^{k} TCM_{js} \left(\frac{n_{ij}}{\sum n_{ij}} \right),$$

donde n_{ij} es la población de la localidad i en el grupo de edad j (j = 1, 2,…,k) y TCM_{js} es la tasa cruda de mortalidad para el grupo de edad j en la población estándar.

Así, en la estandarización directa, la población de estudio proporciona las tasas y la población estándar provee los pesos. En la estandarización indirecta, la población estándar proporciona las tasas y la población de estudio proporciona los pesos.

1.2.3. Cociente de mortalidad estandarizado (*CME*)

El primer paso en el mapeo de enfermedades es el cálculo del *CME*, como una estimación del riesgo relativo dentro de cada área (θ_i). El *CME* (o *SMR* por su sigla en inglés) se calcula a partir de los casos observados (o_i) en la localidad i y los casos esperados (e_i) para la misma localidad:

$$CME_i = \frac{o_i}{e_i},$$

Los casos esperados se calculan aplicando el método de estandarización indirecta que toma las tasas específicas por estrato de edad de la población estándar. Estas tasas se promedian utilizando como ponderaciones los tamaños de los estratos de la población estudiada.

Bajo el supuesto anterior, de que el número de muertes en los diferentes estratos de edad no están correlacionados, mientras que los errores de muestreo asociados con la población estándar son insignificantes, puede calcularse el error estándar de los *CMEs* a partir de la expresión:

$$DS(CME_i) = \frac{\sqrt{o_i}}{e_i},$$

donde o_i son los casos observados en la localidad i y e_i los casos esperados para la misma localidad (Breslow y Day, 1987).

Comúnmente, una cuestión estadística de interés es determinar una gama de posibles valores para el verdadero *CME*, razonablemente consistentes con los datos observados. Los límites de confianza exactos para el *CME* se obtienen determinando, en primer lugar, los límites inferior y superior para la media $\mu_i = E(o_i)$ de la distribución de Poisson, correspondiente a los casos observados, que llamaremos L y U respectivamente. Como se explicará más adelante, se asume en general que modelo estadístico adecuado para los casos observados (o_i) en la localidad i es la distribución de Poisson $o_i \sim$Poisson (μ_i) donde $\mu_i = \theta_i e_i$. Luego, los límites inferior (LI) y superior (LS) para el *CME* se calculan como L/e_i y U/e_i respectivamente (Haenszel *et al*., 1962).

1.3. Autocorrelación espacial

1.3.1. Introducción

La importancia del espacio dentro del estudio de variables epidemiológicas es incuestionable. El problema de la dependencia espacial, y en particular de la autocorrelación espacial, ha sido objeto de una gran cantidad de estudios en el área de las Ciencias de la Salud (Haining, 1990; Cressie, 1993; Elliot, 2001; Lawson, 2001). En un sentido estricto, los conceptos de dependencia y autocorrelación espacial no son sinónimos, siendo la autocorrelación espacial una expresión más débil de la dependencia espacial, relativa únicamente a los primeros momentos de la distribución conjunta de una variable. Sin embargo, en lo que sigue, ambos conceptos serán utilizados indistintamente

La Estadística presenta la mayor parte de sus resultados bajo el supuesto de independencia y, cuando se viola esta condición, los resultados que se obtienen ya no son válidos. Un caso típico de violación de la condición de independencia se presenta en el estudio de las series temporales. Los modelos de dependencia temporal han sido objeto de un estudio profundo y se han presentado estructuras que expresan el comportamiento de las variables con respecto al tiempo. Menos desarrollo han tenido los modelos que intentan captar la dependencia de una variable en función de su "entorno", en buena parte debido a la complejidad que involucran las estructuras de dependencia espacial.

En Estadística espacial, el problema que se plantea es la pérdida de la condición de independencia de las observaciones tomadas en una determinada área. La autocorrelación espacial puede definirse como la influencia de valores similares de una variable en espacios geográficos cercanos, es decir, cuando una variable tiende a asumir valores similares en unidades geográficamente cercanas (Anselin, 2001). La propiedad básica de los datos espacialmente autocorrelacionados es que no existe aleatoriedad en el espacio sino que están espacialmente relacionados entre sí (Lee y Wong, 2001).

La autocorrelación espacial puede ser positiva o negativa. Si la presencia de un fenómeno determinado en una región lleva a que se extienda ese mismo fenómeno hacia el resto de regiones que la rodean, favoreciendo así la concentración del mismo, nos encontraremos ante un caso de autocorrelación positiva. Por el contrario, existirá autocorrelación negativa cuando la presencia de un fenómeno en una región impida o dificulte su aparición en las regiones circundantes o contiguas a ella, es decir, cuando unidades geográficas cercanas son netamente más disímiles entre ellas que entre regiones alejadas en el espacio. Por último, cuando la variable analizada se distribuye de forma aleatoria, no existirá autocorrelación espacial (Moreno y Vayá, 2000).

Existen dos causas principales que pueden inducir a la aparición de dependencia espacial: la existencia de errores de medida y de fenómenos de interacción espacial. Los errores de medida pueden surgir, entre otros aspectos, como consecuencia de una escasa correspondencia entre la extensión espacial del fenómeno bajo estudio y las unidades espaciales de observación (Haining, 1990). Por ejemplo, podría ocurrir que el fenómeno bajo estudio requiriese de información relativa a unidades para los cuales no se dispone de datos estadísticos, debiendo utilizar en su lugar información relativa a unidades muy desagregadas o unidades con un mayor nivel de agregación. Por otro lado, la existencia de fenómenos de interacción espacial y de jerarquías espaciales también tienen como consecuencia la aparición de un esquema de autocorrelación espacial.

Si se centra el análisis en datos sobre localizaciones irregulares, a cada dato es posible asociarle una coordenada correspondiente a la latitud y longitud de un punto que sea representativo de la zona, elegido a criterio del investigador. Esta superficie debe dotarse de una estructura de vecindades que establezca las relaciones espaciales de los datos.

Supondremos una superficie que ha sido particionada en n zonas y definiremos a n_j como el conjunto formado por todas las zonas que son vecinas de la zona j, esto es $n_j = \{k; k \text{ es vecino de } j\}$. En estas zonas existirá una estructura de vecindades, determinada a través la una matriz $W=\{w_{ij}\}$ de orden n x n. La matriz W define si dos zonas son vecinas, de tal manera que si $w_{ij} = 0$ la zona i no es vecina de la zona j y si $w_{ij} \neq 0$ el coeficiente mide la intensidad de la relación. Respecto a cómo construir la matriz W, cabe destacar que no existe una definición unánimemente aceptada, si bien se debe cumplir que sus pesos sean no negativos y finitos (Anselin, 1980). La matriz W no tiene necesariamente que ser simétrica, pudiendo plantearse relaciones unidireccionales o bidireccionales de diferente intensidad. (López Hernández y Palacios Sanchez, 2000).

Una de las razones más importantes por las que se debe detectar la autocorrelación espacial en un conjunto de observaciones es por las consecuencias que tiene sobre los procesos de inferencia en modelos de regresión lineal como estimación ineficiente de los coeficientes del modelo, sesgo o subestimación de la varianza residual, R^2 sobreestimados, valores erróneos de t y F, inconsistencia. Algo similar ocurre en los modelos lineales generalizados.

1.3.2. Matriz de vecinos *W*

Es posible detectar una cierta similitud entre los conceptos de autocorrelación espacial y temporal en la medida en que, en ambos casos, se produce un incumplimiento de la hipótesis de independencia entre las observaciones muestrales. Sin embargo, existe una importante diferencia entre ellas. La dependencia temporal es únicamente unidireccional, el pasado explica el presente. La dependencia espacial, en cambio, es multidireccional, una región puede estar afectada no sólo por otra contigua a ella sino por otras que la rodean, al igual que ella puede influir sobre aquéllas (Anselin y Griffith, 1988). La solución al problema de la multidireccionalidad en el contexto espacial puede resolverse mediante la definición de la denominada matriz *W* de contactos o de pesos espaciales. Esta matriz permite incorporar el espacio dentro del análisis y ocupa una posición central dado que, esencialmente, define el conjunto de vecinos para cada localización. El primer paso para cuantificar la estructura de dependencia espacial en un conjunto de datos es definir, para el conjunto de puntos o áreas, la relación espacial existen entre ellos (Haining, 2003)

Como se indicó en el apartado anterior, suponiendo que el tamaño muestral es igual a *n*, la matriz *W* será de orden $n \times n$ y puede representarse de la siguiente manera:

$$W = \begin{bmatrix} 0 & w_{1,2} & \cdots & w_{1,j} & \dots & w_{1,n} \\ w_{2,1} & 0 & \cdots & w_{2,j} & \dots & w_{2,n} \\ \vdots & \vdots & \ddots & \dots & \dots & \dots \\ w_{i,1} & w_{i,2} & \cdots & 0 & \dots & w_{i,n} \\ \vdots & \vdots & \vdots & \dots & \dots & \dots \\ w_{n,1} & w_{n,2} & \cdots & w_{n,j} & \dots & 0 \end{bmatrix}$$

donde las columnas y filas corresponden a las observaciones de corte transversal y los pesos w_{ij} ($i, j = 1, 2, \ldots, n$) aproximan la relación entre dos localizaciones *i* (filas) y *j* (columnas). La diagonal principal está formada por ceros, estableciendo que ninguna observación puede estar relacionada consigo misma (la misma observación no puede ser vecina de sí misma). Se asume que *W* es no negativa, de tal manera que los pesos $w_{ij} \geq 0$ para $i \neq j$.

La estructura de vecindades es frecuentemente construida desde la geografía o geometría, usando los conceptos de contigüidad y distancia. Suponiendo que las observaciones se distribuyen sobre un mapa (Figura 1.2), hay varias alternativas para establecer el conjunto de vecinos de la celda *a*. Una posibilidad es considerar vecinos a aquellas celdas que poseen un borde común, siendo para nuestro caso, cada celda *b* del mapa i.- (Figura 1.2). Otra posibilidad es considerar como vecinos a aquellas celdas que poseen un vértice común, tal como se muestra en el mapa ii.- (Figura 1.2). Las elecciones de estos conjuntos de vecinos son denominadas, respectivamente, “criterio tipo torre” y “criterio tipo alfil”, en analogía al

movimiento de las piezas de ajedrez. De igual forma, pueden seleccionarse vecinos mediante una combinación de ambos criterios, dando lugar al "criterio tipo reina" (Anselin, 1988).

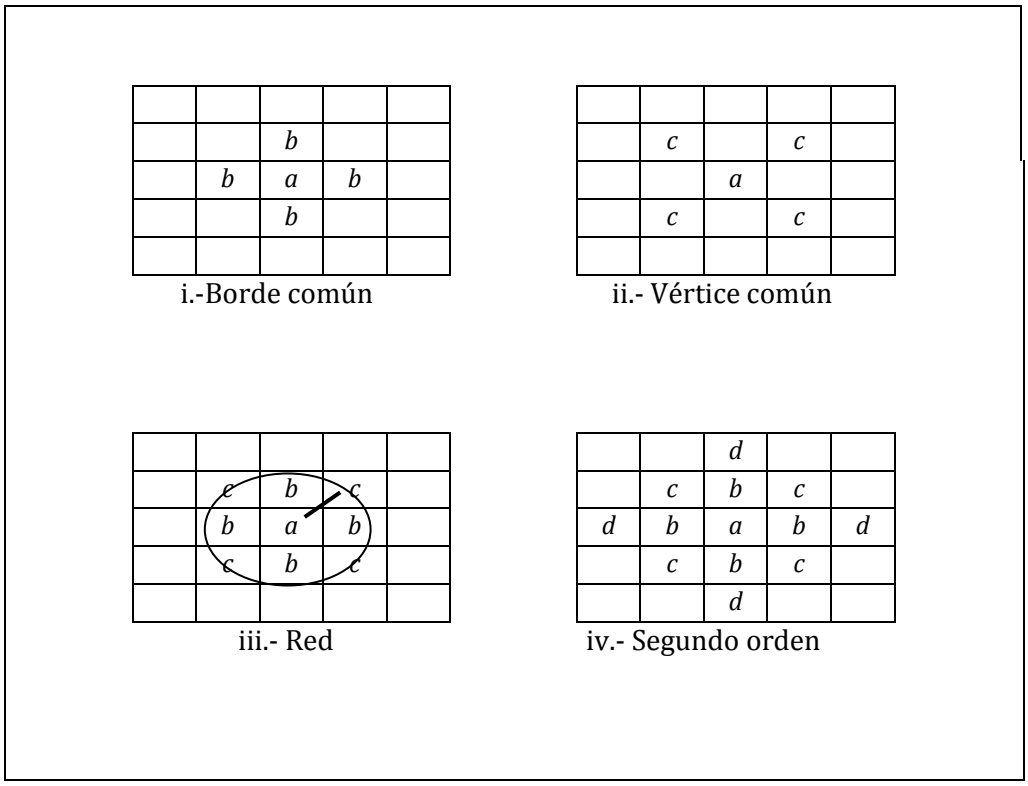
Las áreas o polígonos de un mapa pueden transformarse en puntos, y viceversa. La elección de un punto representativo de un polígono es típicamente resuelto por criterios geométricos mediante el punto central o centroide del polígono. Dado el centroide de cada celda, como muestra el gráfico iii.- (Figura 1.2), es posible definir una red de puntos o nodos. En este caso, se aplica un criterio de distancia para definir a los vecinos de *a,* tal que cada centroide que se encuentre dentro de una distancia máxima al centroide de *a*, será considerado vecino. Cuando las unidades espaciales son puntos regular o irregularmente distribuidos sobre el sistema, como el caso de las ciudades en una jerarquía urbana, el concepto de contigüidad se define sobre la noción de distancia.

Un problema con la elección de vecinos por medio de la distancia (y contigüidad) es la existencia de puntos aislados que pueden no contener vecino alguno para un radio determinado. Esto sucede habitualmente cuando la densidad de los puntos sobre el plano no es regular o cuando se encuentran algunos nodos distribuidos por agrupamientos (*clusters*) y otros aislados. Para salvar este problema se suele determinar un radio *d* de amplitud tal que asegure que cada observación tienen al menos un vecino.

Un criterio alternativo es el de *k*–vecinos más cercanos. En este caso, considerando la distancia geométrica entre las regiones, se seleccionan los *k* vecinos más cercanos de cada punto.

Por otro lado, es posible considerar varios órdenes de contigüidad o vecindad, como un grupo de vecinos de segundo orden para la celda *a*, tal como en la gráfica iv.- (Figura 1.2). Este conjunto de vecinos, identificado por las letras *c* y *d*, incluye a los vecinos de los vecinos de *a* definidos por el criterio tipo torre. Para distinguir el orden de vecindad se añade un supra-índice a la matriz $W^{(k)}$, $\forall\ k \geq 2$, siendo *k* el orden de vecindad.

Figura 1.2. Diferentes criterios de contigüidad sobre un mapa regular.



Fuente: Anselin L. (1988). Spatial Econometrics: Methods and Models.

Una vez establecida la lista de conjuntos de vecinos en el área de estudio, se debe proceder a asignar ponderaciones espaciales a cada relación. Existen diferentes criterios para la construcción de los pesos de la matriz *W*, siguiendo, por ejemplo, alguna hipótesis de interacción. Cada hipótesis resultará en una matriz de ponderaciones diferentes. Cuando se desconoce sobre el proceso espacial asumido, es conveniente mantener una elección binaria, con $w_{i,j} = 1$ cuando *i* y *j* son vecinos, y $w_{i,j} = 0$ cuando no lo son (Bavaud, 1998). No obstante, también es posible construir una matriz de pesos no binaria, donde los elementos $w_{i,j}$ expresan el grado de interacción espacial potencial entre cada par posible de localizaciones. En efecto, como alternativa a los pesos binarios, pueden considerarse funciones que combinan la distancia, el perímetro y otras características geográficas de las unidades espaciales.

López Hernández *et al.* (2000), presentan el criterio de definir dos localidades como vecinas si tienen frontera en común. En este caso la intensidad de esta relación es constante (independiente, por ejemplo, de la longitud de la frontera) y la misma para todas ellas. Además, definen dos criterios alternativos para determinar las vecindades de cada zona y evaluar la presencia o ausencia de correlación espacial:

• Dos localidades son vecinas si se encuentran a menos de determinada distancia. En este caso la intensidad de la relación entre las distintas zonas no es constante, como en el caso

anterior, y la intensidad de la relación será inversamente proporcional a la distancia Euclídea entre dos capitales. Concretamente:

$$w_{ij} = \begin{cases} 0 & si\ d_{ij} > d \\ \dfrac{1}{d_{ij}} & si\ d_{ij} \leq d, \end{cases}$$

donde (*i, j*) denota las localizaciones vecinas, d_{ij} indica la distancia Euclídea entre las localidades *i* y *j* y *d* indica la distancia máxima fijada para que dos localizaciones sean consideradas vecinas.

• Como en el caso anterior, dos localidades son vecinas si se encuentran a menos de determinada distancia. Sin embargo, el modelo que determina la intensidad de la relación entre zonas vecinas es seleccionado con el fin de presentar una relación no simétrica. La relación existente entre dos zonas será inversamente proporcional a la distancia y directamente proporcional al cociente entre los tamaños de las poblaciones. La expresión de la matriz de conexiones es la siguiente:

$$w_{ij} = \begin{cases} 0 & si\ d_{ij} > d \\ \dfrac{1}{d_{ij}} \dfrac{p_j}{p_i} & si\ d_{ij} \leq d, \end{cases}$$

donde p_i *y* p_j hacen referencia a la cantidad de habitantes de las localidades *i* y *j.*

Con esta matriz establecemos el criterio por el cual las localizaciones más pobladas ejercen más influencia que las menos pobladas dentro de una misma vecindad.

Algunos otros criterios alternativos presentados en la literatura son:

• $w_{ij} = \frac{1}{|x_i - x_j|}$ donde *x*: variable socioeconómica, por ejemplo, el PBI per cápita (Case *et al*., 1993).

• $w_{ij} = \left[d_{ij}\right]^{-a}$ donde d_{ij} es la distancia entre dos puntos o regiones (*i, j*) y *a* es un coeficiente positivo (Cliff y Ord, 1981).

• $w_{ij} = d_{ij}^{-2}$ donde d_{ij} es la distancia entre dos puntos o regiones (*i, j*) (Anselin, 1980).

• $w_{ij} = \gamma_{ij}\alpha_i\beta_{i(j)}$ donde γ_{ij} es un factor de contigüidad binario (1,0), α_i es la proporción del área de la unidad *i* respecto al total de área de todas las unidades del sistema y $\beta_{i(j)}$ es la proporción del perímetro de la unidad *i* en contacto con la unidad *j* (Dacey ,1969)

Sin embargo, algunas de estas especificaciones plantean posibles problemas de endogeneidad que deberán ser considerados en el momento de la estimación del modelo. Este tipo de problema puede surgir con propuestas que utilizan variables socio-económicas, como el nivel de empleo o el producto bruto *per cápita*, para la elección de los pesos. Exceptuando estos casos, los procedimientos presentados pueden encasillarse como procedimientos exógenos, es decir, aquellos que determinan la estructura de la matriz en función, únicamente, del arreglo espacial de los datos.

Por otro lado, es lógico suponer que la fuerza de las relaciones entre vecinos se atenúa con la distancia, tal como consideran Cliff y Ord (1981), de manera que los pesos sean proporcionales a la inversa de la distancia entre puntos. No obstante, si lo único que se conoce acerca de las relaciones de vecindad es su existencia o ausencia, este paso puede ser potencialmente engañoso. Asimismo, si se conoce que existen flujos entre áreas vecinas, que describen la estructura de los pesos espaciales, debe considerarse su utilización como pesos generales.

Una vez elegidos los pesos espaciales, lo habitual es trabajar con alguna transformación de la matriz, en el modelo espacial, ya que mejora las propiedades estadísticas de los estimadores y sus estadísticos. En particular, Tiefelsdorf y Griffith (2007), han considerado los siguientes sistemas de codificación para *W*:

• Normalización por fila, donde los nuevos pesos son obtenidos como:

$$w_{ij}^{*} = \frac{w_{ij}}{\sum_j w_{ij}},$$

de tal forma que la suma de cada fila de la matriz sea igual a la unidad: $\sum_j w_{ij}^{*} = 1$.

• Estandarización global, que calcula los nuevos pesos como:

$$c_{ij} = \frac{n.w_{ij}}{\sum_i \sum_j w_{ij}}.$$

• Normalización de los pesos espaciales, donde:

$$u_{ij} = \frac{w_{ij}}{\sum_i \sum_j w_{ij}},$$

de tal forma que la suma sea igual a la unidad: $\sum_i \sum_j u_{ij} = 1$.

• Varianza de estabilización de los pesos espaciales, tal que:

$$s_{ij} = \frac{n.s_{ij}^*}{\sum_i \sum_j s_{ij}^*},$$

donde $s_{ij}^* = \frac{w_{ij}}{\sqrt{\sum_j w_{ij}^2}}$. .

Tal como se mostrará en el apartado de Resultados, la elección de los criterios para definir vecinos así como el esquema de codificación elegido para los pesos influyen en las conclusiones obtenidas. Los enlaces correspondientes a áreas con muchos vecinos se pueden ponderar hacia arriba o hacia abajo, dependiendo de la elección del estilo (Bivand *et al.*, 2008).

1.3.3. Detección de Autocorrelación Espacial

La autocorrelación espacial puede definirse como un fenómeno que ocurre cuando la distribución espacial de la variable de interés exhibe algún patrón sistemático (Cliff y Ord, 1981) o como la existencia de una relación funcional entre lo que ocurre en un punto determinado del espacio y lo que ocurre en otro lugar (Moreno y Vayá, 2000). Es decir, una variable se encontrará espacialmente autocorrelacionada cuando los valores observados en una región dependan de los valores observados en regiones vecinas, de forma que se produzca una cierta continuidad geográfica en la distribución de esta variable, por ejemplo, sobre un mapa (Coro, 2003). Su presencia implica que deben formularse modelos estadísticos más complejos, que incorporen explícitamente el efecto del espacio.

En estudios epidemiológicos, con frecuencia, es de interés indagar si el riesgo relativo $\square_i$ del evento de interés (ocurrencia de enfermedad o muerte) se correlaciona espacialmente en una zona determinada. Generalmente, se verifica si los riesgos de áreas contiguas geográficamente son más similares o más diferentes de lo esperado, bajo la hipótesis que establece que la distribución del riesgo no está relacionada con la localización espacial de las áreas (Assunção y Reis, 1999). La cuestión, entonces, radica en conocer aquellas herramientas que permiten detectar la presencia de autocorrelación espacial. Desde una perspectiva descriptiva, una primera aproximación cualitativa puede realizarse mediante el análisis exploratorio espacial (Haining, 2003). Este análisis permite la visualización mediante diferentes gráficos del comportamiento de la variable bajo estudio, siendo el mapa un elemento central. Existe una amplia variedad de mapas y formas para describir datos continuos sobre polígonos irregulares, una de las más usuales es el denominado "mapa de coropletas", que representa la distribución espacial de una variable o atributo mediante diferentes tonalidades. El número de tonalidades corresponde a los diferentes intervalos y los mismos pueden ser definidos por el usuario. Sin embargo, los resultados

que puedan deducirse de la inspección de un mapa son altamente sensibles, entre otros aspectos, al número de intervalos definidos para representar la variable bajo estudio. En consecuencia, es necesario llevar a cabo un estudio exhaustivo que permita contrastar la existencia de un esquema de dependencia espacial estadísticamente significativo en la distribución espacial de una variable. Con este objetivo existen diversos tests de autocorrelación, los cuales pueden dividirse en dos grupos: contrastes globales y contrastes locales de autocorrelación espacial, que serán presentados a continuación.

1.3.3.1. Autocorrelación Espacial Global

Para contrastar si se cumple la hipótesis de que una variable se encuentra distribuida de forma totalmente aleatoria en el espacio o, por el contrario, existe una asociación significativa de valores similares o disímiles entre regiones vecinas, han sido propuestos un conjunto de estadísticos de dependencia espacial. El más ampliamente utilizado es el de Moran (Moran, 1948). Otros estadísticos comúnmente empleados son el de Geary (Geary, 1954) y el de Getis y Ord (Getis y Ord, 1992).

Los valores de estos estadísticos pueden tener algún interés en sí mismos, pero no son directamente interpretables. El enfoque generalmente adoptado es estandarizar el valor observado, bajo la hipótesis nula de que no existe dependencia espacial para los pesos espaciales elegidos. En general, la prueba es unilateral, evaluando si el estadístico observado es significativamente mayor que su valor esperado.

El estadístico *I* de Moran se calcula como:

$$I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} (y_i - \bar{y}) w_{ij} (y_j - \bar{y})}{S_0 \sum_{i=1}^{N} (y_i - \bar{y})^2},$$

donde w_{ij} denota los elementos de la matriz de pesos espaciales W correspondientes a las localidades (i, j), y_i los valores que toma la variable Y de interés en la localización i, $\bar{y}$ la media de la variable Y y $S_0 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$.

En general, I oscila entre -1 y 1, con grandes valores positivos indicando similitud entre vecinos y valores cercanos a cero indicando ausencia de autocorrelación espacial.

Los momentos del estadístico de Moran, bajo la hipótesis nula de aleatoriedad, son:

$$E\,(I) = \frac{-1}{(N-1)},$$

$$V\,(I) = \frac{\left(3S_0^2 + S_1 R^2 - NS_2\right)}{S_0\,(N-1)(N+1)} \frac{1}{(N-1)^2}.$$

Si I es mayor que el valor esperado, la distribución de los riesgos relativos puede ser caracterizada por una autocorrelación espacial positiva, lo que significa que el riesgo en cada área *i* tiende a ser similar en ubicaciones espacialmente contiguas. Si por el contrario, *I* es menor al valor esperado, estaremos en presencia de una autocorrelación espacial negativa, lo que implica que el valor asumido por el riesgo en cada área *i* tiende a ser diferente de los valores correspondientes a áreas espacialmente contiguas.

Su distribución asintótica es normal:

$$\sqrt{N}\,[I - E(I)] \underset{as}{\sim} N[0; V(I)].$$

Para el cómputo del contraste de autocorrelación se puede utilizar cualquier definición de la matriz de pesos *W,* siendo habitual proceder previamente a la estandarización de la misma. Sin embargo, los resultados obtenidos pueden variar, a veces significativamente, en función de la matriz *W* especificada. En el caso de utilizar una matriz de contigüidad física, es recomendable replicar el cálculo del estadístico *I* de Moran para criterios de contigüidad de órdenes superiores, a fin de determinar si el esquema de autocorrelación espacial detectado entre regiones vecinas es extensible a regiones alejadas en el espacio (Moreno y Vayá, 2000).

En Epidemiología, la distribución de probabilidad del estadístico *I,* bajo hipótesis nula, se determina suponiendo que los riesgos relativos θ_i son variables aleatorias independientes e idénticamente distribuidas (*i.i.d.*), con una distribución normal. Esto implica que el valor esperado de los riesgos relativos es constante en todas las áreas, lo que parece estar en conflicto con la situación común de riesgos relativos diferentes pero espacialmente correlacionados. En la práctica, sin embargo, este supuesto subyacente es generalmente violado. Por otro lado, las zonas con poblaciones pequeñas tienen riesgos relativos más variables y, por lo tanto, son más propensas a asumir un valor extremo. Dado que la densidad de población tiende a mostrar una estructura espacial, también lo hace el riesgo relativo (Assunção y Reis, 1999).

Concretamente, el test de Moran permite probar la hipótesis nula H_0) el riesgo relativo es espacialmente constante o los riesgos son heterogéneos sin correlación espacial contra la alternativa H_1) los riesgos relativos son heterogéneos con correlación espacial. Un resultado del test de Moran significativo implica que las zonas cercanas tienden a tener riesgos

similares, produciendo un mapa con *clusters* de valores similares. Sin embargo, si se acepta H_0 podemos tener riesgos homogéneos o heterogéneos.

Diferentes experimentos basados en simulaciones Monte Carlo (Anselin y Florax, 1995) revelan que la aproximación a la distribución normal funciona razonablemente bien con tamaños muestrales medios (n > 50). El comportamiento del estadístico empeora sensiblemente cuando se añaden problemas como heterocedasticidad, atípicos, etc., lo que significa que los resultados deben interpretarse con cautela.

El estadístico *I* de Moran puede aplicarse tanto al *CME*, tomando en cuenta la distribución espacial de la población, como a los casos observados (o_i). Sin embargo, en este último caso, podría existir autocorrelación espacial sólo por la distribución espacial de la población subyacente ya que, en general, mayor es el número de casos cuanto mayor es la población.

Assunção y Reis (1999) proponen un ajuste al estadístico *I* de Moran, para diferentes tamaños poblacionales, basado en un enfoque bayesiano que permite realizaciones con diferentes riesgos entre las áreas. Suponiendo que los riesgos relativos tienen una esperanza y varianza *a priori,* denominadas β y α respectivamente, Marshall (1991) propone estimaciones de estos parámetros, de manera que la esperanza marginal de los θ_i es la misma pero las varianzas difieren entre las áreas y aumentan a medida la población disminuye. Esto permite estandarizar los *CME*s de los diferentes departamentos, obteniendo valores que denominaremos z_i.

El Índice Empírico de Bayes (*IEB*) se calcula como:

$$IEB = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} z_i z_j}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \sum_{i=1}^{N} (z_i - \bar{z})^2} .$$

Como el *I* de Moran, el *IEB* tenderá a ser positivo si los riesgos están correlacionados espacialmente y permite probar la hipótesis nula H_0) el riesgo relativo es espacialmente constante o los riesgos son heterogéneos sin correlación espacial. Este nuevo índice, obtenido por permutación, tiene mayor potencia que el de Moran, aplicado directamente sobre las tasas observadas.

1.3.3.2. Detección de Autocorrelación Espacial Local

Uno de los rasgos que caracterizan a los estadísticos de asociación presentados anteriormente es que son válidos para contrastar la presencia de un esquema de autocorrelación espacial global, dado que analizan todas las regiones de la muestra de forma conjunta. Por ello, dichos estadísticos no son sensibles a situaciones donde predomina una importante inestabilidad en la distribución espacial de la variable objeto de estudio (procesos no estacionarios espacialmente), existiendo por ejemplo *clusters* o agrupaciones de regiones localizadas en áreas específicas del territorio que concentran valores más elevados o bajos de lo que cabría esperar ante una distribución homogénea, mientras que la aleatoriedad domina en el resto del territorio. Es decir, no contemplan la posibilidad de que el esquema de dependencia detectado a nivel global (por ejemplo, ausencia de autocorrelación espacial) pueda no mantenerse en todas las unidades del espacio analizado. La estacionariedad de un proceso espacial implica que sus propiedades estadísticas no cambian a través del espacio. En sentido estricto, un proceso espacial es estacionario cuando cualquier distribución conjunta de variables aleatorias sobre un subconjunto de puntos en el espacio depende únicamente de la posición relativa de las diferentes localizaciones, la cual se encuentra determinada por su orientación relativa (ángulo) y las distancias respectivas (Anselin ,1988).

Para superar la limitación de las medidas de autocorrelación espacial global, que ofrecen una imagen "promedio" de la distribución espacial de la variable de interés y, por lo tanto, puede ocultar características interesantes del fenómeno en estudio, durante la última década se han presentado varias medidas de autocorrelación espacial local que pueden utilizarse con diferentes propósitos. Cuando se aplican a los conjuntos de datos que carecen de autocorrelación espacial global, pueden ser capaces de revelar una o más áreas que presentan una desviación significativa de aleatoriedad espacial. Cuando se aplican a los conjuntos de datos donde está presente la autocorrelación espacial global, pueden ayudar a identificar los lugares que más contribuyen a la pauta general de la agrupación espacial (Sokal *et al.,* 1998). De manera más general, se emplean las estadísticas locales para detectar la agrupación espacial significativa alrededor de ubicaciones individuales.

Uno de los estadísticos que puede calcularse es el estadístico local de Moran I_i, basado en el test tradicional I de Moran (Anselin y Florax, 1995) Así, se obtendrá un valor de dicho estadístico para cada región de la muestra, pudiendo analizar la situación de cada unidad espacial por separado.

El I_i de Moran se calcula a partir de la siguiente expresión:

$$I_i = \frac{z_i}{\sum_{i=1}^{N}(z_i)^2/N}\sum_{j=1}^{N} w_{ij}z_j \ ,$$

donde w_{ij} denota los elementos de la matriz de pesos espaciales *W* correspondientes a las localidades (*i*, *j*), z_i los valores que toma la variable *Y* de interés estandarizada en la localización *i* y z_j los valores que toma la variable *Y* de interés estandarizada en las localizaciones *j* vecinas de *i.*

Siguiendo una hipótesis de distribución aleatoria, la esperanza del citado estadístico local es:

$$E\ (I_i) = \frac{-\sum_{j=1}^{N} w_i}{(N-1)}.$$

De forma similar a los estadísticos anteriores, se puede asumir la hipótesis de que la I_i estandarizada se distribuye según una normal N(0,1). Tras su estandarización, un valor positivo (negativo) de I_i indicará la existencia de un *cluster* de valores similares (disímiles) de la variable analizada alrededor de la región *i*.

A partir de la I_i es posible conocer la contribución exacta que presenta cada región en el valor del estadístico global de dependencia *I* de Moran, pudiendo de esta forma detectar observaciones *outliers*, es decir, observaciones con una contribución excepcional al mismo.

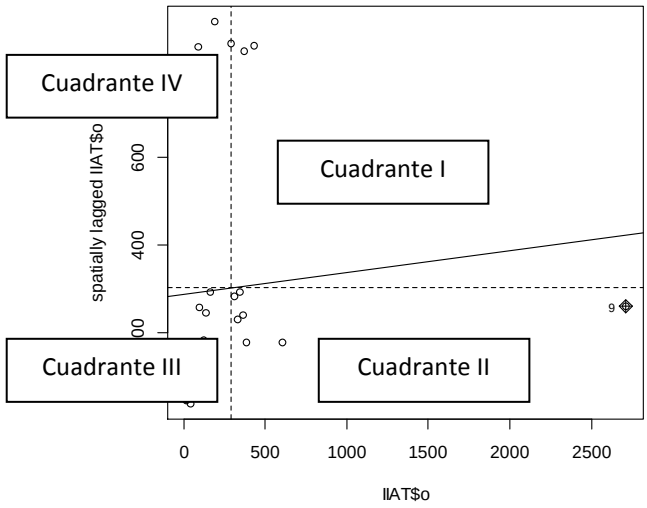
Sokal *et al*. (1998) demuestran que las pruebas de significación para la autocorrelación espacial local son problemáticas y, por lo tanto, la *p-valores* reportados deben considerarse solamente como indicación aproximada de la significación estadística, especialmente en la presencia de autocorrelación espacial global. Sin embargo, los estadísticos (sobre todo en su forma estandarizada) son informativos cuando se emplean de manera exploratoria.

1.3.3.3. Scatterplot

Otro instrumento útil en el análisis del grado de dependencia espacial de una variable y que suministra información similar a la obtenida con el cómputo del estadístico *I* de Moran es la observación del denominado Scatterplot de Moran. Este tipo de gráfico representa en el eje de abscisas las observaciones de la variable normalizada y en el de ordenadas el retardo espacial de dicha variable también normalizado (obtenido tras multiplicar la matriz *W* por la variable *Y*). De este modo, los cuatro cuadrantes reproducen diferentes tipos de dependencia espacial. Si la nube de puntos está dispersa en los cuatro cuadrantes es indicio de ausencia de correlación espacial, tal como puede observarse en la Figura 1.3 presentada a modo de ejemplo. Si, por el contrario, los valores se encuentran concentrados sobre la

diagonal que cruza los cuadrantes I y III , existe una elevada correlación espacial positiva de la variable, de forma que su pendiente es igual al valor obtenido para el contraste de la *I* de Moran. La dependencia será negativa si los valores se concentran en los dos cuadrantes restantes.

Figura 1.3. Ejemplo de gráfico *Scaterplott* de Moran.



Fuente: elaboración propia

El cuadrante I representa agrupamientos espaciales de altos valores de la variable alrededor de localizaciones con altos valores (alto-alto), mientras que el cuadrante III representa agrupamientos espaciales de bajos valores de la variable alrededor de localizaciones con bajos valores (bajo-bajo). Asimismo, el cuadrante II representa agrupamientos espaciales de bajos valores de la variable alrededor de localizaciones con altos valores (bajo-alto), en tanto que el cuadrante IV representa agrupamientos espaciales de altos valores de la variable alrededor de localizaciones con bajos valores (alto-bajo). Estas cuatro categorías dan lugar a una cierta forma de alisado espacial que puede visualizarse fácilmente en un mapa (Anselin y Bao, 1997).

1.4. Modelos

1.4.1. Introducción

Una característica, asociada específicamente a la epidemiología espacial, es que los datos son frecuentemente de tipo discreto, por lo que los modelos desarrollados para ser aplicados en esta área están asociados a distribuciones de probabilidad discretas, para conteos dentro de regiones arbitrarias (Lawson, 2001).

En el estudio de la distribución espacial de enfermedades existen modelos básicos que usualmente se aplican, al menos como punto de partida, en el análisis de los conteos de casos (Lawson, 2001). Estos modelos, basados en la verosimilitud, involucran ciertos supuestos referidos a los datos que se examinan, como que los casos dentro de la población de estudio se comportan de forma independiente con respecto a la propensión a enfermedad o muerte (Lawson *et al.*, 2003). Sin embargo, estos supuestos fundamentales no siempre se cumplen. Por ejemplo, en el caso de enfermedades infecciosas, cuando se consideran dentro del área de estudio, zonas con población cero o cuando varios casos ocurren en la misma localización. Por ello es necesario recurrir a modelos más complejos, que incluyen efectos aleatorios para describir la heterogeneidad no observada de los datos o modelar la falta de independencia de las observaciones (Cressie, 1993). El tratamiento de estos efectos no observados dentro de las unidades espaciales de análisis, comúnmente llamados efectos aleatorios, ha sido tratado extensamente en la literatura, tanto de metodología estadística como de aplicaciones epidemiológicas (Breslow y Clayton, 1993; Clayton, 1991; Lawson, 2001; Richardson, 2003). Los efectos aleatorios pueden tomar diferentes formas y se deben usar métodos adecuados para obtener estimaciones correctas, bajo modelos que incluyan estos efectos (Waller *et al.*, 1997; Lawson *et al.*, 2003).

Generalmente, el riesgo relativo θ_i es el foco de atención cuando se aplican modelos en el mapeo de enfermedades. La modelización espacial de riesgos ha hecho uso repetidamente de distintas herramientas para conferir estructura de dependencia espacial a las observaciones objeto de modelización. La aproximación frecuentista incluye métodos que persiguen la estimación de parámetros dentro de un modelo de estructura jerárquica. Estos métodos asumen que los efectos aleatorios tienen una distribución compuesta o *a priori*. Como se ha expresado, un supuesto común cuando se examinan datos de conteo dentro de regiones (o_i) es que o_i ~Poisson ($e_i\mu_i$) independientes, donde e_i es el conteo esperado y θ_i se considera un parámetro constante de riesgo relativo. Asimismo, se puede asumir que θ_i ~Gamma (ν, α) y la estimación del parámetro del modelo puede realizarse por máxima verosimilitud (Bock y Aitkin, 1981; Aitkin, 1996). Bajo la perspectiva Bayesiana, la

incertidumbre o falta de información sobre el parámetro puede ser incorporada a través de distribuciones previas, considerando este parámetro como una variable aleatoria.

Una alternativa a las especificaciones anteriores es considerar modelo multinivel con respuesta Poisson que puede ajustarse a partir del enfoque frecuentista (Rasbash *et al.,* 2000; Raudenbush *et al.,* 2000; Rabe-Hesketh *et al.,* 2001) o Bayesiano. El primero generalmente involucra algunas aproximaciones; por ejemplo el paquete MLwiN (Rasbash *et al.,* 2000) usa métodos métodos de cuasi-verosimilitud (Goldstein, 1991; Goldstein y Rasbash, 1996) para transformar el problema, de manera que pueda ajustarse usando el algoritmo general iterativo de mínimos cuadrados. Otros enfoques clásicos emplean aproximaciones de Laplace (Raudenbush *et al*., 2000) y cuadratura Gaussiana (e.g. Rabe-Hesketh *et al*., 2001). Asimismo, la confluencia de la estadística Bayesiana y la potencia de computación han permitido realizar análisis espaciales en el campo de las Ciencias de la Salud. La difusión de los métodos *MCMC* (Markov Chain Monte Carlo) y su aplicación mediante el *software* WinBUGS suponen toda una revolución en éste y en muchos otros campos (Gilks et al., 1993; López-Abente e Ibáñez, 2001).

Específicamente, los Modelos generalizados mixtos y para variables latentes (*GLLAMMs*) (Rabe-Hesketh y Skrondal, 2001) ofrecen una solución multinivel para respuestas de tipo continuas, conteos, datos de supervivencia, dicotómicos, ordenados y categóricos. Ejemplos de modelos de esta clase son los modelos generalizados lineales multinivel y los modelos lineales generalizados mixtos. Al igual que en el análisis clásico de regresión, el objetivo de los modelos multinivel (*MM*) es ajustar la relación entre una variable de respuesta y un conjunto de variables explicativas. La diferencia es que los *MM* involucran unidades de observación a diferentes "niveles" (Rabe-Hesketh and Skrondal, 2012). Incluyen modelos jerárquicos y modelos con efectos aleatorios, y son particularmente útiles cuando se dispone de datos que poseen niveles de agrupamiento dentro de su estructura, por ejemplo, clases dentro de escuelas o pedanías anidadas en departamentos. Si bien es posible que dentro de un área o *cluster* particular las observaciones sean independientes, es de esperar que las observaciones sean más parecidas dentro del agrupamiento que en agrupamientos diferentes, debido al entorno social y geográfico compartido. En consecuencia, cuando se desea modelar este tipo de información es importante considerar la estructura subyacente y, en particular, la correlación entre observaciones del mismo *cluster* o área (Goldstein, 1995). Los modelos multinivel son, además, particularmente adecuados para ajustar observaciones tomadas en diversos momentos para cada sujeto (medidas repetidas) y tienen la ventaja de que pueden utilizarse cuando existen datos faltantes, asumiendo que son completamente al azar (Lawson *et al*, 2003).

Los *MM* comenzaron a aplicarse en la década de 1980, considerando variables de respuesta continuas y simétricas, en el campo de la Educación donde los datos tienen, naturalmente, una estructura jerárquica, por ejemplo, estudiantes anidados dentro de clases o escuelas (Aitkin *et al.,* 1981; Aitkin and Longford, 1986). En el campo de las Ciencias de la Salud, si bien las respuestas son generalmente conteos, es posible encontrar algunos ejemplos de aplicación de estos modelos para variables continuas. Este es el caso de Leyland y McLeod (2000), que consideraron el tratamiento del *CME* como una variable de respuesta continua, para el período 1979-1992 en 403 distritos de Inglaterra y Gales.

Cuando se considera el conteo de casos observados de muerte, los supuestos de un proceso Poisson indican que, a cada área, le corresponde una media diferente. Sin embargo, para simplificar, se considera que la esperanza de los conteos en el área *i* es una función de un parámetro dentro de un modelo jerárquico, sin considerar la distribución continua subyacente del riesgo. Este supuesto implica una simplificación considerable y se asume, entonces, que la distribución de los casos observados en la localidad *i* (o_i) es:

$$o_i \sim \text{Poisson}\,(e_i\theta_i)$$

donde e_i son los casos esperados y θ_i es un parámetro de riesgo relativo para la misma localidad. Así, como primera aproximación, se considera que los o_i son independientes y procedentes de una distribución de Poisson cuya media es e_i ?i.

El estimador de máxima verosimilitud de ?i, que se llama cociente de mortalidad estandarizado (*CME*), es:

$$\hat{\theta}_i = \frac{o_i}{e_i}.$$

El *CME* puede interpretarse como el *odds* de estar en el grupo enfermo *versus* estar en el grupo sano, de manera que si el estimador es mayor que 1, entonces hay un exceso en el riesgo de esa región. Por lo tanto, será de interés identificar las regiones en las que el riesgo relativo es significativamente mayor que 1 (Banerjee *et al*., 2003; Haining, 2003, y Lawson *et al*., 2003).

Como se indicó antes, desafortunadamente, la varianza de este estimador es proporcional a $1/e_i$. En consecuencia, las estimaciones para enfermedades raras o áreas con poblaciones pequeñas, donde el número de casos esperados es realmente bajo, pueden conducir a estimadores muy imprecisos. En el extremo, un valor para e_i cercano a cero generaría un *CME* muy grande. Asimismo, *CME*s de cero no distinguen variaciones en los conteos esperados. Para abordar este problema, Rabe-Hesketh y Skrondal (2005) proponen emplear

un Modelo Poisson con intercepto aleatorio (*smoothing*) en combinación con predicción Bayesiana empírica. Además, los *CMEs* no consideran que áreas geográficas cercanas tienden a tener cocientes similares. Este inconveniente puede resolverse permitiendo que los interceptos aleatorios estén espacialmente correlacionados (Skrondal y Rabe-Hesketh, 2012).

El modelo de Poisson es estricto, en el sentido que impone que la media y la varianza sean iguales. En consecuencia, cuando los datos exhiben algún tipo de sobredispersión, es poco probable que la distribución de Poisson sea la correcta. En aplicaciones espaciales es importante, además, distinguir dos formas básica de extra-variación. Primero, como en el caso no espacial, un extra-variación independiente y no correlacionada espacialmente (Besag *et al.,* 1991). Segundo, una sobredispersión entre unidades espaciales autocorrelacionadas, que reconoce diferentes causas. Por ejemplo, la enfermedad bajo análisis puede presentar un agrupamiento espacial, lo que sucede con numerosas enfermedades infecciosas y con otras que aparentemente no lo son (Cuzick y Hills, 1991; Glick, 1979). Además, la autocorrelación espacial puede ser inducida por la existencia de efectos ambientales o de exposición no observados.

Díaz *et al.* (2009, 2010[a]) aplicaron modelos multinivel para estudiar la distribución espacial de cáncer en la Provincia de Córdoba. En primer lugar ajustaron un modelo Poisson mixto (Rabe-Hesketh y Skrondal, 2005) a la variable respuesta Y_{ij} definida como la parte entera del índice de incidencia de cáncer estandarizado, ajustado por edad de acuerdo a los estándares internacionales, considerando el sexo y el efecto departamento, mediante un modelo de nivel 1 con intercepto aleatorio. La Provincia de Córdoba se encuentra dividida en 26 departamentos, cada uno subdividido en pedanías, es decir, unidades administrativas más pequeñas (salvo el departamento Capital). Teniendo en cuenta, entonces, que las pedanías (i) se anidan en departamentos (j), ajustaron un modelo un modelo Poisson de dos niveles, con sobredispersión, incluyendo un intercepto aleatorio que varía entre departamentos (*superclusters*):

$$\ln(\mu_{ij}) = \ln(e_{ij}) + \beta_1 + \xi_{ij}^{(1)} + \xi_j^{(2)},$$

donde $\xi_{ij} \sim N(0, \psi^{(1)})$ y $\xi_j \sim N(0, \psi^{(2)})$.

Teniendo en cuenta el marco teórico general desarrollado, se introducirán primeramente los modelos *GLLAMM*s, con las particularidades propias del tratamiento de datos de conteo. Posteriormente se realizará el abordaje de las diferentes estrategias de modelación, considerando las tres grandes perspectivas indicadas por Lawson *et al.* (2003) para el

estudio de la distribución geográfica de enfermedades: i)estimación del riesgo relativo en áreas pequeñas, ii) estudios de asociación geográfica y iii)aglomeraciones de casos.

1.4.2. *GLLAMMs* para datos de conteo

Una respuesta continúa y_i para el área i podría ser modelada usando un modelo de regresión estándar como:

$$y_i = \mathbf{X\beta} + \xi_i \ ,$$

Sin embargo, no es razonable asumir que las desviaciones ξ_i respecto de y_i, para una población con media β, no estén correlacionanadas en áreas cercanas. Esta dependencia puede modelarse dividiendo el residual ξ_i en dos componentes de error, específicos del área i y no relacionados: un componente ζ_i y un componente ϵ_i.

$$\xi_i \equiv \zeta_i + \epsilon_i \ .$$

Así ζ_i, llamado frecuentemente *efecto aleatorio* o *intercepto aleatorio,* mide la desviación del área i con respecto a la media general β y no está correlacionado entre áreas. Tiene esperanza cero y varianza ψ, interpretable como la variabilidad entre áreas y representa las diferencias entre zonas debido a características no incluidas como variables en el modelo.

El componente ϵ_i, denominado *residuo,* es la desviación aleatoria de la respuesta y_i con respecto a la media del área i ($\beta + \zeta_i$). Este componente tiene esperanza cero y varianza constante θ, interpretable como la variabilidad dentro de un área. Además, se asume que el residual total tiene esperanza cero:

$$E(\zeta_i + \epsilon_i) = 0,$$

lo que implica que la esperanza de la media de las respuestas, llamada *estructura de medias,* es:

$$E(y_i) = \beta.$$

Si la estructura de medias ha sido correctamente especificada, el estimador $\hat{\beta}$ del parámetro β será consistente e insesgado (esto último si la distribución de $\zeta_i + \epsilon_i$ es simétrica).

Para determinar la estructura de covarianza, se asume que el intercepto aleatorio ζ_i y el residuo ϵ_i no están correlacionados. En consecuencia, la varianza de las respuestas resulta:

$$Var(y_i) = E\{(\zeta_i + \epsilon_i)^2\} = E\left(\zeta_i^2\right) + 2E(\zeta_i\epsilon_i) + E\left(\epsilon_i^2\right).$$

Como $Cov(\zeta_i, \epsilon_i) = 0$, luego:

$$Var(y_i) = Var(\zeta_i + \epsilon_i) = \psi + \theta.$$

Si la estructura de medias y covarianza son correctas, los estimadores de todos los parámetros serán consistentes y asintóticamente eficientes.

Cuando se analizan conteos y_i para diferentes áreas *i*, caracterizados por covariables, la esperanza μ_i es modelada generalmente usando un modelo log-lineal. Así, en forma compacta, este modelo puede expresarse como:

$$\mu_i \equiv \mathrm{E}(\mathrm{y_i} \,|\, \mathbf{X}) = \exp(\mathbf{X\beta}),$$

donde la función exponencial impide que la esperanza de los conteos sea negativa. El modelo puede escribirse alternativamente como:

$$\ln(\mu_i) = \mathbf{X\beta},$$

donde *ln* es la función de enlace y la varianza condicional de los conteos, dadas las covariables es $Var(y_i \,|\, \mathbf{X}) = \mu_i$.

Rabe-Hesketh y Skrondal (2012), sugieren modelar la dependencia de las observaciones pertenecientes a áreas cercanas incluyendo un intercepto aleatorio en el modelo Poisson de regresión. Este modelo mixto simple para los conteos y_i es:

$$\mu_i \equiv \mathrm{E}(\mathrm{y_i} \,|\, \mathbf{X}, \zeta_i) = \exp(\mathbf{X\beta} + \zeta_i), \qquad (1)$$

donde **X** es una matriz *m* x *p* de *p* -1 covariables, **β** es un vector de parámetros *p* x 1 y ζ_i es un intercepto aleatorio que representa la heterogeneidad no observada entre áreas.

En este modelo, el intercepto aleatorio ζ_i específico del área, se incluye para modelar la dependencia dentro de cada zona y se asume que, dadas las covariables, los interceptos ζ_i :

- son variables aleatorias independientes e idénticamente distribuidas (*i.i.d.*),
- siguen una distribución Normal $\zeta_i \,|\, \mathbf{X} \sim N(0, \psi)$,
- tienen esperanza cero:

$$E(\zeta_i \,|\, \mathbf{X}_i) = 0, \qquad (2)$$

lo que implica, asimismo, que $Corr(\varsigma_i, \mathbf{X}_i)=0$, es decir, que ζ_i y las covariables x_i son independientes y los ζ_i son independientes entre áreas,

- poseen varianza homocedástica:

$$Var(\varsigma_i | \mathbf{X}_i)=\psi,$$

lo que implica que $Var(\varsigma_i)=\psi$,

- no están correlacionados para diferentes áreas *i* y *i'*:

$$Cov(\varsigma_i, \varsigma_{i'} | \mathbf{X}_i, \mathbf{X}_{i'})=0 \quad \text{si i} \neq \text{i}'.$$

De la expresión (1) se deduce que la regresión condicional es lineal (modelo condicional o específico del área):

$$E(y_i | \mathbf{X}, \varsigma_i)=\exp\{\mathbf{X\beta}+\varsigma_i\},$$
$$E(y_i | \mathbf{X\beta}_i, \varsigma_i)=\exp\ (\mathbf{X\beta}).\exp(\varsigma_i).$$

El supuesto (2) implica que el modelo marginal o de media poblacional también es lineal:

$$E(y_i | \mathbf{X}_i)=\exp\{(\psi/2)+\mathbf{X\beta}\},$$
$$Var(y_i | \mathbf{X}_i)=E(y_i | \mathbf{X}_i)+\left[E(y_i | \mathbf{X}_i)\right]^2\{\exp(\psi)-1\}.$$

Si $\psi > 0$, la varianza es mayor que la media marginal y existe sobredispersión.

El método clásico para estimar los parámetros del modelo es el de Máxima Verosimilitud (*MV*). La función de verosimilitud es la densidad de probabilidad conjunta de las respuestas observadas y_i (i = 1,..., n) para las *i* áreas, como función de los parámetros **β y** ψ.

$$\Pr(y_1,...,y_n | \mathbf{X}, \zeta_i)=\prod_{i=1}^{n_i}\Pr(y_i | \mathbf{X}, \zeta_i).$$

Para obtener la distribución marginal conjunta de las respuestas, no condicionada a ζ_i, integramos respecto al intercepto aleatorio:

$$\Pr(y_1,...,y_n | \mathbf{X})=\int \Pr(y_i,...,y_n | \mathbf{X}, \varsigma_i)\phi(\varsigma_i;0,\psi)d\varsigma_i \ , \qquad (3)$$

donde $\phi\ (\varsigma_i; 0, \psi)$ es la densidad normal de ζ_i con media 0 y varianza ψ.

El producto de la contribución a la verosimilitud para todas las áreas, es la comúnmente denominada *verosimilitud marginal* y la idea es encontrar los estimadores de los parámetros $\widehat{\boldsymbol{\beta}}$ y $\widehat{\boldsymbol{\psi}}$ que maximicen la función de verosimilitud (Rabe-Hesketh y Skrondal,

2012). Como las áreas son independientes, la *verosimilitud marginal* se obtiene como el producto de la distribución marginal conjunta de las respuestas:

$$L(\boldsymbol{\beta},\ \psi)=\prod_{i=1}^{n}\Pr(y_1,\ldots,y_N\,|\,\mathbf{X}).$$

Se aplica un proceso interativo, que comienza con valores iniciales para los parámetros y continua hasta alcanzar el máximo, utilizando Newton-Raphson o el algoritmo *EM (Expectation-Maximization).*

La integral sobre ζ_i en (3) puede ser aproximada por una suma de R términos, reemplazando ζ_i por localizaciones e_r y la densidad Normal por un peso w_r para $r = 1, \ldots, N$, denominada *Cuadratura ordinaria (Ordinary quadrature aproximation)*:

$$\Pr(y_1,\ldots,y_n\,|\,\mathbf{X},\varsigma_i)\approx\sum_{r=1}^{R}\Pr(y_1,\ldots,y_n\,|\,\mathbf{X},\varsigma_i=\mathrm{e_r})w_r.$$

Sin embargo, si la función a integrar tiene un fuerte pico, como sucede a menudo cuando se analizan conteos, la aproximación anterior tiene mal desempeño (Rabe-Hesketh, Skrondal y Pickles, 2002, 2005). Por ello, se sugiere emplear una aproximación mejorada conocida como *Cuadratura adaptativa (Adaptative quadrature aproximation),* en la cual las localizaciones se reescalan y transladan:

$$e_{ri}=a_i+b_ie_r,$$

donde *a* y *b* son constantes específicas del área.

1.4.3. Estimación del riesgo relativo en áreas pequeñas o *disease mapping*

El objetivo de mapeo de enfermedades es proporcionar una representación de la distribución espacial del riesgo de una enfermedad en el área de estudio, que suponemos se divide en varias regiones más pequeñas. El riesgo puede reflejar muertes reales debido a la enfermedad (mortalidad), como en el caso de este trabajo o, el número de personas que sufren de la enfermedad (morbilidad) en un cierto período de tiempo para la población en riesgo.

Por lo tanto, los datos básicos deben incluir a la población en situación de riesgo y el número de casos en cada zona y se suelen caracterizar en función de una serie de estratos, definidos usando variables como sexo, grupo de edad y otras importantes. Al considerar los datos clasificados en diferentes grupos, la importancia o efecto de cada variable pueden ser explorada y algunos potenciales factores de confusión pueden ser removidos (Elliott *et. al*,

2000) antes de hacer cualquier otro análisis de los datos. Además, implícitamente, se supone que no hay interacción entre el riesgo y los estratos de la población, es decir, el riesgo relativo depende sólo de la región en estudio.

A continuación se presentan dos modelos para la estimación del riesgo relativo en áreas pequeñas, que luego serán aplicados empleando los datos descriptos en la sección de Materiales y Métodos, además de los mapas de probabilidad que complementan el análisis. Estos modelos son el Modelo Poisson con intercepto aleatorio (*PIA*) y el Modelo Poisson-Gamma (*P-G*).

Clayton y Kaldor (1987) asumen que el riesgo relativo θ_i, específico para cada área, es una variable aleatoria. Condicional a θ_i, los conteos observados (o_i) son variables independientes e idénticamente distribuidas (*i.i.d.)* Poisson con media $\mu_i = e_i\theta_i$, esto es:

$$o_i|\theta_i, e_i \sim \text{Poisson}\ (e_i\theta_i).$$

Para datos de este tipo Rabe-Hesketh y Skrondal (2005) proponen el siguiente modelo Poisson:

$$\ln(\mu_i) = \ln(e_i) + \beta_0 + \zeta_i,$$

donde los interceptos aleatorios representan la heterogeneidad no observada entre áreas $\zeta_i \sim \text{i.i.d}$ y $\zeta_i \sim N(0, \psi)$ y $\ln(e_i)$ es una variable *offset* (covariable con coeficiente 1). El propósito de ésta variable *offset* es asegurar un modelo que incluya el *CME*:

$$\ln(\mu_i) - \ln(e_i) = \ln(\mu_i/e_i) = \beta_0 + \zeta_i,$$

ya que $CME_i = \mu_i/e_i$.

Habiendo obtenido estimaciones por *MV* de los parámetros del modelo es deseable asignar valores a los efectos aleatorios ζ_i para áreas particulares. El método de predicción sugerido por Rabe-Hesketh y Skrondal (2005) para la predicción de los *CMEs* es el Empírico Bayesiano (*Empirical Bayes Prediction*) que provee valores más estables para áreas con poblaciones pequeñas. Así, la estimación Bayesiana de ζ_i usa no sólo las respuestas μ_i, sino también la distribución a priori de ζ_i (normal con media cero y varianza estimada) para predecir valores de los efectos aleatorios individuales:

$$Posterior(\zeta_i|\mu_i) \propto \text{Prior}\ (\zeta_i).Likelihood(\mu_i|\mathbf{X}, \zeta_i)$$

Específicamente, la estimación del valor esperado a posteriori del *CME* para un área *i* será:

$$\hat{CME}_i = \text{E}\left[\exp(\hat{\beta}_0 + \zeta_i)\middle| o_i, e_i\right] = \int \exp(\hat{\beta}_0 + \zeta_i) Posterior(\zeta_i|o_i, e_i) d\zeta_i.$$

Como se mencionara con anterioridad, el uso de una distribución de Poisson implica supuestos que no siempre se pueden sostener, fundamentalmente la igualdad entre media y varianza. Es común que los datos presenten sobredispersión, esto es, cuando la varianza es superior a la media. Una manera sencilla de permitir una varianza mayor es utilizar una distribución Binomial Negativa en lugar de la Poisson.

Rabe-Hesketh y Skrondal (2012), proponen un modelo Binomial Negativo, en el marco de los *GLLAMMs*, especificando una distribución Gamma para los interceptos aleatorios de cada departamento. En esta formulación, que se conoce como modelo de Poisson-Gamma, los recuentos o_i, condicionados a θ_i, son realizaciones independientes a partir de una distribución de Poisson cuya media es $e_i \theta_i$:

$$o_i|\theta_i, e_i \sim \text{Poisson}\,(e_i\theta_i)\,,$$

$$\theta_i \sim \text{Gamma}\,(\nu, \alpha)\,.$$

El riesgo relativo θ_i es considerado como una variable aleatoria que se extrae de una distribución Gamma con media ν/α y varianza ν/α^2. Como consecuencia, o_i se distribuye siguiendo una Binomial negativa con parámetro de tamaño ν y probabilidad $\alpha / (\alpha + e_i)$, que se estiman generalmente a través de la estimación empírica Bayesiana (Clayton y Kaldor, 1987).

Además, la distribución *a posteriori* de los o_i, es decir, su distribución dados los datos observados $\{o_i\}_{i=1}^{N}$, también se puede derivar y es una Gamma con parámetros $\nu + o_i$ y $\alpha + e_i$. En otras palabras, la información proporcionada por la observación de los datos actualiza el conocimiento previo o los supuestos sobre θ_i. La esperanza *a posteriori* de θ_i, que es el estimador máximo verosímil de θ_i, es entonces:

$$\widehat{\theta_i} = E\,[\theta_i\,|o_i, e_i] = \frac{\upsilon + o_i}{\alpha + e_i},$$

que también puede ser expresada como un compromiso entre la media previa de la los riesgos relativos y los CME_i:

$$E\,[\theta_i\,|o_i, e_i] = \frac{e_i}{\alpha + e_i}.CME_i + \left(1 - \frac{e_i}{\alpha + e_i}\right).\frac{\upsilon}{\alpha}.$$

Dos cuestiones deben tenerse en cuenta en relación a este estimador. En primer lugar, cuando e_i es pequeño, como sucede a menudo en las zonas poco pobladas, una pequeña variación en o_i puede producir cambios significativos en el valor de CME. Por esta razón, de acuerdo con $E\,[\theta_i\,|o_i, e_i]$, el CME_i tendrá un peso bajo, en comparación con la media *a priori*.

En segundo lugar, la información para construir las estimaciones posteriores está referida a cada área, mientras que ν y α son los mismos para todas las regiones.

En el modelo con intercepto aleatorio distribuido normalmente, la varianza marginal, dadas las covariables $\mathbf{X_i}$ resulta $Var(y_i|\mathbf{X}_i)=E(y_i|\mathbf{X}_i)+\left[E(y_i|\mathbf{X}_i)\right]^2\{\exp(\psi)-1\}$ con $\psi>0$, cuando la varianza es mayor que la media marginal. El factor $(\exp(\psi)-1)$ multiplica el cuadrado de la esperanza marginal, para obtener el componente aditivo de sobredispersión. Sin embargo, este modelo no tiene un expresión cerrada para la verosimilitud y debe ajustarse usando integración numérica. Una aproximación más eficiente computacionalmente puede lograrse, como se ha indicado, especificando una distribución Gamma (con parámetro de escala k y parámetro de forma $1/k$) para el exponencial del intercepto aleatorio ζ_i , con media 1 y varianza k. Así, la varianza marginal tiene una forma cuadrática similar a la del modelo anterior $Var\,(y_i|\mathbf{X}_i) = E(y_i|\mathbf{X}_i) + [E(y_i|\mathbf{X}_i)]^2 k$, donde k reemplaza al factor $\exp(\psi)-1$.

Por último, en el contexto de este tipo de estudios, los mapas de probabilidad (Choynowski, 1959) son una manera conveniente de representar el significado de los valores observados. Estos mapas muestran el *p-valor* del número de casos observados bajo el modelo considerado, de manera que las zonas con recuentos observados más bajos que lo esperado, en base al tamaño de la población, tienen *p-valores* inferiores, y aquellas con recuentos observados mayores que lo esperado tienen *p-valores* más grandes.

1.4.4. Estudios de asociación geográfica o *ecological analysis*

Se refieren al análisis de la relación entre la distribución espacial de los casos de mortalidad o incidencia y de los factores explicativos. Usualmente se realizan en un nivel espacial agregado, y se refieren a la incidencia regional comparada con mediciones de los factores explicativos a nivel de región u otro nivel de agregación (Greenberg *et al*., 1996).

En lo que sigue, se desarrollan dos modelos que, en el marco de este tipo de estudios, permiten incorporar el efecto de covariables, además de modelar la heterogeneidad no observada entre áreas. En primer lugar, se presenta una extensión del Modelo Poisson con efectos aleatorios, analizado en la sección anterior, incorporando variables explicativas. Luego, se introduce el Modelo Autorregresivo Simultáneo (*SAR*) que modela las interacciones espaciales de los datos mediante la incorporación, en la estructura de covarianza, de la dependencia espacial (Wall, 2002). Estos modelos serán ajustados

posteriormente con los datos de mortalidad por cáncer de mama y próstata de la Provincia de Córdoba, tal como se describe en el capítulo siguiente.

Breslow y Clayton (1993) proponen un modelo Poisson simple, con un intercepto aleatorio por departamento, y covariables, que en forma compacta puede expresarse como:

$$\theta_i = \exp\{\mathbf{X\beta} + \varsigma_i\}.$$

Este modelo es reescrito por Rabe-Hesketh y Skrondal (2005) como:

$$\ln(\mu_i) = \ln(e_i) + \mathbf{X\beta} + \zeta_i .$$

Tal como se explicó anteriormente, éste ajuste en combinación con la predicción Bayesiana empírica permite realizar estimaciones más estables del *CME*. Sin embargo, el modelo *PIA* no incorpora en su estructura, de manera explícita, la autocorrelación espacial (Lawson *et al.,* 2003). Dentro de los modelos que sí lo hacen se encuentra el Modelo Autorregresivo Simultáneo (*SAR*), que utiliza una regresión sobre los valores de las otras áreas para modelar la dependencia espacial. En efecto, los términos de error ϵ_i son modelados de modo que dependen unos de otros de la siguiente forma:

$$r_i = \sum_{i=1}^{M} b_{ij} \, . r_j + \epsilon_i \, .$$

Así, los ϵ_i se utilizan para representar los errores residuales, que se supone son independientes y tienen distribución normal con media cero y matriz de covarianza $\sum_{\epsilon}$. con elementos $\sigma^2_{\epsilon_i}$, *i*= 1,….,*M* (generalmente, sin embargo, se considera la misma varianza). Los valores b_{ij} se utilizan para representar la dependencia espacial entre áreas, con lo que b_{ii} debe ser cero, de manera que a cada área no se le aplique una regresión sobre sí misma.

Si los términos de error se expresan según $r = \mathbf{B}\left(\mathbf{Y} - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right) + \boldsymbol{\epsilon}$, el modelo también puede escribirse como:

$$\mathbf{Y} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{B}\left(\mathbf{Y} - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right) + \boldsymbol{\epsilon}\,,$$

donde $\boldsymbol{\epsilon} = (\mathbf{I} - \mathbf{B})\left(\mathbf{Y} - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)$, **B** es una matriz que contiene los parámetros de dependencia b_{ij} e I es la matriz de identidad de la dimensión requerida. Es importante señalar, además, que la matriz (**I – B)** debe ser no singular.

Bajo este modelo, **Y** se distribuye según una normal multivariada con media:

$$E(\mathbf{Y}) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\,,$$

y matriz de covarianza:

$$Var(\mathbf{Y}) = (\mathbf{I} - \mathbf{B})^{-1} \sum_{\epsilon} (\mathbf{I} - \mathbf{B}^{\mathrm{T}})^{-1}.$$

Frecuentemente $\sum_{\epsilon}$. depende de un solo parámetro σ^2, por lo que $\sum_{\epsilon}$.= σ^2. **I**, lo que implica:

$$Var(\mathbf{Y}) = \sigma^2 (\mathbf{I} - \mathbf{B})^{-1} \left(\mathbf{I} - \mathbf{B}^{\mathrm{T}}\right)^{-1}.$$

Una reparametrización útil de este modelo se puede obtener haciendo $\mathbf{B} = \lambda \mathbf{W}$, donde λ es un parámetro de autocorrelación espacial y **W** es la matriz de pesos espaciales. Con esta especificación, la varianza de **Y** se convierte en:

$$Var(\mathbf{Y}) = \sigma^2 (\mathbf{I} - \lambda \mathbf{W})^{-1} \left(\mathbf{I} - \lambda \mathbf{W}^{\mathrm{T}}\right)^{-1}.$$

Este modelo se puede estimar de manera eficiente por máxima verosimilitud, tal como se mostrará en el capítulo de Resultados.

1.4.5. Aglomeraciones de casos o *disease clustering*

Un *cluster* de enfermedad hace referencia a la aparición de más casos que los esperados para una determinada enfermedad, dentro de área geográfica o un período de tiempo.

Como quedará expuesto en la sección Resultados de éste trabajo, los mapas de enfermedad proporcionan una primera visión de la distribución espacial de la enfermedad, pero puede ser necesario localizar zonas donde el riesgo de mortalidad o enfermedad tiende a ser inusualmente más alto de lo esperado. Besag y Newell (1991) distinguen entre métodos de agrupamiento y métodos de evaluación de los riesgos en torno a fuentes de contaminación. Los primeros, que se ocupan de evaluar la presencia de grupos (Wakefield *et al*., 2000; Haining, 2003; Waller y Gotway, 2004), son los que se desarrollarán en éste trabajo. Los métodos de evaluación, en cambio, analizan el riesgo en relación a una fuente específica (Lawson *et. al*, 2003) y, si bien exceden los alcances de ésta tesis, presentan una línea futura de estudio que puede resultar de interés.

Antes de realizar cualquier análisis sobre la presencia de grupos debe evaluarse, entonces, si existen diferencias significativas entre los riesgos relativos, ya que ésta heterogeneidad puede estar relacionada con muchos factores diferentes. Por ejemplo, puede existir una fuente de contaminación en la zona, provocando un aumento del riesgo en las áreas cercanas. En otros casos, la heterogeneidad responde a un factor de riesgo variable espacialmente, y los riesgos elevados se relacionan con una mayor exposición a este factor de riesgo.

Luego de haber calculado, para cada área, los casos esperados y observados, puede aplicarse una prueba Chi-cuadrado para determinar si existen diferencias significativas (globales) entre estas dos cantidades. El estadístico correspondiente está dado por:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_{ii})^2}{e_{ii}},$$

donde es el *CME* global = $\sum_i o_i / \sum_i e_i$ y asintóticamente, sigue una distribución chi-cuadrado con *n* grados de libertad (Wakefield *et al*., 2000).

Potthoff y Whittinghill (1966) proponen otra prueba de homogeneidad de medias para diferentes variables con distribución de Poisson, que puede ser utilizada para probar la homogeneidad de los riesgos relativos. La hipótesis alternativa indica que los riesgos relativos se han extraído de una distribución Gamma con media ν/α y varianza ν/α^2:

$$H_0: \theta_1 = \cdots = \theta_n = \nu$$

$$H_1: \theta_i \sim \text{Gamma}\ (\nu, \alpha)$$

El estadístico es:

$$PW = e + \sum_{i=1}^{N} \frac{o_i(o_i - 1)}{e_i}.$$

La hipótesis alternativa de esta prueba es que los o_i se distribuyen siguiendo una distribución Binomial negativa, como se explicó antes y, por lo tanto, esta prueba también se puede considerar como un test de exceso de dispersión.

Tango (1995) presenta una prueba similar a la Chi-cuadrado para la detección de *clusters*, comparando el número observado y esperado de casos en cada región. Señala que pueden considerarse diferentes tipos de interacciones entre las regiones vecinas y propone una medida basada en la distancia entre dos regiones.

Concretamente, es estadístico es:

$$T = (r - p)^{T} W\ (r - p) \begin{cases} r^T = \dfrac{o_1}{o+} + \cdots + \dfrac{o_N}{o+} \\ p^T = \dfrac{e_1}{e+} + \cdots + \dfrac{e_N}{e+} \\ W = w_{ij}\ , \end{cases}$$

donde $w_{ij} = \exp\{-d_{ij}/\phi\}$, d_{ij} es la distancia entre las regiones *i* y *j* a partir de sus centroides, ϕ es una constante (positiva) que refleja la fuerza de la dependencia entre áreas.

1.4.6. Bondad de Ajuste

Una vez que se ha ajustado un modelo, es necesario evaluar en qué medida es correcto dicho ajuste a los datos. Se considera que un modelo no presenta un buen ajuste si la variabilidad residual es grande, sistemática o no se corresponde con la variabilidad descrita por el modelo (Ryan, 1997). Además, al ajustar distintos modelos a un mismo conjunto de datos, es necesario utilizar criterios para la comparación de los ajustes y, por tanto, para la selección de un modelo.

En los *GLLAMMs* planteados, el interés se centra en probar las hipótesis:

$$H_0: \psi = 0 \ \ vs. \ H_a: \psi > 0$$

La hipótesis nula equivale a $\zeta_i = 0$, es decir, que no hay efectos aleatorios en el modelo. Si no se rechaza H_0, puede emplearse un modelo de regresión clásico en lugar del modelo *GLLAMM*.

Uno de los estadísticos más utilizados para probar estas hipótesis es el Estadístico *-2 Log-Likelihood:*

$$L=2(l_1 - l_0) \ .$$

Donde l_1 corresponde al modelo con intercepto aleatorio y l_0 al modelo sin ζ_i. Este estadístico se distribuye, bajo hipótesis nula, como una combinación $0{,}50\chi^2(0) + 0{,}50\chi^2(1)$ y el *p-valor* se obtiene dividiendo por 2 el correspondiente a una χ^2 con 1 grado de libertad. Este test se denomina *Test de cociente de verosimilitud* y puede emplearse para modelos anidados, por ejemplo, con igual estructura de medias pero diferente estructura de covarianza, como sucede con los modelos Poisson y Poisson con intercepto aleatorio ajustados en este trabajo.

Para comparar modelos y escoger la mejor opción, dos indicadores comúnmente usados son el criterio de información de Akaike (*AIC*) y el criterio de información Bayesiano o criterio de Schwarz (*BIC*). El *AIC* es una suma ponderada del logaritmo de la verosimilitud del modelo y el número de coeficientes ajustados. El *AIC* no sólo recompensa la bondad de ajuste, sino también incluye una penalidad, que es una función creciente del número de parámetros estimados.

$$AIC = -2 \log likelihood + 2k,$$

donde *k* es el número de parámetros estimados.

El *BIC*, estrechamente vinculado con el *AIC*, también introduce un término de penalización para el número de parámetros en el modelo, aunque mayor que en el criterio de información de Akaike.

$$BIC = -2\log likelihood + \ln(n)k$$

donde *n* es el número de observaciones.

Bajo estas expresiones de *AIC* y *BIC*, el mejor modelo resulta ser aquel con menor valor para el indicador.

Los criterios presentados serán utilizados para evaluar la bondad de ajuste de los modelos propuestos en ésta investigación y realizar análisis comparativos del desempeño de los diferentes ajustes.

1.5. Epidemiología de cáncer

1.5.1. Introducción

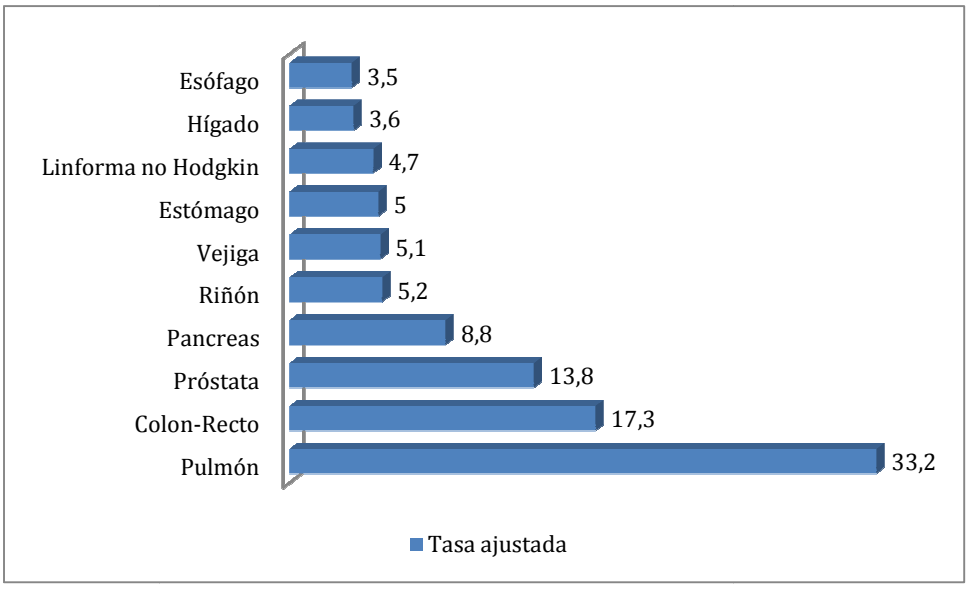
Los patrones de enfermedad en las poblaciones dependen tanto de las características individuales de los sujetos que las componen, como del modo en que estos individuos se relacionan, de sus sociedades y del ambiente en el que habitan (Bhopal, 2002). Se entiende, en consecuencia, que la salud tiene múltiples determinantes, y que su estudio requiere considerar una multiplicidad de factores intervinientes, a varios niveles, en la ocurrencia de enfermedad.

Las enfermedades no transmisibles (*ENT*) son en la actualidad la principal causa de mortalidad mundial. De los 57 millones de defunciones que se produjeron en el año 2008 en todo el mundo, 36 millones, esto es, casi las dos terceras partes, se debieron a *ENT*, principalmente enfermedades cardiovasculares, cáncer, diabetes y enfermedades pulmonares crónicas (Instituto Nacional de Cáncer. Ministerio de Salud de la Nación. Argentina, 2014).

El cáncer es la segunda causa de muerte en países desarrollados y figura entre las tres principales en países en desarrollo (World Health Organization, 2014). Fue responsable de 7,6 millones de muertes, de las cuales más de las dos terceras partes ocurrieron en países de ingresos bajos y medios (Instituto Nacional de Cáncer. Ministerio de Salud de la Nación. Argentina, 2014).

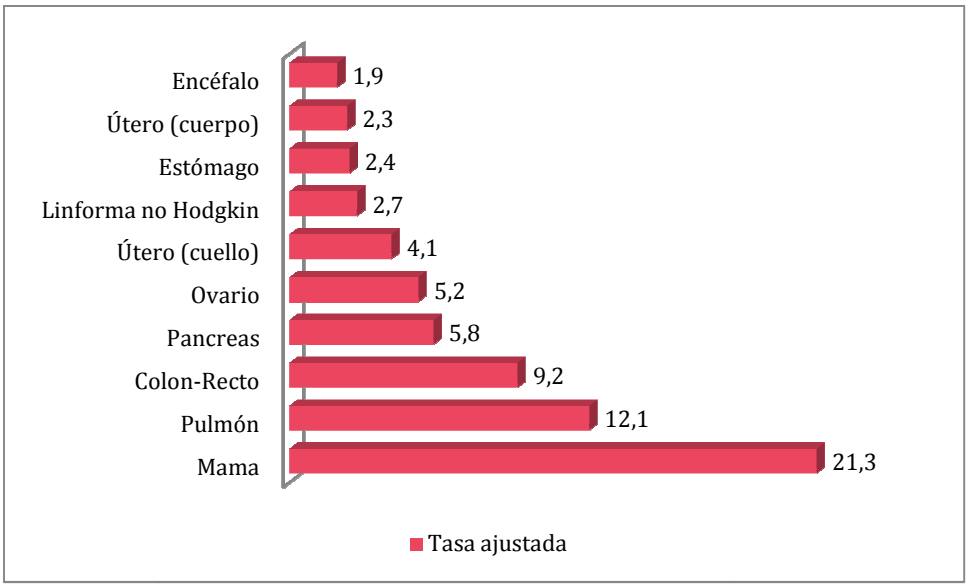
La Organización Panamericana de la Salud (*OPS*) incluye a Argentina entre uno de los países de América del Sur con mayor tasa de mortalidad por cáncer estandarizada por edad (*PAHO*, 2007). En el año 2012 murieron por cáncer casi 62.000 hombres y mujeres en este país. La región Centro, al ser la más poblada, registró más del 70% de estas defunciones. El cáncer de pulmón se encuentra en el primer lugar de importancia en todas las regiones, salvo en Cuyo donde es el cáncer de mama el más considerable. Le siguen en orden de significación el cáncer colorrectal y el de mama; excepto en el Noroeste y en el Sur, donde se observa en tercer lugar al cáncer de próstata y estómago, respectivamente. El cáncer de útero/cuello que no se registra dentro de los diez primeros a nivel país, se encuentra entre las primeras cinco causas de muerte por cáncer más importantes de la región Noreste; mientras que en la región Centro desaparece del ranking seleccionado. Para medir el impacto de esta enfermedad y sin discriminar por sexos, es necesario resaltar que el cáncer de pulmón fue responsable del 15% del total de defunciones por cáncer. Le siguen en orden de importancia el cáncer colorrectal y el de mama (11% y 9% respectivamente) (Instituto Nacional de Cáncer. Ministerio de Salud de la Nación. Argentina, 2014). Sin embargo, se observan diferencias importantes por sexo, tal como muestran los gráficos siguientes:

Figura 1.4. Tasas de mortalidad bruta ajustada por edad por 100.000 habitantes para los principales sitios tumorales registrados en el quinquenio 2007-2011. Argentina. Hombres.



Fuente: Atlas de Mortalidad por Cáncer en Argentina. (2007-2011)

Figura 1.5. Tasas de mortalidad bruta ajustada por edad por 100.000 habitantes para los principales sitios tumorales registrados en el quinquenio 2007-2011. Argentina. Mujeres.



Fuente: Atlas de Mortalidad por Cáncer en Argentina. (2007-2011)

A nivel local, la población total de la Provincia de Córdoba, registró una tasa cruda de mortalidad por cáncer de 154 por 100.000 habitantes para el año 2009 (Atlas de Mortalidad por Cáncer en Argentina, 2007-2011). Considerando los distintos sitios tumorales, se ha reportado que los cánceres con mayor tasas de mortalidad ajustada, por 100.000 habitantes, para el quinquenio 2007-2011, fueron los de pulmón (28,9), colon (15,0) y próstata (14,6) en la población masculina, y los de mama (18,2), colon (8,9) y pulmón (8,5) en la población femenina (Atlas de Mortalidad por Cáncer en Argentina, 2007-2011).

Un gran porcentaje de *ENT* son prevenibles y comparten los mismos factores de riesgo. Se estima que los cinco principales riesgos para la salud están relacionados con el comportamiento y la alimentación: índice alto de masa corporal, bajo consumo de frutas y hortalizas, inactividad física, consumo de tabaco e ingesta excesiva de alcohol. Estos factores causan el 30% de las muertes por cáncer. Aunque la edad es un factor de riesgo significativo, el consumo de tabaco es el más importante, causando el 22% de las muertes mundiales por cáncer en general y 71% de las muertes por cáncer de pulmón. Los cánceres causados por infecciones víricas, como el virus de las hepatitis B y C o por el Virus Papiloma Humanos, son responsables de hasta un 20% de las muertes por cáncer en los países de ingresos bajos y medios. En Argentina, las *ENT* son responsables de más del 60% del total de las defunciones que se producen anualmente en el país, 20% de las cuales corresponden

a tumores. Esto representa aproximadamente 60.000 muertes por año, de las cuales más del 90% se produce en personas mayores de 44 años de edad (Instituto Nacional de Cáncer. Ministerio de Salud de la Nación. Argentina, 2014).

Un factor ambiental que se reconoce como factor subyacente a los patrones de ocurrencia de diversas enfermedades, incluido el cáncer, es la dieta. La importancia de la nutrición en el proceso de salud-enfermedad, y en particular en el proceso carcinogénico, es un hecho que actualmente cuenta con evidencia suficiente y convincente en relación a múltiples aspectos (Pou, 2012). A partir de la evidencia epidemiológica existente sobre la relación entre la nutrición y el cáncer, se ha reportado que alrededor del 30% de la ocurrencia de cáncer a nivel mundial está asociada con la dieta (German Cancer Society, 2004). Esto significa que, aproximadamente 30% de todos los cánceres pueden ser prevenidos mediante una alimentación apropiada, actividad física y mantenimiento de un peso corporal adecuado (Donaldson, 2004).

Asimismo, otros estudios han demostrado un aumento de los riesgos de mortalidad e incidencia por cáncer, asociados a diferentes factores de exposición ambiental. Aballay *et al.* (2011) encontraron una asociación positiva entre la incidencia de cáncer de pulmón y el aumento del contenido de arsénico en las aguas subterráneas, tanto para hombres como para mujeres, en la provincia de Córdoba.

A pesar de que algunos autores proporcionan una clara evidencia de los efectos sobre la salud de las relaciones sociales y el nivel socioeconómico (Seeman y Crimmins, 2001), son pocos los estudios que consideran variables sociodemográficas en relación a la tasas de mortalidad por cáncer, en especial en países de medianos a bajos ingresos.

El monitoreo de la variación geográfica en la distribución de enfermedades y la investigación para comprender las razones subyacentes de dicha variación son habitualmente un punto de partida importante en Epidemiología. Se han identificado muchos factores de riesgo importantes como resultado de los hallazgos en el análisis de patrones geográficos de las enfermedades. El cáncer no es una excepción y mientras el mapeo de enfermedades infecciosas es una práctica bien establecida entre los epidemiólogos, la creación de mapas de enfermedades no transmisibles, como el cáncer, está menos desarrollada. En efecto, el cáncer muestra variaciones espaciales y el conocimiento de su patrón de ocurrencia es esencial para identificar grupos de población vulnerables, así como para desarrollar políticas de salud adecuadas para la prevención, el seguimiento y el control (Díaz *et al*, 2010[a]). En Argentina, sin embargo, existen pocos estudios de naturaleza epidemiológica vinculados a la distribución espacial de la incidencia

y mortalidad de cáncer (Díaz *et al*., 2009; Díaz *et al* 2010[a]). Si bien se han publicado atlas de mortalidad por cáncer en nuestro país (Análisis de Mortalidad por cáncer en Argentina, 1980-2006 y Atlas de Mortalidad por Cáncer en Argentina, 2007-2011), no constituyen una representación completa de la carga de esta enfermedad, debido principalmente a la falta de informaciones fiables de los registros de cáncer de base poblacional que aporten datos sobre la incidencia (aparición de nuevos casos) de esta enfermedad. Es más, no todas las provincias del país poseen registros, de carácter obligatorio, de la ocurrencia de esta patología.

En Córdoba existe, desde el año 2003, el Registro Provincial de Tumores (*RPT*), el cual abarca toda la provincia, cubriendo de este modo el 9% de la población argentina. La mayor parte de los estudios basados en datos del *RPT*, consideran la dimensión temporal para los tipos de cáncer más comunes, pero son pocos los antecedentes de análisis del patrón espacial de los datos.

Díaz *et al*. (2010[a]), empleando los datos correspondientes al año 2004, describen la existencia de un fuerte efecto espacial y de anidamiento para la incidencia del cáncer total en Córdoba. Asimismo, otros trabajos empleando estos datos permitieron identificar diversos factores de riesgo y lograron definir regiones vulnerables y no vulnerables (Díaz *et al*., 2010[b]). No obstante, son pocos los antecedentes de análisis de patrones espaciales asociados a datos de mortalidad por cáncer.

1.5.2. Cáncer de próstata

El cáncer de próstata es el segundo más común en hombres (*GLOBOCAN*, 2002) y es, a su vez, la tercera causa de muerte por cáncer en los hombres en Argentina (Instituto Nacional de Cáncer. Ministerio de Salud de la Nación. Argentina, 2014). La supervivencia promedio para este tipo de cáncer es relativamente alta, aunque es notablemente mayor en los países de altos ingresos (76% en los países de altos ingresos, en comparación con 45% en los países de bajos ingresos) (Niclis, 2011[b]).

El riesgo de cáncer de próstata aumenta con la edad y se diagnostica en muy pocas personas de 50 años o menos (Grönberg, 2003). Después de esta edad, la incidencia y las tasas de mortalidad aumentan casi exponencialmente. Niclis *et al.* (2011[b]) mostraron una alta correlación entre el envejecimiento y la mortalidad por cáncer de próstata. Asimismo, existe evidencia de que el calcio de origen lácteo y la carne roja juegan algún papel en el desarrollo de este tipo de cáncer. Por otra parte, los patrones que reducen el riesgo incluyen un mayor consumo de pescado, cereales, legumbres y verduras (Niclis, 2011[a]).

Publicaciones sobre la tendencia de las tasas de mortalidad por cáncer de próstata en Argentina (1980 a 2001) mostraron un incremento del 2,6% anual desde 1984, presentando una tendencia a la baja desde 1998. En efecto, la mortalidad por cáncer de próstata disminuye a partir del año 2000, luego de alcanzar su máximo valor de 16,41 por 100.000 hombres en 1998. El porcentaje estimado de cambio anual fue de aproximadamente 1% en el último decenio (Instituto Nacional de Cáncer. Ministerio de Salud de la Nación. Argentina, 2014). En la provincia de Córdoba, en cambio, las tasas de mortalidad estandarizadas por edad han disminuido desde 1980, en torno al 2% por año (Atlas of cancer mortality trends. Argentina, 1980-2001). Un trabajo más reciente (Niclis *et al.,* 2011[b]) muestra que, después de alcanzar un pico a mediados de 1990, éstas tasas se redujeron significativamente en Córdoba y en Argentina, un 1,9% y un 1,6% cada año respectivamente.

1.5.3. Cáncer de mama

El cáncer de mama es, en Argentina, la causa más común de muerte por tumores malignos en mujeres. En el año 2012 se registraron 5.530 defunciones por esta causa en nuestro país, lo que representa el 18,9% del total de muertes por cáncer en mujeres (Instituto Nacional de Cáncer. Ministerio de Salud de la Nación. Argentina, 2014).

En la provincia de Córdoba, los tumores mamarios representan 25% del total de todos los tumores y la primera causa de muerte por cáncer entre las mujeres (Díaz *et al,* 2009). Niclis *et al*. (2010) reportaron, para el período 1986–2006, una tendencia decreciente en la mortalidad por cáncer de mama en la provincia y un significativo efecto de cohorte, con un menor riesgo de morir por éste tumor en mujeres nacidas entre 1956 y 1966.

Al analizar la tendencia temporal de la tasa de mortalidad por cáncer de mama en la provincia de Córdoba, en el período 1986–2011, se destaca un incremento hasta el año 1996, momento a partir del cual comienzan a descender. El porcentaje estimado de cambio anual fue de -0,1% para el período 1996-2011 (Instituto Nacional de Cáncer. Ministerio de Salud de la Nación. Argentina, 2014). Si bien los resultados obtenidos no muestran un cambio significativo, se advierte una desaceleración en la tendencia decreciente de las tasas de mortalidad hacia el año 2001 (Tumas *et al.,* 2015).

El cáncer de mama está fuertemente ligado a factores sociales y culturales que cambian con el tiempo, como la urbanización (Hall *et al*, 2005) y a factores vinculados a la dieta (Fondo para la Investigación Mundial del Cáncer/Instituto Americano para la

Investigación del Cáncer, 2007). La dieta, además, difiere generalmente por región, principalmente entre zonas urbanas y rurales (Popkin, 2004).

1.6. Objetivos

1.6.1. Objetivo General

El objetivo general de éste trabajo es presentar un marco metodológico, dentro de la modelación estadística, que permita la identificación de patrones espaciales en las series de mortalidad de cáncer (mama y próstata) en la Provincia de Córdoba.

1.6.2. Objetivos Específicos

- Realizar un análisis exploratorio, utilizando herramientas que permite detectar la presencia de autocorrelación espacial.

- Proponer modelos para estimar el riesgo relativo en los diferentes departamentos y detectar factores socio-económicos significativos.

- Estudiar la bondad de ajuste de los modelos propuestos mediante estadísticos generalizados y técnicas exploratorias.

- Realizar análisis comparativos del desempeño de los diferentes modelos.

CAPÍTULO 2: MATERIALES Y MÉTODOS

2.1. Datos y fuentes de información

Se trabajó con los datos de defunciones obtenidos de la Dirección de Estadísticas e Información de Salud de la Nación Argentina y de la Dirección de Estadísticas y Censos de la Provincia de Córdoba. Asimismo, se obtuvieron estimaciones poblacionales, para cada año intercensal, por interpolación exponencial a partir de la información censal 1980, 1991, 2001 y 2010 (INDEC).

A partir de ésta información, se calcularon las tasas de mortalidad (por 100.000 habitantes) específicas por sexo y estandarizadas por edad según el método directo (población mundial de referencia) y el método indirecto (tasas de la población de referencia), para la provincia de Córdoba, sus 26 departamentos y para la serie temporal 1986-2011, considerando:

- Cáncer de Mama (CIE-9 174 y CIE-10 C50)
- Cáncer de Próstata (CIE-9 185 y CIE-10 C61)

Además, para el año 2001 se incorporaron al análisis, a partir de la información del Censo Nacional de Población, Hogares y Viviendas, las siguientes variables socio-demográficas por departamento:

- Porcentaje de la población total sin cobertura de salud (obra social y/o plan de salud privado o mutual).
- Porcentaje de la población total desocupada.
- Porcentaje de la población con necesidades básicas insatisfechas.

Según la metodología utilizada en *La pobreza en la Argentina* (INDEC, 1984), los hogares con necesidades básicas insatisfechas son los que presentan al menos uno de los siguientes indicadores de privación:

- Hacinamiento: hogares con más de tres personas por cuarto;
- Vivienda: hogares en una vivienda de tipo inconveniente (pieza de inquilino, vivienda precaria u otro tipo, excluyendo casa, departamento y rancho);
- Condiciones sanitarias: hogares que no tienen ningún tipo de retrete;
- Asistencia escolar: hogares que tienen algún niño en edad escolar (6 a 12 años) que no asisten a la escuela;
- Capacidad de subsistencia: hogares que tienen cuatro o más personas por miembro ocupado y, además, cuyo jefe no haya completado tercer grado de escolaridad primaria.

2.2. Metodología

2.2.1. Cálculo de tasas y *CMEs*

Para cada sitio tumoral se calcularon, por departamentos, las tasas crudas y las tasas ajustadas por edad (*TMEs*), según población mundial. Las tasas se expresaron en defunciones por 100.000 personas en riesgo por año. Para los ajustes de las tasas de mortalidad se utilizó el paquete estadístico *Stata* (*Data analysis and Statistical software*) versión 13 que, a través del comando *dstdize*, permite calcular las tasas ajustadas por edad por el método directo; la población de referencia corresponde a la de OMS, presentada con anterioridad.

Posteriormente se calcularon los cocientes de mortalidad estandarizados (*CMEs*) que, como se indicó, representan la relación entre el número observado de defunciones y el número esperado de las mismas en un departamento. El número esperado se calculó aplicando la estandarización por el método indirecto, empleando el comando *istdize* de *Stata* 13. Así, las

tasas específicas (por edad) para toda la Argentina, se aplicaron a la estructura por edad de los departamentos de la provincia de Córdoba. Dichas tasas se presentan en la Tabla 2.1.

Tabla 2.1. Tasas de mortalidad específicas por grupo de edad para Argentina (por 100.000). Período 1997-2011. Cáncer de próstata y mama.

Grupo de edad	Tasa Cáncer de próstata	Tasa Cáncer de mama
0-19	0,05	0,05
20-24	0,05	0,25
25-29	0,10	1,05
30-34	0,05	3,85
35-39	0,10	9,15
40-44	0,25	17,8
45-49	0,95	30,3
50-54	4,05	44,75
55-59	12,15	58,40
60-64	29,75	71,95
65-69	70,75	84,40
70-74	139,10	100,50
75-79	262,35	129,25
80+	591,55	229,40

Fuente: Atlas de mortalidad por cáncer en Argentina 1997-2001 y 2007-2011

En todos los casos se presentaron, además, los intervalos del 95% de confianza para las *TMEs* y los *CMEs*. El intervalo de confianza exacto se calculó para cada *CME* estimado suponiendo un proceso de Poisson, como se describe en Breslow y Day (1987). Dichos intervalos se representaron gráficamente, mediante Diagramas de puntos, con sus respectivas bandas de confianza, empleando el software *InfoStat.*

Finalmente, y a los fines de visualizar la distribución geográfica por departamentos, de las tasas de mortalidad estandarizadas y los cocientes de mortalidad, se representaron sus valores en el mapa. Su utilizaron escalas monocromáticas, tal como sugieren Pickle *et al.,* (1999), para inducir a una menor variabilidad y los límites de clases se definieron empleando cinco cuantiles determinados a partir de la distribución observada.

2.2.2. Análisis de la autocorrelación espacial

Siguiendo la organización propuesta en la sección Marco Teórico, la primera cuestión a resolver es la construcción de la estructura de vecindades. Dado que no existe una teoría que guíe la construcción de esta matriz de contactos para los departamentos de Córdoba, se consideraron los diferentes criterios geográficos habitualmente utilizados.

Empleando el comando *poly2nb* del paquete *spdep* del software *R*, se determinó una lista de vecinos basada en regiones con límites contiguos, que comparten uno o más puntos límites, es decir, se aplicó el criterio de contigüidad tipo reina. También se determinaron los 4-vecinos más cercanos, empleando la función *knearneigh* del software *R,* teniendo en cuenta que es recomendable analizar si el esquema de autocorrelación espacial detectado entre regiones vecinas es extensible a regiones alejadas en el espacio (Moreno y Vayá, 2000).

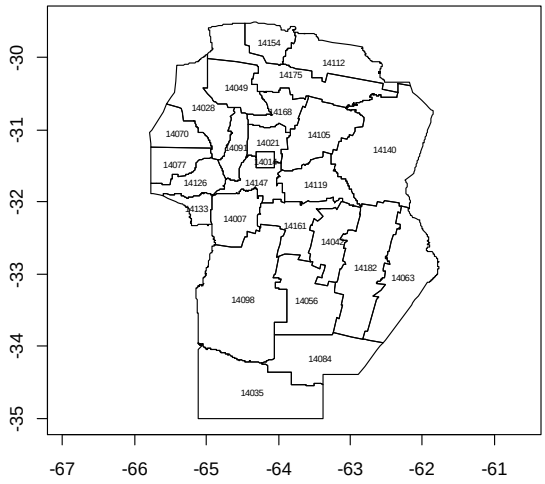
Posteriormente, se construyeron los mapas de contactos (Figura 2.1). En el caso de áreas irregulares, como puede observarse, el criterio de vecinos más cercanos arroja resultados similares a los obtenidos con el criterio tipo reina. Para aplicar los criterios, previamente se definió el centroide de cada departamento (Figura 2.2). Una alternativa al centroide es identificar la ciudad cabecera del departamento, lo que modifica en mayor medida la estructura de vecindad a partir del criterio de 4-vecinos más cercanos. El número de vecinos con el criterio de contigüidad tipo reina varía, en tanto que para el criterio de vecinos más cercanos es constante e igual a 4 (Tabla 2.2). Por su parte, el número de enlaces no nulos, para la contigüidad tipo reina es de 122, mientras que con el criterio de los 4 vecinos más cercanos resultó igual a 104.

Figura 2.1. Mapas de contactos para diferentes criterios de contigüidad. Provincia de Córdoba.

Contigüidad tipo reina 4-vecinos más cercanos

Fuente: elaboración propia

Figura 2.2. . Mapa de límites y centroides de cada departamento. Provincia de Córdoba.



Fuente: elaboración propia

Tabla 2.2. Número de vecinos de cada departamento según criterio de contigüidad tipo reina. Provincia de Córdoba.

Departamento	Número de vecinos	Departamento	Número de vecinos
Calamuchita	5	Rio Cuarto	5
Capital	2	Rio Primero	6
Colón	5	Rio Seco	3
Cruz del Eje	6	Rio Segundo	6
General Roca	2	San Alberto	6
General San Martin	4	San Javier	3
Ischilín	4	San Justo	6
Juarez Celman	5	Santa María	9
Marcos Juarez	3	Sobremonte	2
Minas	2	Tercero Arriba	6
Pocho	3	Totoral	5
Pte. Roque Saenz Peña	5	Tulumba	7
Punilla	6	Unión	6

Fuente: elaboración propia

Como se ha mencionado previamente, el uso de criterios geográficos genera un tratamiento exógeno de la matriz de contactos, evitando problemas de inferencia. La selección, entre las estructuras de contactos, puede ser realizada en una etapa más avanzada, una vez definido

el modelo espacial más adecuado (Herrera, *et al,* 2012). Usando la función *nbdists* de *R*, puede calcularse una lista de vectores de distancias correspondiente al objeto vecino. El mayor valor será la distancia mínima necesaria para asegurarse de que todas las áreas están vinculadas con, al menos, un vecino.

Posteriormente se determinaron las ponderaciones espaciales w_{ij}, empleando la función *nb2listw* de *R,* que toma un objeto de la lista de vecinos y lo convierte en un peso. El software *R* considera varios estilos de conversión, siendo el predeterminado aquel donde los pesos están estandarizados para que su suma sea igual a uno, esto es, la estandarización por fila (*W**). En este caso, las ponderaciones en zonas con pocos vecinos serán más grandes que las que se originan en las zonas con muchos vecinos.

Incorporando el argumento *style="B"* se generó una matriz de pesos binaria, asignado un 1 para el caso en que dos observaciones sean vecinas y 0 en caso contrario. En este caso, las sumas de los pesos para las áreas diferirán según el número de vecinos que tengan.También se utilizaron otros dos argumentos que asignan pesos iguales para todos los enlaces: *style="C"* que proporciona una matriz estandarizada globalmente y *style="U"* que estandariza los pesos de manera que la suma total sea 1.

Finalmente, se aplicó el esquema de codificación de estabilización de varianza propuesta por Tiefelsdorf *et al*. (1999) y que se aparece como *style = "S"* en el paquete *R.* Con este esquema, en comparación con *W**, los pesos varían menos pero las sumas por filas de los pesos varían más (en *W** son siempre iguales a 1). Asimismo, se registra una menor variabilidad en las sumas por filas con respecto a los estilos *B*, *C* y *U*.

Se ajustaron pruebas de autocorrelación espacial a los casos observados y a los *CMEs,* globales (*I* de Moran y *IEB*) y locales (*I* de Moran), empleando la estandarización por filas de la matriz *W.* Si bien no existe una razón contundente que justifique este hecho, la posibilidad de ponderar por igual la influencia total que recibe cada región de sus vecinos, con independencia del número total de vecinos de cada una de ellas, explicaría su uso generalizado. No obstante, tal y como expone Anselin (1988), la estandarización de *W* no siempre es adecuada, especialmente cuando ésta se basa en un concepto de distancia dado que, en este caso, la matriz estandarizada carecería de significado. Sin embargo, dicha estandarización facilita la interpretación de los coeficientes autorregresivos del modelo estimado al asimilarlos a un coeficiente de correlación, asegurando además que los parámetros espaciales estimados sean comparables entre los distintos modelos propuestos. También se presentaron los gráficos Scatterplot correspondientes a los casos observados o_i para cada tipo de sitio tumoral.

2.2.3. Modelos

Empleando el comando *gllamm* del software *Stata* se ajustó el modelo Poisson con intercepto aleatorio, sin covariables, para la Provincia de Córdoba, período 1986-2011. Para la estimación de los parámetros se empleó la aproximación mejorada *Cuadratura adaptativa,* utilizando como valor de a_i la media a posteriori de ζ_i y como b_i la desviación estándar a posteriori. Este modelo, combinado con la predicción de los *CMEs* a través del método empírico Bayesiano, que se realizó con el comando *gllapred* de *Stata,* es sugerido por Rabe-Hesketh y Skrondal (2005).

Para detectar la presencia de sobredispersión se aplicaron las pruebas de puntuación desarrolladas por Dean (1992), previstas en la librería *DCluster* de *R*, cuyas hipótesis son H0: $\psi = 0$ vs. Ha: $\psi > 0$.

En relación al modelo Poisson-Gamma, teniendo en cuenta que ν y α son desconocidos, fue necesario estimarlos. Esto se realizó, a partir de los datos, utilizando el método de los momentos propuesto por Clayton y Kaldor (1987) para producir estimaciones Empíricas Bayesianas (*EB*), implementadas en el paquete *DCluster* de *R*. Estas estimaciones se calculan por medio de un procedimiento iterativo, utilizando dos ecuaciones basadas en la media y en varianza estimada.

Posteriormente se estimaron los riesgos relativos para cada modelo, construyendo los respectivos mapas. Asimismo, utilizando la función *probmap* de *R,* se obtuvieron los *p-valores* correspondientes a los diferentes modelos, empleados luego en la construcción de los mapas de probabilidad.

En el ajuste del modelo Poisson con intercepto aleatorio, se consideraron en una segunda etapa, como covariables:

- Porcentaje de la población total sin cobertura de obra social y/o plan de salud privado o mutual (sin cob.),
- Porcentaje de la población con necesidades básicas insatisfechas (nbi),
- Porcentaje de la población total desocupada (desocup)

disponibles para el año 2001 y para toda la provincia, a partir del Censo Nacional de Población, Hogares y Viviendas.

Se analizó en primer lugar la correlación entre éstas covariables, siendo alta entre el Porcentaje de la población con necesidades básicas insatisfechas y las otras variables (0,92 con el Porcentaje de la población total sin cobertura de obra social y -0,78 con el Porcentaje de la población total desocupada).

La correlación entre el Porcentaje de la población total sin cobertura de obra social y el Porcentaje de la población total desocupada es, en cambio, bastante baja (0,017). Es por ello, que se decidió mantener éstas dos covariables en el modelo.

En *R* fue posible ajustar además, para el año 2001, el modelo *SAR* mediante el uso de la función *spautolm,* en el paquete *spdep*. Waller y Gotway (2004) proponen tomar como variable dependiente una transformación del *CME*:

$$z_i = log\,\frac{1000(CME+1)}{n_i}$$

Para determinar si existe correlación espacial residual luego de ajustar el modelo, se analizó el *p-valor* de la prueba de cociente de verosimilitud y se aplicó el test de Moran sobre los residuos del modelo.

También se aplicaron pruebas de homogeneidad de los riesgos relativos, para los datos de mortalidad del período 1986-2011, a fin de detectar la existencia de *clusters,* empleando el paquete *DCluster* de *R* (Gómez- Rubio *et al.*, 2005), que utiliza diferentes modelos para calcular la significación de los valores observados. Se realizó la prueba Chi-cuadrado así como el test de homogeneidad propuesto por Potthoff and Whittinghill (1966). Finalmente, se aplicó el test de Tango para detectar la existencia global de *clusters*.

Dado que el modelo Poisson estándar se encuentra anidado en el modelo Poisson con intercepto aleatorio, se utilizó el Test de cociente de verosimilitud para realizar comparaciones, empleando el comando *lrtest* de *Stata.* El *AIC* y el *BIC*, calculados empleando el comando *estimates stats* de *Stata*, se aplicaron también para evaluar la bondad de ajuste de los modelos.

CAPÍTULO 3: RESULTADOS

Tras la implementación de la metodología antes descripta, se presentan los resultados de esta tesis, organizados en los siguientes ejes:

3.1. Distribución espacial de las Tasas de mortalidad estandarizadas por edad (por 100.000). Provincia de Córdoba. Período 1986-2011. Cáncer de Próstata y Mama.

3.2. Distribución espacial de los Cocientes de mortalidad estandarizados. Provincia de Córdoba. Período 1986-2011. Cáncer de Próstata y Mama.

3.3. Análisis de la autocorrelación espacial.

3.4. Estimación del riesgo relativo en áreas pequeñas (*disease mapping*)

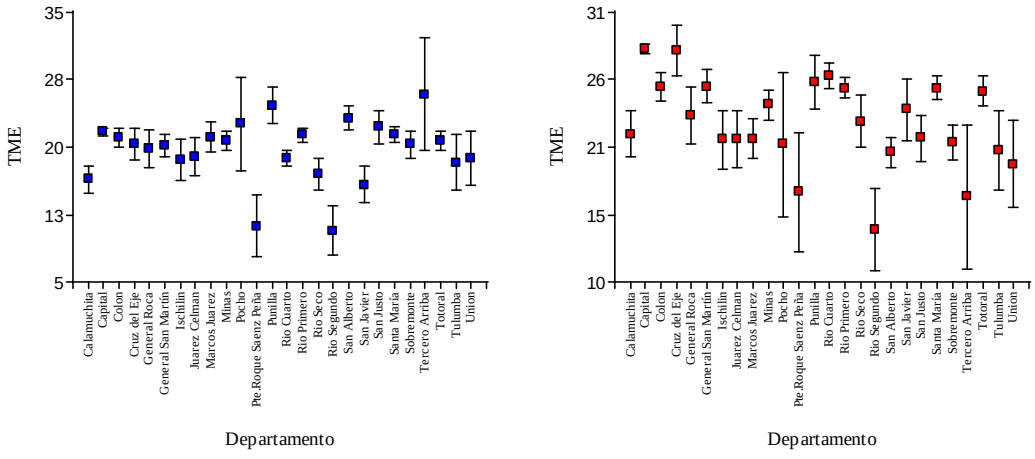
3.5. Estudios de asociación geográfica (*ecological analysis*)

3.6. Aglomeraciones de casos (*disease clustering*)

3.1. Distribución espacial de las Tasas de mortalidad estandarizadas por edad (por 100.000). Provincia de Córdoba. Período 1986-2011. Cáncer de Próstata y Mama.

Se presentan a continuación, por departamentos, las tasas de mortalidad crudas, las tasas de mortalidad estandarizadas por edad y los intervalos del 95% de confianza para las tasas estandarizadas, por 100.000 habitantes, para cáncer de próstata y mama, en la Provincia de Córdoba, período 1986-2011 (Tablas 3.1. y 3.2.). A los efectos de visualizar la amplitud de los respectivos intervalos del 95% de confianza para la *TME*, se realizaron gráficos (Figura 3.1).

Figura 3.1.Tasas de mortalidad estandarizadas por edad e intervalos de confianza (por 100.000). Provincia de Córdoba. Período 1986-2011.



Cáncer de próstata. Hombres Cáncer de mama. Mujeres

Fuente: elaboración propia

Tabla 3.1.Tasas de mortalidad crudas (*TCMs*), tasas de mortalidad estandarizadas por edad (*TMEs*), desvío estándar (DE(*TME*))e intervalos de confianza (IC) para las tasas estandarizadas (por 100.000). Provincia de Córdoba. Período 1986-2011. Cáncer de próstata. Hombres.

Departamento	*TCM*	*TME*	DE(*TME*)	IC(95%) para *TME*	
				LI	LS
Calamuchita	20,90	16,45	1,50	13,51	19,38
Capital	17,77	21,74	0,42	20,92	22,56
Colón	18,28	21,09	1,09	18,95	23,23
Cruz del Eje	21,38	20,38	1,71	17,03	23,74
General Roca	21,01	19,86	2,07	15,81	23,91
General San Martin	21,74	20,19	1,30	17,65	22,73
Ischilín	21,76	18,64	2,28	14,18	23,09
Juarez Celman	21,59	18,96	2,12	14,80	23,12
Marcos Juarez	23,06	21,10	1,66	17,85	24,34
Minas	28,66	20,75	1,09	18,61	22,89
Pocho	29,04	22,64	5,20	12,45	32,82
Pte. Roque Saenz Peña	15,80	11,25	3,41	4,57	17,94
Punilla	30,64	24,66	2,10	20,54	28,78
Rio Cuarto	24,13	18,81	0,90	17,04	20,58
Rio Primero	23,84	21,37	0,82	19,76	22,98
Rio Seco	16,74	16,95	1,79	13,45	20,46
Rio Segundo	9,95	10,73	2,68	5,47	15,99
San Alberto	24,55	23,26	1,37	20,58	25,95
San Javier	15,78	15,80	2,01	11,86	19,73
San Justo	24,27	22,27	1,86	18,63	25,91
Santa María	25,39	21,48	0,88	19,76	23,20
Sobremonte	18,21	20,27	1,47	17,40	23,15
Tercero Arriba	28,13	25,90	6,29	13,57	38,23
Totoral	24,56	20,70	1,14	18,46	22,94
Tulumba	17,31	18,31	3,05	12,32	24,29
Unión	24,03	18,68	3,00	12,79	24,56

Fuente: elaboración propia

LI: límite inferior; LS: límite superior para el Intervalo del 95% de confianza (IC(95%)).

Tabla 3.2.Tasas de mortalidad crudas, tasas de mortalidad estandarizadas por edad (*TMEs*), desvío estándar (DE(*TME*)) e intervalos de confianza (IC) para las tasas estandarizadas (por 100.000). Provincia de Córdoba . Período 1986-2011. Cáncer de mama. Mujeres.

Departamento	*TMC*	*TME*	DE(*TME*)	IC(95%) para *TME*	
				LI	LS
Calamuchita	26,71	21,49	1,79	17,98	25,00
Capital	31,86	28,14	0,39	27,64	28,90
Colón	26,44	25,21	1,08	23,09	27,32
Cruz del Eje	31,77	28,04	1,96	24,20	31,87
General Roca	27,25	22,96	2,20	18,66	27,26
General San Martin	32,88	25,22	1,32	22,64	27,93
Ischilín	28,95	21,08	2,29	16,59	25,57
Juarez Celman	24,39	21,14	2,23	16,76	25,52
Marcos Juarez	26,79	21,16	1,58	18,06	24,26
Minas	35,91	23,78	1,16	21,52	26,05
Pocho	25,14	20,70	5,66	9,6	31,80
Pte. Roque Saenz Peña	21,73	17,01	4,65	7,90	26,12
Punilla	35,01	25,58	2,12	21,42	29,74
Rio Cuarto	37,04	26,04	1,01	24,06	28,02
Rio Primero	32,61	25,13	0,83	23,51	26,76
Rio Seco	23,88	22,50	2,04	18,49	26,50
Rio Segundo	12,83	14,05	3,24	7,69	20,40
San Alberto	24,70	20,07	1,19	17,75	22,40
San Javier	25,33	23,45	2,42	18,71	29,19
San Justo	23,45	21,18	1,81	17,63	24,72
Santa María	34,96	25,13	0,90	23,37	26,88
Sobremonte	23,26	20,82	1,36	18,17	23,48
Tercero Arriba	20,28	16,64	5,63	6,13	27,15
Totoral	33,08	24,87	1,20	22,52	27,21
Tulumba	21,99	20,29	3,10	14,20	26,37
Unión	23,77	19,18	3,36	12,59	25,77

Fuente: elaboración propia

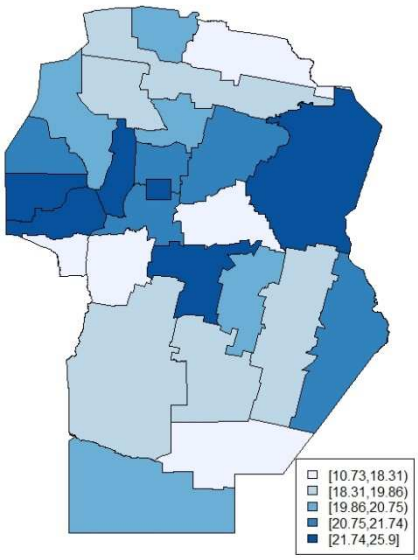
LI: límite inferior; LS: límite superior para el Intervalo del 95% de confianza (IC(95%)).

La distribución espacial de las *TMEs* (Figura 3.2) se presenta en mapas, donde se han utilizado diferentes colores para cada tipo de tumor y distintas tonalidades, dentro de cada color, para indicar la magnitud de las tasas, según los cuantiles.

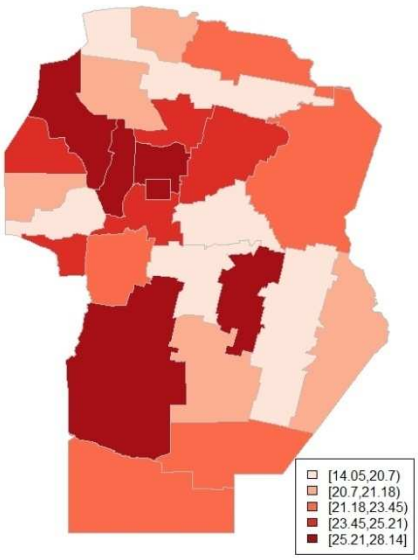
Para el cáncer de próstata, la distribución se encuentra concentrada fundamentalmente en los departamentos del centro de la provincia. Se observa que los departamentos de Pocho, San Alberto, Punilla, Tercero Arriba, Capital y San Justo, registraron tasas en el quintil superior. Las *TMEs* más bajas se registraron en los departamentos Río Segundo y Presidente Roque Sáenz Peña. En el cáncer de mama, se destaca un gradiente, que abarca el noroeste provincial, además del departamento General San Martín y Río Cuarto, presentando tasas que se ubican en los quintiles superiores de la distribución. Las menores *TMEs* se registraron en los departamentos Río Segundo, Tercero Arriba y Presidente Roque Sáenz

Peña, aunque los dos últimos presentan intervalos de confianza bastante amplios. Por otra parte, se destaca que las tasas de mortalidad en el departamento Capital son elevadas, en ambos sexos y en los dos cánceres estudiados.

Figura 3.2.Mapeo de las Tasas de mortalidad estandarizadas por edad (por 100.000). Provincia de Córdoba. Período 1986-2011.



Cáncer de próstata. Hombres



Cáncer de mama. Mujeres

Fuente: elaboración propia

3.2. Distribución espacial de los Cocientes de mortalidad estandarizados. Provincia de Córdoba. Período 1986-2011. Cáncer de Próstata y Mama.

A partir de los casos observados (o_i) y los casos esperados (e_i) para cada departamento de la Provincia de Córdoba, se calcularon los Cocientes de mortalidad estandarizados y sus intervalos del 95% de confianza, para el período 1986-2011, considerando cada tipo de tumor (Tablas 3.3. y 3.4.).

Tabla 3.3. Casos observados (o_i), casos esperados (e_i), cocientes de mortalidad estandarizados (*CMEs*) e intervalos de confianza (IC) para los cocientes estandarizados. Provincia de Córdoba. Período 1986-2011. Cáncer de próstata. Hombres.

Departamento	Casos observados (o_i)	Casos Esperados (e_i)	*CME* (en %)	IC(95%) para *CME*	
				LI	LS
Calamuchita	121	149,85	80,75	67,00	96,48
Capital	2711	2533,39	107,01	103,02	111,12
Colón	374	359,20	104,12	93,83	115,23
Cruz del Eje	142	140,23	101,26	85,29	119,65
General Roca	93	96,73	96,14	77,60	117,78
General San Martin	311	315,19	98,67	88,01	110,27
Ischilín	80	85,37	93,71	74,31	116,63
Juarez Celman	163	155,92	104,54	89,11	121,88
Marcos Juarez	363	354,57	102,38	92,12	113,47
Minas	19	17,07	111,31	67,00	172,79
Pocho	11	20,51	53,63	26,78	95,98
Pte. Roque Saenz Peña	138	112,86	122,28	102,72	144,46
Punilla	434	465,21	93,29	84,72	102,50
Rio Cuarto	381	642,19	59,33	53,52	65,59
Rio Primero	90	105,93	84,96	68,32	104,43
Rio Seco	16	28,70	55,75	31,87	90,54
Rio Segundo	289	250,25	115,48	102,55	129,59
San Alberto	62	79,66	77,83	59,67	99,78
San Javier	144	131,74	109,31	92,18	128,69,
San Justo	604	567,30	106,47	98,15	115,31
Santa María	192	190,14	100,98	87,20	116,31
Sobremonte	17	13,58	125,18	72,91	200,38
Tercero Arriba	329	321,21	102,42	91,65	114,11
Totoral	36	39,95	90,11	63,12	124,76
Tulumba	39	43,10	90,49	64,34	123,69
Unión	342	313,76	109,00	97,75	121,19

Fuente: elaboración propia

LI: límite inferior; LS: límite superior para el Intervalo del 95% de confianza (IC(95%)).

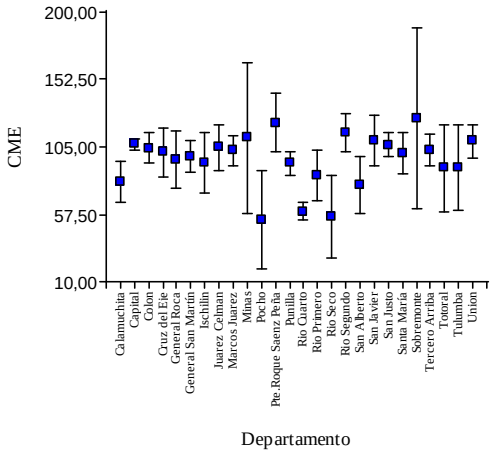
Tabla 3.4 .Casos observados (o_i), casos esperados (e_i), cocientes de mortalidad estandarizados (*CMEs*) e intervalo de confianza (IC) para los cocientes estandarizados. Provincia de Córdoba. Período 1986-2011. Cáncer de mama. Mujeres.

Departamento	Casos observados (o_i)	Casos Esperados (e_i)	CME (%)	IC(95%) para *CME*	
				LI	LS
Calamuchita	154	163,28	94,32	80,01	110,44
Capital	5324	4191,03	127,03	123,64	130,49
Colón	562	492,71	114,06	104,83	123,89
Cruz del Eje	216	170,37	126,78	110,44	144,87
General Roca	116	112,71	102,92	85,05	123,44
General San Martin	481	437,92	109,84	100,24	120,11
Ischilín	96	104,29	92,05	74,57	112,42
Juarez Celman	189	197,25	95,82	82,64	110,50
Marcos Juarez	477	443,36	107,59	98,15	117,69
Minas	15	16,73	89,66	50,19	147,89
Pocho	14	18,11	77,31	42,26	129,71
Pte. Roque Saenz Peña	158	135,92	116,24	98,82	135,85
Punilla	725	606,84	119,47	109,32	.128,49
Rio Cuarto	978	868,46	112,61	105,66	119,90
Rio Primero	126	128,39	98,14	81,75	116,85
Rio Seco	19	31,81	59,73	35,97	93,29
Rio Segundo	300	328,53	91,32	81,27	102,26
San Alberto	99	92,56	106,96	86,93	130,22
San Javier	146	159,95	91,28	77,07	107,34
San Justo	861	750,74	114,69	107,15	122,61
Santa María	245	260,67	93,99	82,59	106,52
Sobremonte	11	13,42	81,97	40,89	146,58
Tercero Arriba	462	415,48	111,20	101,29	121,81
Totoral	44	47,29	93,04	67,61	124,92
Tulumba	35	41,25	84,85	59,10	117,99
Unión	447	397,46	112,46	102,28	123,39

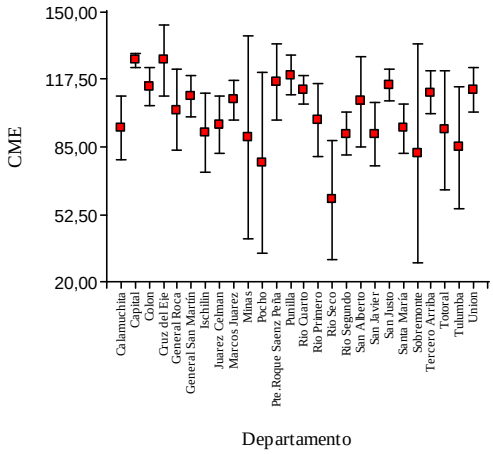
Fuente: elaboración propia

Nota: LI: límite inferior; LS: límite superior para el Intervalo del 95% de confianza (IC(95%)).

Figura 3.3. Cocientes de mortalidad estandarizados e intervalo de confianza. Provincia de Córdoba. Período 1986-2011.



Cáncer de próstata. Hombres

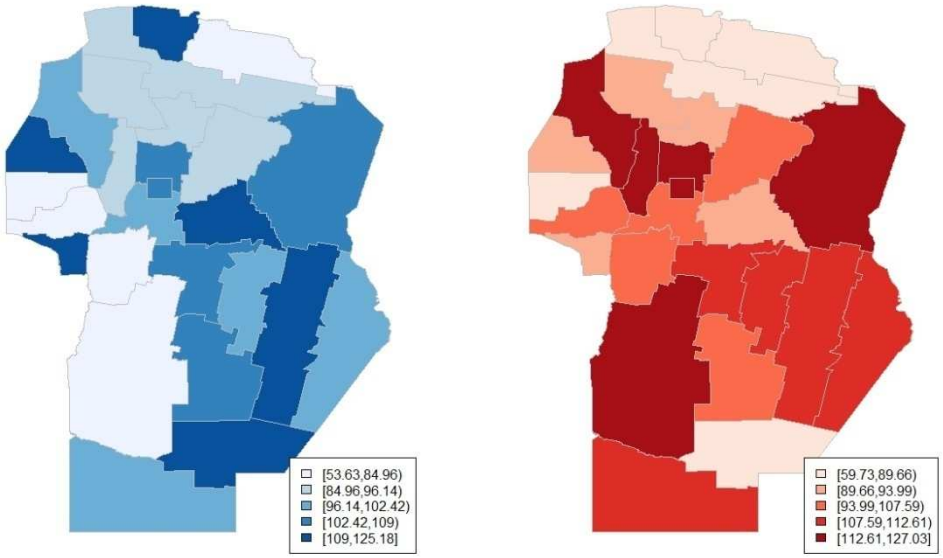


Cáncer de mama. Mujeres

Fuente: elaboración propia

La distribución espacial (figura 3.4), para el cáncer de próstata, exhibe una concentración de valores elevados del *CME* hacia el centro-este provincial. Se trataría de una zona en riesgo, teniendo en cuenta que los casos observados superan a los esperados, en función de la estructura poblacional. Dentro de ésta zona, los departamentos Minas y Sobremonte presentaron valores elevados del *CME*, además de amplios intervalos de confianza, en comparación con el resto de los departamentos. Sin embargo, hay que tener en cuenta que son departamentos con baja densidad poblacional, lo que condiciona las interpretaciones. Los *CMEs* más bajos se presentaron en los departamentos Pocho, Río Seco y Río Cuarto, los dos primeros con un reducido tamaño poblacional. En relación al cáncer de mama, a partir del mapa, se puede distinguir un gradiente en el noroeste de la provincia, que involucra a los departamentos Cruz del Eje, Punilla, Capital y Cólon, con *CMEs* en el quintil superior. El departamento Capital registró el *CME* más alto, seguido de Cruz del Eje. Conforman un área en riesgo, además, los departamentos de Tercero Arriba, General San Martín, Unión y Marcos Juarez. El menor *CME* corresponde al departamento Río Seco, aunque con un intervalo de confianza poco preciso, seguramente asociado al tamaño pequeño de la población de este departamento. Sobremonte y Tulumba, también registraron bajos valores del *CME*, aunque son departamentos particulares por la baja densidad poblacional que presentan.

Figura 3.4. Mapeo de los Cocientes de mortalidad estandarizados en porcentaje. Provincia de Córdoba. Período 1986-2011.



Cáncer de próstata. Hombres Cáncer de mama. Mujeres

Fuente: elaboración propia

3.3. Análisis de la autocorrelación espacial

Habiendo aplicado los criterios de contigüidad tipo reina y 4 vecinos más cercanos, se calcularon las principales medidas descriptivas correspondientes a las distancias entre vecinos. Los resultados (Tabla 3.5) evidencian escasas diferencias entre ambos criterios. El criterio tipo reina presenta un máximo superior al criterio de vecinos más cercanos, aunque las medidas de tendencia central registran diferencias que no superan los 0,10 puntos. Posteriormente se determinaron las ponderaciones espaciales para diferentes estilos de conversión (Tablas 3.6 y 3.7). Para el criterio de contigüidad tipo reina, los pesos mínimo y máximo difieren entre sí, salvo cuando se trabaja con la matriz de pesos binaria *B* o la matriz estandarizada *U.* En el criterio de vecinos más cercanos, los pesos máximos y mínimos coinciden, lo que resulta lógico ya que el número de vecinos es fijo e igual a 4. La suma total de los pesos coincide con el número de departamentos para los estilos de conversión *W**, *C* y *S.* Para la matriz binaria, en cambio, dicha suma es igual al número de enlaces no nulos y para la matriz *U* igual a la unidad. Los valores mínimos y máximos de las ponderaciones correspondientes a cada criterio de contigüidad se presentan en forma

comparativa (Tabla 3.8). Como puede observarse, el criterio tipo reina asigna un peso mínimo superior, siendo las ponderaciones máximas iguales con ambos criterios.

Tabla 3.5. Estadísticos descriptivos para las distancias entre vecinos según criterio de contigüidad. Provincia de Córdoba. Período 1986-2011.

Criterio de contigüidad	Mínimo	Primer Cuartil	Mediana	Tercer Cuartil	Media	Máximo
Tipo reina	0,27	0,56	0,72	0,90	0,77	2,34
Vecinos más cercanos	0,27	0,58	0,82	1,06	0,85	1,80

Fuente: elaboración propia

Tabla 3.6. Estadísticos descriptivos según estilo de conversión, para el criterio de contigüidad tipo reina. Provincia de Córdoba. Período 1986-2011.

Estilo de conversión	Peso mínimo	Peso máximo	Mínimo de la suma de los pesos por fila	Máximo de la suma de los pesos por fila	Suma total de los pesos
W*	0,111[(1)]	0,500[(2)]	1,000	1,000	26
B	1,000	1,000	2,000	9,000	122
C	0,213[(3)]	0,213	0,426	1,918	26
U	0,008[(4)]	0,008	0,016	0,074	1
S	0,157	0,333	0,666	1,412	26

Fuente: elaboración propia

[(1)] Surge de dividir 1 por 9 que es el número máximo de vecinos para este criterio.

[(2)] Surge de dividir 1 por 2 que es el número mínimo de vecinos para este criterio

[(3)] Surge de dividir el número de departamentos n= 26 por el número de enlaces no nulos.

[(4)] Surge de dividir 1 por el número de enlaces no nulos.

Tabla 3.7. Estadísticos descriptivos según estilo de conversión, para el criterio de contigüidad de 4 vecinos más cercanos. Provincia de Córdoba. Período 1986-2011.

Estilo de conversión	Peso mínimo	Peso máximo	Mínimo de la suma de los pesos por fila	Máximo de la suma de los pesos por fila	Suma total de los pesos
W*	0,250[(1)]	0,250	1,000	1,000	26
B	1,000	1,000	4,000	4,000	104
C	0,250	0,250	1,000	1,000	26
U	0,010	0,010	0,038	0,038	1
S	0,250	0,250	1,000	1,000	26

Fuente: elaboración propia

[(1)] Surge de dividir 1 por 4 que es el número de vecinos para este criterio.

Tabla 3.8. Estadísticos descriptivos según criterio de contigüidad, para la matriz de distancias (*D*). Provincia de Córdoba. Período 1986-2011.

Criterio de contigüidad	Peso mínimo	Peso máximo	Mínimo de la suma de los	Máximo de la suma de los
Tipo reina	554,2	3746,0	1711,0	13970,0
Vecinos más cercanos	428,2	3746,0	2812,0	10480,0

Fuente: elaboración propia

Tal como se discutió en el apartado anterior, se puede concluir que existe una tendencia hacia la concentración en el espacio de valores similares del *CME*, siendo poco probable una distribución aleatoria de dicha variable (Figura 3.4). No obstante, si bien esta conclusión parece razonable, se aplicaron pruebas de autocorrelación espacial, globales y locales, empleando la estandarización por filas de la matriz *W* (Tablas 3.9, 3.10 y 3.11). El índice global de Moran, aplicado sobre los casos observados correspondientes al cáncer de próstata, arrojó un valor positivo, mayor que el valor esperado y estadísticamente significativo a un nivel de 0,10. Esto indica que la distribución de los casos observados para este tipo de tumor, en la Provincia de Córdoba, puede ser caracterizada por una autocorrelación espacial positiva, es decir, que los casos observados en cada área *i* tienden a ser similares en ubicaciones espacialmente contiguas. Sin embargo, el análisis de la autocorrelación espacial a partir de los casos observados no tiene en cuenta la población subyacente, por lo que resulta conveniente analizar también el *I* Moran calculado a partir del *CME*, que no resultó significativo en este caso. Además, el *IEB* indicó que el riesgo relativo es espacialmente constante entre los departamentos o los riesgos son heterogéneos sin correlación espacial, para este tipo de cáncer. Para ambos criterios de contigüidad, en relación al cáncer de mama, no resultó estadísticamente significativa la autocorrelación espacial global para la variable correspondiente a los casos observados. El índice de Moran aplicado al *CME*, en cambio, indicó una concentración de valores similares, es decir, una autocorrelación espacial positiva, en la Provincia de Córdoba, de los cocientes de mortalidad estandarizados correspondientes al cáncer de mama. El *IEB* no resultó significativo estadísticamente para el *CME*, de manera que no pudo probarse la existencia de correlación espacial.

Tabla 3.9. Valores asociados al test global de Moran para o_i, según criterio de contigüidad. Provincia de Córdoba. Período 1986-2011.

Criterio de contigüidad	Cáncer de Próstata Hombres				Cáncer de Mama Mujeres			
	I	***E(I)***	***DS(I)***	**Valor *p***	***I***	***E(I)***	***DS(I)***	**Valor *p***
Tipo reina	0,049	-0,040	0,056	0,055	1,235	-0,040	0,049	0,109
Vecinos más cercanos	0,049	-0,040	0,054	0,049	0,017	-0,040	0,048	0,116

Fuente: elaboración propia

Tabla 3.10. Valores asociados al test global de Moran para el *CME*, según criterio de contigüidad. Provincia de Córdoba. Período 1986-2011.

Criterio de contigüidad	Cáncer de Próstata Hombres				Cáncer de Mama Mujeres			
	I	***E(I)***	***DS(I)***	**Valor *p***	***I***	***E(I)***	***DS(I)***	**Valor *p***
Tipo reina	-0,196	-0,040	0,119	0,905	0,062	-0,040	0,121	0,198
Vecinos más cercanos	-0,150	-0,040	0,013	0,829	0,114	-0,040	0,116	0,093

Fuente: elaboración propia

Tabla 3.11. Valores asociados al test *IEB*, según criterio de contigüidad. Provincia de Córdoba. Período 1986-2011

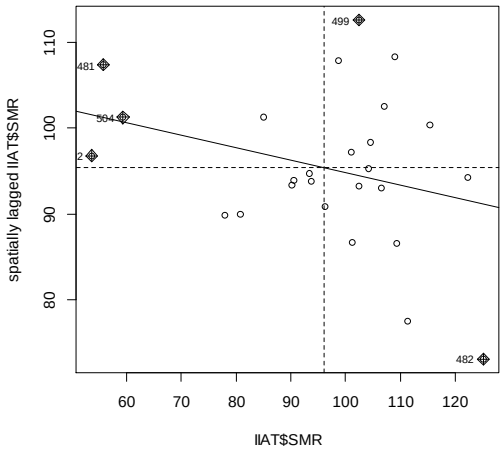
Criterio de contigüidad	Cáncer de Próstata Hombres		Cáncer de Mama Mujeres	
	IEB	**Valor *p***	***IEB***	**Valor *p***
Tipo reina	-0,028	0,548	0,879	0,548
Vecinos más cercanos	-0,038	0,589	2,993	0,251

Fuente: elaboración propia

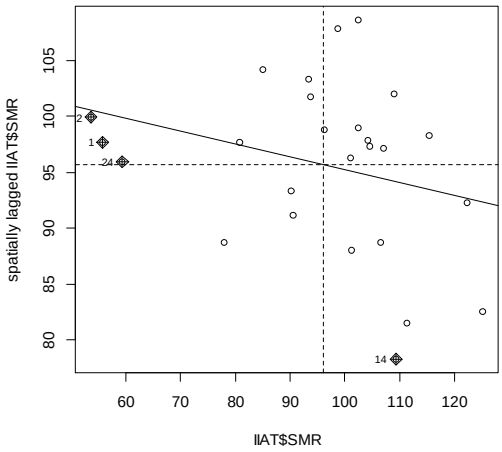
Con el objeto de detectar *clusters* que concentran valores más elevados o bajos de lo que cabría esperar ante una distribución homogénea, se calculó también el estadístico local de Moran. El análisis de los resultados del índice local no arrojó, para ninguno de los tipos de tumores, la presencia de regiones localizadas que concentren valores similares correspondientes a los *CME*. Los gráficos Scatterplot para los cocientes de mortalidad, en ambos tipos de tumores (Figura 3.5) permitieron visualizar, en todos los casos, una nube de puntos dispersa en los cuatro cuadrantes, lo que indica ausencia de correlación espacial. Esta conclusión resulta coherente con los bajos valores obtenidos para el estadístico *I* de Moran. Se detectaron, además, observaciones influyentes que varían según el criterio de

contigüidad empleado, para ambos tipos de cáncer, señalizadas de manera diferencial en los gráficos.

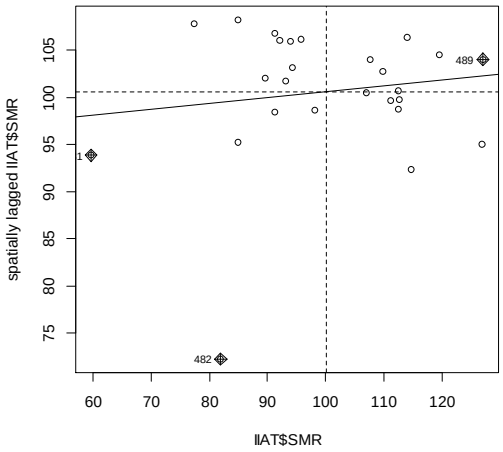
Figura 3.5. Scatterplot para los *CME*. Provincia de Córdoba. Período 1986-2011.



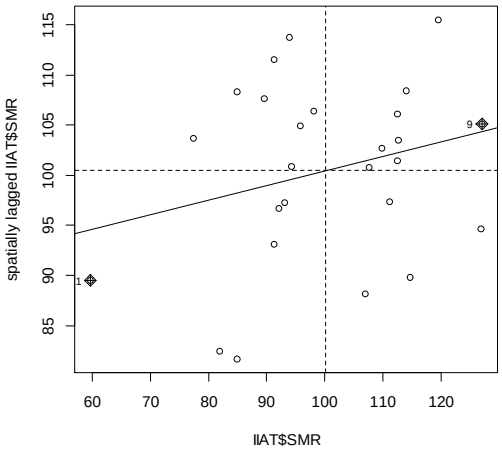
Criterio Tipo Reina
Cáncer de Próstata. Hombres



Criterio Vecinos más cercanos
Cáncer de Próstata. Hombres



Criterio Tipo Reina
Cáncer de Mama. Mujeres



Criterio Vecinos más cercanos
Cáncer de Mama. Mujeres

Fuente: elaboración propia

3.4. Estimación del riesgo relativo en áreas pequeñas (*disease mapping*).

En primer lugar se ajustó el modelo Poisson con intercepto aleatorio (*PIA*), sin covariables, para el período 1986-2011. Este modelo, se combinó con la predicción de los *CMEs* a través del método empírico Bayesiano. Las estimaciones *MV* (Tabla 3.12) arrojaron una varianza del intercepto aleatorio de 0,022 para el cáncer de próstata y 0,009 para el cáncer de mama. En ambos casos, la estimación de la varianza resulta mayor que cero, confirmando la existencia de superdispersión. Dado que el modelo Poisson estándar se encuentra anidado en el modelo Poisson con intercepto aleatorio, se aplicó el Test de cociente de verosimilitud (Tabla 3.13). Como puede observarse, para ambos tipos de tumores, el modelo Poisson debe ser rechazado en favor del modelo Poisson con intercepto aleatorio.

Tabla 3.12. Estimaciones para el modelo Poisson con intercepto aleatorio. Provincia de Córdoba. Período 1986-2011.

	Cáncer de próstata. Hombres			**Cáncer de mama. Mujeres**		
	Estimaciones	**Desvío estándar**	**Valor *p***	**Estimaciones**	**Desvío estándar**	**Valor *p***
β_0 (const)	-0,040	0,035	0,254	0,054	0,026	0,037
ψ	0,022			0,009		

Fuente: elaboración propia

Tabla 3.13. Resultados del Test de cociente de verosimilitud para los modelos Poisson y *PIA.* Provincia de Córdoba. Período 1986-2011.

	Cáncer de Próstata. Hombres	**Cáncer de Mama. Mujeres**
-2 Log-Likelihood	117,35	81,14
p-valor	0,000	0,000

Fuente: elaboración propia

Como se ha explicado, el uso de una distribución de Poisson implica supuestos que no siempre se pueden sostener, fundamentalmente la igualdad entre media y varianza. A los fines de confirmar la presencia de sobredispersión se aplicaron pruebas de puntuación. Para ambos tipos de tumores el valor *p* del test resultó de 0,000, confirmando la presencia de una varianza superior a la media. Una forma de resolver esta sobredispersión es ajustando un Modelo Poisson-Gamma (*P-G*) y estimando los parámetros ν y α que son desconocidos. Estas estimaciones se calcularon por medio de un procedimiento iterativo

(Tabla 3.14). Para el cáncer de próstata, los valores de los estimadores dan una media previa de los riesgos relativos de 0,972 y para el cáncer de mana 1,057, en ambos casos muy cercanas a 1. Estos valores resultan considerablemente superiores a los factores de sobredispersión del modelo *PIA*, que resultaron 0,022 para próstata y 0,001 para mama. El Modelo Poisson-Gamma aparece como el que mejor ajusta a los datos, en función de los indicadores de Bondad de ajuste (Tabla 3.15).

Tabla 3.14. Valores estimados de ν y α. Modelo *P-G.* Provincia de Córdoba. Período 1986-2011.

	Cáncer de Próstata. Hombres	**Cáncer de Mama. Mujeres**
$\hat{v}$	52,826	97,157
$\hat{\alpha}$	54,348	91,882

Fuente: elaboración propia

Tabla 3.15. Indicadores *AIC* y *BIC* para los diferentes modelos. Provincia de Córdoba. Período 1986-2011.

	Cáncer de próstata. Hombres			**Cáncer de mama. Mujeres**		
	Poisson	***PIA***	***P-G***	**Poisson**	***PIA***	***P-G***
AIC	360,77	245,43	244,14	322,25	243,11	243,04
BIC	362,03	247,94	246,66	323,50	245,62	245,55

Fuente: elaboración propia

Habiendo obtenido las estimaciones de los parámetros para cada modelo, se efectuaron las correspondientes estimaciones de los riesgos relativos (Tabla 3.16), que se representaron en los mapas, a fin de visualizar su distribución espacial (Figura 3.6). La distribución geográfica de los riesgos estimados para el cáncer de próstata, resulta muy similar para los modelos Poisson con intercepto aleatorio y Poisson- Gamma, siendo algo superiores en el primer caso. En general, las estimaciones con el modelo Poisson son más altas que las registradas con los otros modelos. Los mayores riesgos se registran en la zona este de la Provincia, involucrando los Departamentos San Justo, Río Segundo, Unión y Pte. Roque Saenz Peña. También presentan riesgos superiores al 100% los Departamentos Capital y San Javier. Este gradiente coincide, en términos generales, con la zona de riesgo detectada al analizar la distribución espacial de los *CMEs* (Figura 3.4.). Para el cáncer de mama, las estimaciones de los riesgos también son muy similares entre los modelos Poisson con intercepto aleatorio y Poisson- Gamma, aunque en este caso resultan mayores las correspondientes al segundo modelo. La zona este, en este caso, también concentra los

mayores riesgos, compartiendo ésta característica con departamentos como Capital, Colón, Sobremonte, Minas y San Javier que, fuera del área, presentan riesgos elevados.

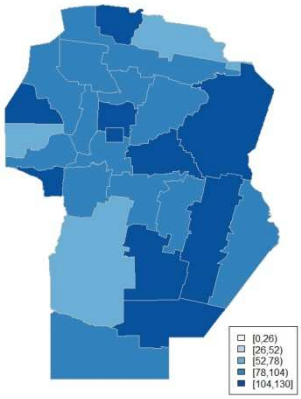
Tabla 3.16. Riesgos relativos ($\hat{\theta}_i$), en porcentaje, bajo diferentes modelos. Provincia de Córdoba. Período 1986-2011.

Departamento	Cáncer de próstata. Hombres.			Cáncer de mama. Mujeres.		
	Poisson	*PIA*	*P-G*	Poisson	*PIA*	*P-G*
Calamuchita	81,10	84,68	85,25	81,49	98,88	81,49
Capital	107,48	106,82	107,11	109,75	126,55	84,93
Colón	104,57	103,22	103,30	98,55	112,61	109,22
Cruz del Eje	101,70	100,04	100,29	109,54	119,00	103,03
General Roca	96,56	96,24	96,28	88,92	104,29	89,90
General San Martin	99,10	98,35	98,29	94,90	109,02	94,16
Ischilín	94,12	94,70	95,28	79,53	99,06	84,96
Juarez Celman	105,00	102,67	103,30	82,78	99,33	85,40
Marcos Juarez	102,82	101,67	103,00	82,90	107,15	92,51
Minas	111,79	100,76	101,31	77,46	103,71	89,08
Pocho	53,87	84,35	85,26	66,79	101,79	87,22
Pte. Roque Saenz Peña	122,81	115,10	114,34	100,43	111,71	96,65
Punilla	93,70	93,56	93,27	103,22	117,50	101,53
Rio Cuarto	59,59	62,40	62,18	97,29	111,87	96,60
Rio Primero	85,33	88,61	89,26	84,79	101,60	87,42
Rio Seco	55,99	82,03	83,23	51,60	95,60	81,04
Rio Segundo	115,99	112,62	112,33	78,89	94,94	81,51
San Alberto	78,17	85,06	86,26	92,41	106,32	91,76
San Javier	109,78	106,00	106,31	78,86	97,16	83,31
San Justo	106,93	105,71	106,01	99,09	113,60	98,12
Santa María	101,42	100,06	100,29	81,20	97,45	83,74
Sobremonte	125,73	103,40	103,08	70,82	103,26	88,62
Tercero Arriba	102,87	101,65	101,30	96,07	110,08	95,09
Totoral	90,50	93,69	94,27	80,39	101,93	87,52
Tulumba	90,88	93,73	94,27	73,31	100,07	85,65
Unión	109,47	107,40	107,62	97,17	111,05	95,95

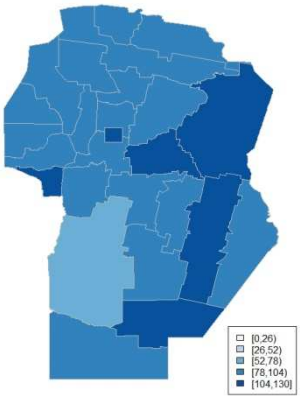
Fuente: elaboración propia

Figura 3.6. Mapeo de los riesgos relativos estimados para diferentes modelos. Provincia de Córdoba. Período 1986-2011.

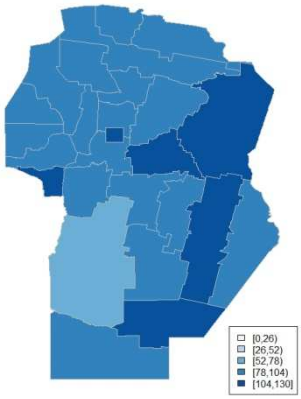
Cáncer de próstata. Hombres.



Modelo Poisson.

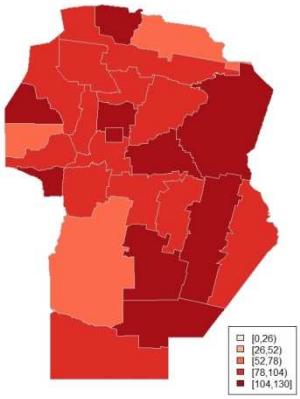


Modelo Poisson con intercepto aleatorio

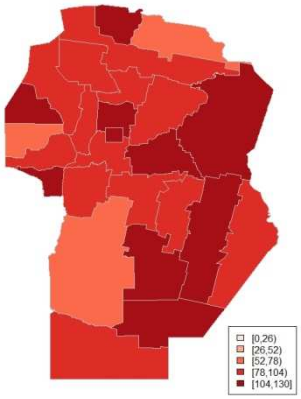


Modelo Poisson-Gamma

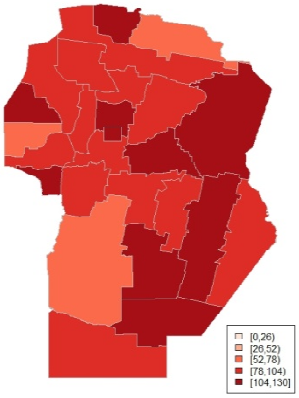
Cáncer de mama. Mujeres



Modelo Poisson



Modelo Poisson con intercepto aleatorio



Modelo Poisson-Gamma

Fuente: elaboración propia

Para complementar el análisis se calcularon los *p-valores* correspondientes a los diferentes modelos (Tabla 3.17), empleados luego en la construcción de los mapas de probabilidad (Figura 3.7). Las zonas con recuentos observados más bajos que lo esperado, en base al tamaño de la población, tienen *p-valores* inferiores y son representadas en el mapa con colores más claros. Las áreas representadas con colores más oscuros en el mapa son

aquellas con recuentos observados mayores que lo esperado y que, por lo tanto, presentan *p*-valores más grandes. Para el cáncer de próstata, no se detectan diferencias importantes entre los *p-valores* correspondientes a los modelos *PIA* y *P-G*. Ambos modelos identifican al departamento Río Cuarto como una zona con menos casos que los esperados. Cuando se analiza el cáncer de mama, se observan mayores diferencias. El modelo Poisson con intercepto aleatorio identifica a Río Cuarto como un departamento de bajo riesgo, mientras que el modelo Poisson-Gamma atribuye los *p-valores* más bajos a los departamentos de Río Segundo y Santa María.

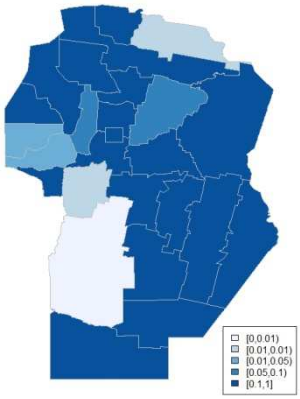
Tabla 3.17. *P-valores* bajo diferentes modelos. Provincia de Córdoba. Período 1986-2011.

Departamento	Cáncer de próstata. Hombres.			Cáncer de mama. Mujeres.		
	Poisson	*PIA*	*P-G*	Poisson	*PIA*	*P-G*
Calamuchita	0,009	0,036	0,041	0,005	0,022	0,021
Capital	0,099	0,999	0,999	0,999	0,999	0,999
Colón	0,814	0,752	0,745	0,375	0,235	0,266
Cruz del Eje	0,601	0,534	0,536	0,914	0,653	0,685
General Roca	0,393	0,388	0,382	0,109	0,130	0,134
General San Martin	0,451	0,122	0,394	0,129	0,083	0,096
Ischilín	0,318	0,344	0,359	0,012	0,061	0,058
Juarez Celman	0,749	0,660	0,679	0,004	0,015	0,015
Marcos Juarez	0,713	0,651	0,679	0,049	0,373	0,045
Minas	0,737	0,272	0,583	0,192	0,389	0,386
Pocho	0,017	0,352	0,363	0,073	0,370	0,363
Pte. Roque Saenz Peña	0,992	0,955	0,945	0,543	0,327	0,353
Punilla	0,090	0,092	0,075	0,809	0,626	0,667
Rio Cuarto	0,000	0,000	0,000	0,200	0,120	0,143
Rio Primero	0,070	0,139	0,151	0,032	0,135	0,069
Rio Seco	0,008	0,304	0,275	0,001	0,232	0,211
Rio Segundo	0,994	0,980	0,976	0,000	0,000	0,000
San Alberto	0,061	0,113	0,133	0,233	0,202	0,211
San Javier	0,877	0,780	0,783	0,002	0,015	0,013
San Justo	0,951	0,923	0,926	0,402	0,259	0,296
Santa María	0,596	0,533	0,535	0,000	0,003	0,002
Sobremonte	0,860	0,621	0,613	0,152	0,419	0,415
Tercero Arriba	0,708	0,644	0,606	0,201	0,126	0,145
Totoral	0,309	0,392	0,402	0,080	0,083	0,210
Tulumba	0,308	0,385	0,394	0,033	0,212	0,203
Unión	0,095	0,916	0,916	0,281	0,175	0,198

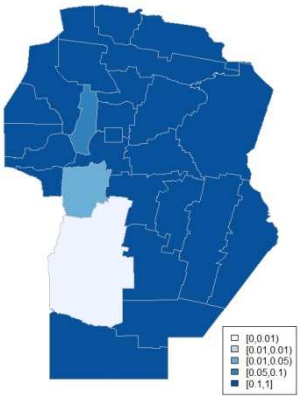
Fuente: elaboración propia

Figura 3.7. Mapas de probabilidad para diferentes modelos. Provincia de Córdoba. Período 1986-2011.

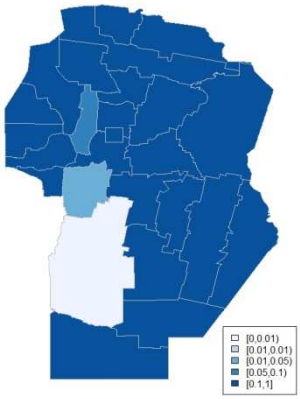
Cáncer de próstata. Hombres.



Modelo Poisson

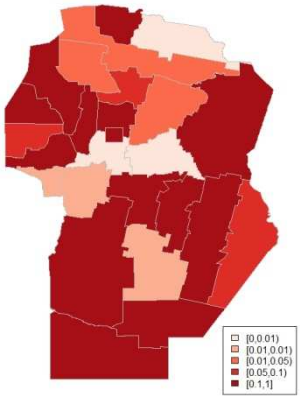


Modelo Poisson con intercepto aleatorio

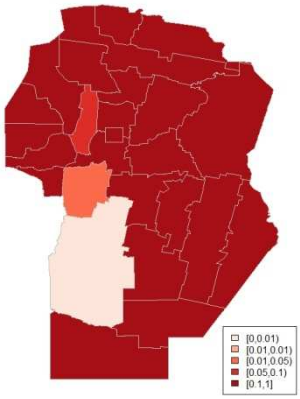


Modelo Poisson-Gamma

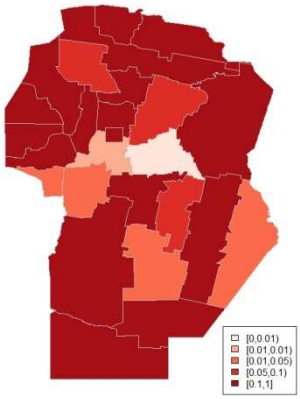
Cáncer de mama. Mujeres.



Modelo Poisson



Modelo Poisson con intercepto aleatorio



Modelo Poisson-Gamma

Fuente: elaboración propia

3.5. Estudios de asociación geográfica (*ecological analysis*)

En el contexto de este tipo de estudio, se ajustó inicialmente un modelo Poisson con intercepto aleatorio, considerando como covariables el porcentaje de la población total sin cobertura de obra social y el porcentaje de la población total desocupada, para el año 2001 (Tabla 3.18). Como en el modelo *PIA* anterior, sin covariables, las estimaciones de las

varianzas asociadas a los interceptos aleatorios evidenciaron la existencia de superdispersión, ya que fueron mayores que cero. Las covariables incluidas en el modelo no resultaron estadísticamente significativas, a un nivel del 5%, para ninguno de los tipos de tumores. El Test de cociente de verosimilitud confirmó la conveniencia del modelo *PIA* en relación al modelo Poisson clásico (Tabla 3.19). Además, se ajustó un Modelo *SAR*, considerando como variable dependiente la transformación z_i de los *CME*s (Tabla 3.20). Resultaron significativas y con coeficiente positivo las variables socioeconómicas, para el cáncer de próstata. Si bien los valores estimados de λ son 0,071 y 0,107, para próstata y mama respectivamente, el *p-valor* asociado es superior en ambos casos al nivel de significación del 0,05, lo que indica que no existe correlación espacial residual luego de ajustar el modelo. Esto puede corroborarse con el test de Moran sobre los residuos del modelo, cuyos *p-valores* fueron 0,154 y 0,203 para ambos tipos de tumor. A partir del indicador *AIC* de Bondad de ajuste (Tabla 3.21), puede concluirse que el modelo *SAR* presenta un mejor ajuste a los datos.

Tabla 3.18. Estimaciones para el modelo Poisson con intercepto aleatorio y covariables. Provincia de Córdoba. Año 2001

	Cáncer de próstata. Hombres			**Cáncer de mama. Mujeres**		
	Estimaciones	**Desvío estándar**	**Valor *p***	**Estimaciones**	**Desvío estándar**	**Valor *p***
β_0 (const)	0,694	0,584	0,905	0,015	0,368	0,968
β_1 (sin cob)	-0,005	0,010	0,635	-0,001	0,047	0,837
β_2 (desocup)	0,007	0,025	0,791	0,012	0,011	0,108
ψ	0,043			0,019		

Fuente: elaboración propia

Tabla 3.19. Resultados del Test de cociente de verosimilitud

	Cáncer de Próstata. Hombres	**Cáncer de Mama. Mujeres**
-2 Log-Likelihood	16,92	79,94
p-valor	0,000	0,000

Fuente: elaboración propia

Tabla 3.20. Estimaciones para el modelo *SAR* con covariables. Provincia de Córdoba. Año 2001.

	Cáncer de próstata. Hombres			Cáncer de mama. Mujeres		
	Estimaciones	**Desvío estándar**	**Valor *p***	**Estimaciones**	**Desvío estándar**	**Valor *p***
β_0 (const)	-1,588	0,457	0,001	-0,704	0,692	0,309
β_1 (sin cob)	0,037	0,006	0,000	-0,007	0,008	0,4074
β_2 (desocup)	0,836	0,016	0,000	0,010	0,027	0,718
λ	-1,588	0,457	0,001	-0,704	0,692	0,309

Fuente: elaboración propia

Tabla 3.21. Indicador *AIC* para los diferentes modelos. Provincia de Córdoba. Año 2001.

	Cáncer de próstata. Hombres		Cáncer de mama. Mujeres	
	PIA	***SAR***	***PIA***	***SAR***
AIC	139,69	16,06	113,23	49,59

Fuente: elaboración propia

3.6. Aglomeraciones de casos (*disease clustering*)

A continuación se presentan los resultados de las pruebas de homogeneidad de los riesgos relativos (Tabla 3.22). En todos los casos se advierte heterogeneidad de los riesgos entre los diferentes departamentos, un nivel de significación del 5%, lo que se considera un punto de partida importante para el análisis sobre la presencia de grupos. Para detectar la existencia global de *clusters* se aplicó el test de Tango, cuyos *p-valores* resultaron 0,158 y 0,261 para el cáncer de próstata y mama respectivamente. A partir de estos resultados no pudo corroborarse, dentro de la Provincia de Córdoba, la existencia de grupos de departamentos con casos observados inusualmente mayores que los esperados, a un nivel de significación de 0,05. No obstante, si se detectaron zonas en riesgo en la provincia, tal como se indicara en los apartados anteriores, hacia el centro-este para el cáncer de próstata y en el noroeste para el cáncer de mama.

Tabla 3.22. *P-valores* asociados a las pruebas de homogeneidad de los riesgos relativos. Provincia de Córdoba. Período 1986-2011.

	p-valor	
	Cáncer de Próstata. Hombres	**Cáncer de Mama. Mujeres**
Chi-cuadrado	0,001	0,001
Homogeneidad	0,001	0,001

Fuente: elaboración propia

CAPÍTULO 4: DISCUSIÓN Y CONCLUSIONES

En este trabajo se identificó que la distribución espacial de la tasa de mortalidad por cáncer de próstata y de mama, en la Provincia de Córdoba, sigue un patrón no aleatorio. Específicamente, el análisis de los mapas correspondientes a los cocientes de mortalidad estandarizados permitió identificar áreas en riesgo, con un número de casos observados superior al esperado, teniendo en cuenta la distribución poblacional.

Respecto al cáncer de próstata, una concentración de valores elevados del *CME* fue detectada hacia el centro-este provincial, donde los departamentos de Presidente Roque Sáenz Peña, Unión y Río Segundo registran los valores más altos. Hacia el oeste de la Provincia, los departamentos Minas, Sobremonte y San Javier también presentaron valores correspondientes al quintil superior de la distribución, siendo los dos primeros departamentos con una baja densidad poblacional. El estudio de la autocorrelación demostró que, si bien la distribución de los casos observados correspondientes al cáncer de próstata en la Provincia de Córdoba, puede ser caracterizada por una autocorrelación espacial positiva, este comportamiento no se observa en relación a los cocientes de mortalidad estandarizados, ajustados por la población de estudio. En efecto, no pudo verificarse que los riesgos relativos fuesen heterogéneos y presenten correlación espacial global significativa. En este sentido, y como se discutirá más adelante, los resultados deben interpretarse con cautela, teniendo en cuenta que se trabaja con un número de unidades muestrales reducido. El índice empírico bayesiano aplicado a los *CME* tampoco resultó significativo estadísticamente, lo que indicaría que el riesgo relativo para éste cáncer es espacialmente constante entre los departamentos o bien, los riesgos son heterogéneos pero no existe correlación espacial significativa. En este sentido, las pruebas de homogeneidad de los riesgos relativos determinaron la existencia de diferencias de significación entre los departamentos de la Provincia en relación a los *CME*. En relación al cáncer de mama, un gradiente en el noroeste de la provincia, que involucra a los departamentos Cruz del Eje,

Punilla, Capital y Cólon, con *CMEs* en el quintil superior fue detectado. Otra región de riesgo establecida por nuestros resultados incluyó a los departamentos de Tercero Arriba, General San Martín, Unión y Marcos Juárez. El hecho de que el índice de Moran aplicado al *CME* fuese positivo y mayor que su correspondiente valor esperado indica una concentración de valores similares, es decir, una autocorrelación espacial positiva de los cocientes de mortalidad estandarizados para éste cáncer.

Si bien en éste trabajo el énfasis estuvo en las cuestiones vinculadas a la modelación, los resultados se visualizan, frecuentemente, en la forma de mapas. La representación y análisis de mapas de mortalidad por enfermedad se han convertido actualmente en una herramienta básica para el análisis de la salud pública regional. No obstante, como afirma Lawson (2001), si bien los mapas de los *CMEs* proporcionan información visual importante sobre la distribución espacial, deben interpretarse en conjunto con la información estadística. El uso de escalas monocromáticas en su construcción reduce la variabilidad, no obstante, existe una cierta arbitrariedad en la elección del número de clases para la paleta de colores, que puede sesgar las interpretaciones. Para reducir la potencial parcialidad en la interpretación, se sugiere complementar el análisis con los mapas de probabilidad, que muestran directamente la variabilidad asociada, tal como se realizó en este trabajo.

En ésta tesis fue utilizado un índice (Moran) ampliamente citado en todas las disciplinas científicas, aunque es importante recordar que las inferencias basadas en éste indicador son asintóticas (distribución normal). Anselin (1995) cuestiona dicha distribución en la medida en que no siempre la aproximación asintótica es válida y, además, porque los momentos de primer y segundo orden utilizados para la estandarización del estadístico son obtenidos bajo una hipótesis nula de que no existe autocorrelación espacial, lo que no siempre se cumple. Efectivamente, la distribución normal funciona razonablemente bien con tamaños muestrales medios ($n > 50$), lo que no se verificaría en este caso ($n=26$ departamentos en la Provincia de Córdoba). En una etapa posterior de esta investigación se propone avanzar en el análisis de los patrones espaciales de la distribución de las tasas de mortalidad por cáncer a nivel de pedanías o a nivel de todo el país, lo que permitiría trabajar con un tamaño de muestra mayor. Asimismo, se proponen explorar otras alternativas, como la técnica no paramétrica *bootstrap* basada en el estadístico de Moran, que asigna aleatoriamente los valores observados a las diferentes áreas y que, como demuestran Lin y Zhang (2007), tiene mayor potencia en comparación con las pruebas asintóticas, especialmente para los casos en que se trabaja con muestras pequeñas y densa contigüidad espacial.

Este trabajo presenta un marco teórico que enfatiza el enfoque de modelación estadística. Si bien es lógico suponer que los conteos de casos se distribuyen Poisson con una esperanza diferente dentro de cada área, generalmente se asume que la esperanza es función de un parámetro de riesgo relativo constante θ_i. De esta manera, $o_i \sim \text{Poisson}(e_i\theta_i)$. En esta definición, el valor esperado de los conteos es una función multiplicativa de los casos esperados e_i, de manera que la log-verosimilitud asociada está dada por $l = \sum_{i=1}^{n} o_i \ln(e_i\theta_i) - \sum_{i=1}^{n} e_i\theta_i$ y por diferenciación es posible obtener el estimador máximo verosímil de θ_i que es simplemente o_i / e_i, el CME_i. Este modelo implica, sin embargo, una serie de supuestos que deben tenerse en cuenta. Fundamentalmente, se asume que todo exceso de riesgo debe ser captado por los e_i, de manera que en su cálculo debe considerarse la información de aquellas variables que, estando disponibles, pueden afectar la distribución del riesgo subyacente. En este sentido, los casos esperados en este trabajo se calcularon aplicando el método de estandarización indirecta que toma las tasas específicas por estrato de edad de la población estándar, las que se promedian utilizando como ponderaciones los tamaños de los estratos de la población estudiada. Esta estandarización se justifica en el hecho de que, en Epidemiología, la mayoría de las tasas (incidencia, prevalencia y mortalidad) son fuertemente dependientes del grupo etario. Sin embargo, podría ser interesante, tal como sugieren Lawson, Brown y Rodeiro (2003) incorporar ajustes por variables del Censo Nacional, particularmente relativas a mediciones socioeconómicas o indicadores de privación.

La falta de información sobre variables socieconómicas en el cálculo de los e_i, sin embargo, pueden suplirse incorporando covariables en el modelo, tal como se realizó en ésta investigación. En efecto, se asume que existen efectos no observados en los o_i, de manera que es necesario incluir efectos aleatorios en el modelo, de manera de describir adecuadamente la distribución del riesgo relativo. En este trabajo, se detectó la presencia de extravariación, para ambos tipos de tumores, aplicando las pruebas de puntuación desarrolladas por Dean (1992). Esta variabilidad extra puede obedecer a diferentes causas, tal como explican Hinde y Demétrio (2005), que incluyen, para datos espaciales, correlación entre las respuestas y omisión de variables no observadas. Ignorar la superdispersión puede tener diversas consecuencias. En primer lugar, los errores estándar obtenidos a partir del modelo pueden ser incorrectos y seriamente subestimados. Además, los cambios en la desviación asociada con los términos del modelo también puede ser demasiado grande, dando lugar a la selección de modelos excesivamente complejos. Finalmente, la interpretación del modelo puede ser incorrecta y las predicciones imprecisas. Habiendo

detectado, como en este caso, que los datos exhiben superdispersión, es necesario aplicar modelos que consideren esta circunstancia.

Para modelar la superdispersión se trabajó, en primer lugar, con un modelo con intercepto aleatorio distribuido normalmente, cuya varianza marginal, dadas las covariables $\mathbf{X_i}$ resulta:

$$Var(y_i|\mathbf{X}_i)=E(y_i|\mathbf{X}_i)+\left[E(y_i|\mathbf{X}_i)\right]^2\{\exp(\psi)-1\}$$

con $\psi>0$, cuando la varianza es mayor que la media marginal. En ésta investigación, la desviación estándar del intercepto aleatorio se estimó en 0,148 para el cáncer de próstata y 0,095 para el cáncer de mama, en ambos casos mayor que cero, confirmando la existencia del fenómeno. Los factores $\exp(\psi)-1$, que multiplican el cuadrado de la esperanza marginal, para obtener el componente aditivo de superdispersión, se estimaron en 0,160 y 0,100 respectivamente.

Sin embargo, este modelo no tiene un expresión cerrada para la verosimilitud y debe ajustarse usando integración numérica. Una aproximación más eficiente computacionalmente puede lograrse especificando una distribución Gamma (con parámetro de forma $1/k$ y parámetro de escala k) para el exponencial del intercepto aleatorio ζ_i , con media 1 y varianza k. Así, la varianza marginal tiene una forma cuadrática similar a la del modelo anterior:

$$Var\ (y_i|\mathbf{X}_i) = E(y_i|\mathbf{X}_i) + [E(y_i|\mathbf{X}_i)]^2 k,$$

donde k reemplaza al factor $\exp(\psi)-1$. Empleando las estimaciones de los parámetros ν y α, $k=\hat{\upsilon}/\hat{\alpha}$. Por lo tanto, queda k= 0,972 para próstata y k= 1,057 para mama, considerablemente superiores al factor de superdispersión del modelo Poisson.

Las estimaciones de los riesgos relativos obtenidas con los modelos Poisson con intercepto aleatorio y Poisson- Gamma se representaron en los mapas, a fin de visualizar su distribución espacial, que resultó muy similar para ambas estrategias de modelación, permitiendo detectar una zona de riesgo al este de la provincia, tanto para cáncer de mama como para el cáncer de próstata.

Las distribuciones a priori y la verosimilitud proveen dos fuentes de información para el mismo problema. La verosimilitud informa sobre el parámetro a través de los datos, en cambio, la distribución a priori informa a través de supuestos. La primera es adecuada cuando el tamaño de muestra es grande, pero cuando, como en este caso, n es pequeño, es más conveniente la segunda. Sin embargo, ninguno de los modelos incorpora explícitamente la estructura de autocorrelación espacial.

En una segunda etapa se desarrollaron dos modelos que permitieron, además de modelar la heterogeneidad no observada entre áreas, incorporar el efecto de covariables socieconómicas, disponibles para el año 2001 y para toda la provincia, a partir del Censo Nacional de Población, Hogares y Viviendas. Luego de analizar la correlación existente entre éstas covariables, se seleccionaron para el ajuste, el porcentaje de la población total sin cobertura de obra social y el porcentaje de la población total desocupada. Estas variables no resultaron estadísticamente significativas en el modelo Poisson, que además, no incorpora en su estructura, de manera explícita, la autocorrelación espacial (Lawson et al., 2003). A partir de ello, se ajustó también un modelo Autorregresivo Simultáneo, considerando como variable dependiente la estandarización de los *CME*s sugerida por Waller y Gotway (2004) para corregir la inestabilidad de las tasas. Para el cáncer de próstata, resultaron significativas y con coeficiente positivo las variables socioeconómicas, no así para el cáncer de mama.

La bondad de ajuste de los modelos se evaluó mediante el Test de cociente de verosimilitud, en el caso de modelos anidados, como el Poisson y el Poisson con intercepto aleatorio, resultando superior el segundo. Para comparar modelos se utilizaron además los indicadores *AIC* y *BIC* concluyendo que, en lo modelos con covariables, el modelo autorregresivo presenta un mejor ajuste a los datos.

Una de las fortalezas del presente estudio radica, justamente, en la implementación de diferentes estrategias de modelación, en auge en la investigación epidemiológica, principalmente para variables respuesta no normales (Skrondal y Rabe-Hesketh, 2003), como en este caso, y se considera superadora, en muchos aspectos, respecto a los clásicos abordajes en este campo.

Finalmente, como tópicos o líneas a profundizar en futuras investigaciones se sugieren:

- Analizar si, a nivel de Argentina, existe un patrón no aleatorio en la distribución espacial de las tasas de mortalidad por cáncer de próstata y mama, similar al detectado para la Provincia de Córdoba.
- Incorporar en la modelación la dimensión temporal, además de la distribución espacial, teniendo en cuenta que se dispone de las tasas de mortalidad para el período 1986-2011.
- Desarrollar el análisis con un nivel de desagregación mayor, trabajando con información de las pedanías de la Provincia, lo que permitiría ajustar un modelo Poisson con dos niveles.
- Indagar sobre otras covariables, fundamentalmente socioeconómicas, que puedan incorporarse al análisis.

- Extender el estudio ajustando diferentes modelos Bayesianos espacio temporales, a fin de efectuar comparaciones con las estimaciones obtenidas.
- Aplicar métodos de evaluación de los riesgos en torno a fuentes de contaminación, con vistas a detectar la presencia de posibles grupos.

CAPÍTULO 5: BIBLIOGRAFÍA

- Aballay L.R., Díaz M.P., Francisca F.M., Muñoz S.E. (2011). *Cancer incidence and pattern of arsenic concentration in drinking water wells in Córdoba, Argentina*. Int J Environ Health Res.
- Ahmad O., Boschi-Pinto C., Lopez A., Murray C. Lozano R., Inoue M. (2001). *Age standardization of rates: a new WHO standard.* GPE Discussion Paper Series Nº 31. World Health Organization.
- Aitkin M., Anderson D., Hinde J. (1981). *Statistical modeling of data on teaching style* (with discussion). Joyleshing sturnal of the Royal Statistical Society 144, 148-161.
- Aitkin M. and Longford N. (1986). *Statistical modeling in school effectiveness studies* (with discussion). Journal of the Royal Statistical Society 149, 1-43.
- Aitkin M. (1996). *A general maximum likelihood analysis of overdispersion in generalized linear models.* Statistics and Computing 6, 251-262.
- Andrew D., Ord J. (1981). *Spatial Processes: Models & Applications*. Pion.
- Anselin, L. (1980). *Estimation Methods for Spatial Autoregressive Structures*. Regional Science Dissertation and Monograph Series, Ithaca, New York.
- Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer.
- Anselin L., Griffith D. (1988). *Do spatial effects really matters in regression analisys?.* Papers of the Regional Science Association vol. 65, pp 11-34.
- Anselin L., Florax R. (1995). *New Directions in Spatial Econometrics*. Berlin: Springer-Verlag.
- Anselin L., Bao S. (1997). *Exploratory Spatial Data Analysis*. En Recent developments in spatial analysis (Eds. Fischer y Getis), Springer-Verlag, Berlín; pp.35-59.
- Anselin L. (2001). *Spatial Effects in Econometric Practice in Environmental and Resource Economics*. American Journal of Agricultural Economics, Agricultural and Applied Economics Association, vol. 83(3), pages 705-710.
- Anselin L. y Moreno R. (2001). *Properties of tests for spatial error components.* ERSA conference papers 01 p.183. European Regional Science Association.
- Armitage P., Berry G. (1987). *Statistical methods and medical research*. Oxford, London: Blackwell Scientific Publications.

- Assunção R., Reis E. (1999). *New proposal to adjust Moran's I for population density.* Departamento de Estadística e CEDEPLAR, UFMG. Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. Statistics in Medicine 18.
- Banerjee S., Wall M., Carlin B.P. (2003). *Frailty modeling for spatially-correlated survival data, with application to infant mortality in Minnesota.* Biostatistics 4.
- Bavaud F. (1998). *Models for spatial weights: A systematic look*. Geographical Analysis, 30:153–171.
- Besag J., York J., Mollié A. (1991). *Bayesian image restoration with two applications in spatial statistics.* Annals of the Institute of Statistical Mathematics 43.
- Besag J., Newell, J. (1991). *The detection of clusters in rare diseases*. Journal of the Royal Statistical Society A, 154:143–155.
- Bhopal R. (2002). *Concepts of Epidemiology: an integrated introduction to the ideas, theories, principles, and methods of epidemiology*. Oxford, NY: Oxford University Press. p. 317.
- Bivand R. S., Müller W., Reder M. (2008). *Power calculations for global and local Moran's I.* Technical report. Department of Applied Statistics. Johannes Kepler University. Linz, Austria.
- Bock R., Aitkin M. (1981). *Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm*. Psychometrika 46, 443-459.
- Bray F., Engholm G., Hakulinen T., Gislum M., Tryggvadóttir L., Storm H.H., Klint A. *Trends in survival of patients diagnosed with cancers of the brain and nervous system, thyroid, eye, bone, and soft tissues in the Nordic countries 1964-2003 followed up until the end of 2006*. Acta Oncol. 2010 Jun. 49(5):673-93.
- Breslow N. and Clayton D. (1993). *Approximate inference in generalized linear mixed models.* Journal of the American Statistical Association 88:9-25.
- Breslow N.E., Day N.E. (1981). *Fundamental Measures of Disease Ocurrence and Association.* Statistical Methods in Cancer Research. Vol. I. The Analysis of Case-Control Studies (IARC Scientific Publications Nº 32). Lyon, Intenational Agency for Research on Cancer. Pp. 42-81.
- Breslow N.E., Day N.E. (1987). *Rates and Rate Standardizartion.* Statistical Methods in Cancer Research. Vol. II. The Design and Analysis of Cohort Studies (IARC Scientific Publications Nº 82). Lyon, Intenational Agency for Research on Cancer. Pp. 48-79.
- Burstein L., Fischer K.H., Miller M.D. (1980). *The multilevel effects of background on science achievement: a cross national comparison*. Sociology of Education; 53:215-25.
- Case, A., Rosen, H. y J. Hines (1993). "Budget Spillovers and Fiscal Policy Interdependence: Evidence from the States," *Journal of Public Economics*, 52, pp. 285-307.

- Catalán-Reyes M.J., Galindo-Villardón M.P. (2003). *Utilización de los modelos multinivel en investigación sanitaria*. Gac. Sanit. 17(Supl. 3):35-52.
- Chiang C.L. (1961). *Standard error of the age-adjusted death rate*. Vital Statistics. Special Reports 47:271-285, USDHEW.
- Choynowski M. (1959). *Map based on probabilities*. Journal of the American Statistical Society, 54:385–388.
- Clayton D., Kaldor J. (1987). *Empirical Bayes estimates of age-standardized relative risks for use in disease mapping.* Biometrics 43: 671-681.
- Clayton D. (1991). *A Monte Carlo method for Bayesian inference in frailty models.* Biometrics 47.
- Cliff A., Haggett P. (1992). *Atlas of disease distributions: analytic approaches to epidemiological data.* Oxford: Blackwell.
- Cliff A. D., J. K. Ord (1981). *Spatial Processes, Models and Applications*. London: Pion.
- Cochran, W. G. (1977). *Sampling Techniques*. 3rd. ed. New York: Wiley.
- Coro C. (2003). *Métodos gráficos del análisis exploratorio de datos espaciales*. Instituto L.R. Klein-Dpto. de Economía Aplicada. Universidad Autónoma de Madrid.
- Cressie N. (1993). *Statistics for spatial data*. John Wiley, New York.
- Dacey M. (1969). *Similarities in the Areal Distributions of Houses in Japan and Puerto Rico*. Área, 3, pp. 35-37.
- Cuzick J., Hills M. (1991). *Clustering and clusters-summary*. In G. Draper (ed.). Geographical Epidemiology of Childhood Leukaemia and Non-Hodgkin Lymphomas in Great Britain 1966-1983. London. HSMO.
- Dean C. (1992). *Testing for overdispersion in Poisson and Binomial regression models*. Journal of the American Statistical Association, 87:451–457.
- Díaz M. P., Osella A., Aballay L., Muñoz S., Lantieri M., Butinof M., Meyer Paz R., Pou S., Eynard A., La Vecchia C. (2009). *Cancer incidence pattern in Cordoba, Argentina*. European Journal of Cancer Prevention. 18:259–266.
- Díaz M. P., Corrente J., Osella, A., Muñoz S., Aballay L. (2010[a]). *Modeling Spatial Distribution of Cancer Incidence in Cordoba, Argentina*. Applied Cancer Research 30(2)245-252.
- Díaz M., García F., Caro P., Díaz M.P. (2010[b]). *Modelos Mixtos Generalizados para el estudio de la asociación entre algunas variables socioeconómicas y las tasas de incidencia de cáncer en localidades de Córdoba, Argentina*. Instituto Interamericano de Estadística. 62, 178, pp. 99-117.

- Diez-Roux A.V., Nieto F., Muntaner C., Tyroler H., Comstock G., Shahar E.,(1997). *Neighborhood environments and coronary heart disease: a multielvel analysis.* Journal Epidemiol. 146:48-63.
- Diez-Roux A. (1998). *Bringing context back into epidemiology: variables and fallacies in multilevel analysis*. Am J Public Health; 88:216-22.
- Dirección de Estadísticas e Información de Salud (2011). *Agrupamiento de causas de mortalidad por división político territorial de residencia, edad y sexo. República Argentina*. Año 2009. Boletín N° 131. Buenos Aires: Ministerio de Salud, Presidencia de la Nación.
- Doll R., Cook P. (1967). *Summarizing indices for comparison of cancer incidence data*. Int J Cancer 2:269-79.
- Doll R. (1980). *The epidemiology of cancer*. Cancer, 45, pp. 2475-2485.
- Doll R., Smith, P. (1982). *Comparison between registries: Age-standardized rates*. In: Waterhouse, J., Muir, C., Shanmugaratnam, K. & Powell, J., eds, Cancer Incidence in Five Continents, Vol. IV (IARC Scientific Publications No. 42), Lyon, International Agency for Research on Cancer, 671-674.
- Donaldson M.S. (2004). *Nutrition and cancer: A review of the evidence for an anti□cancer diet*. Nutrition Journal. 3:19.
- Duncan C., Jones K., Moon G. (1996). *Health-related behaviour in context: a multilevel modelling approach*. Soc Sci Med; 42: 817-30.
- Elliott P., Cuzick J., English D. (1992). *Geographical and Environmental Epidemiology: Methods for Small Area Studies*. New York: Open University Press.
- Elliott P., Martuzzi M., Shadick G. (1995). *Spatial statistical methods in environmental epidemiology: a critique*. Statistical Methods in Medical Research, 4, 137-159.
- Elliot P., Wakefield J., Best N., Briggs D., editors. (2000). *Spatial Epidemiology Methods and Applications*. Oxford University Press.
- Elliott P, Wartenberg D. (2004). *Spatial epidemiology: current approaches and future challenges*. Environmental Health Perspectives; 112(9): 998-1006.
- English D. (1992). *Geographical epidemiology and ecological studies*. En: Elliot P, Cuzick J, English D, Stern R, eds. Geographical and Environmental Epidemiology. Oxford University Press, 3-13.
- Esteve J., Benhamou E., Raymond L. (1994). *Techniques Methods in Cancer Risk.* Statistical Methods in Cancer Research. Vol. IV. Descriptive Epidemiology (IARC Scientific Publications Nº128). Lyon. International Agency for Research on Cancer. Pp 49-105.
- Geary, R.C. (1954). *The contiguity ratio and statistical mapping*. The Incorporated Statistician 5,115-145.
- German Cancer Society (2004). EPIC-NEWSletter. Frankfurt: GCS.

- Getis A., Ord J. (1992). *The Analysis of Spatial Association by Use of Distance Statistics*. Geogmphicd Analysis 24 (July), 189-206.
- Gilks W., Clayton D., Spiegelhalter D., Best N., McNeil A., Sharples L., Kirby, A. (1993). *Modelling complexity: applications of Gibbs sampling in medicine (with discussion).* Journal of the Royal Statistical Society Series B, 55, 39-102.
- Glick, B. J. (1979). *The spatial autocorrelation of cancer mortality.* Social Science and Medicine. 13 D.
- Goldstein H. (1991). *Nonlinear multilevel models with an application to binary response data.* Biometrika 78.
- Goldstein H. (1995). *Multilevel Statistical Models (2nd. edn.).* London. Edward Arnold.
- Goldstein H., Rasbash J. (1996). *Improved approximations for multilevel models with binary response.* Journal of the Royal Statistical Society. 159.
- Greenberg E., Baron J., Karagas M., Stukel T., Nierenberg D., Stevens M., Mandel J., Haile R. (1996). *Mortality associated with low plasma concentration of beta carotene and the effect of oral supplementation.* JAMA.
- Haining R. (1990) *Spatial data analysis in the social and environmental sciences*. Cambridge.
- Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge. University Press.
- Hill A. (1977). *A short textbook of medical statistics*. 9th ed. London: Hodder and Stoughton.
- Hinde J., Demétrio C. (2005). *Overdispersion: Models and Estimation.* A Short Course for SINAPE 1998.
- Keyfitz N. (1996). *Sampling Variance of Standardized Mortality Rates*. Human Biology 38: 309–317.
- Knowelden J., Wilson B., Davies J. (1962). *Cancer incidence of the African population of Kyadondo (Uganda).* The Lancet. Volumen 280.
- Knox E. (1964). *The detection of space-time interactions*. Applied Statistics 13, 25-29.
- Kreft IG. (1998). *An illustration of item homogeneity scaling and multilevel analysis techniques in the evaluation of drug prevention programs*. Eval. Rev.; 22:46-77.
- Ferlay J., Shin HR., Bray F., Forman D., Mathers C., Parkin D. (2008) *Cancer Incidence and Mortality Worldwide: IARC*. Cancer Base No. 10. Lyon, France: International Agency for Research on Cancer. Disponible en la URL http://globocan.iarc.fr.
- Goldstein, H. (1995). *Multilevel Statistical Models*. London: Edward Arnold.
- Grönberg H. (2003). *Prostate cancer epidemiology*. Lancet; 361:859-64.

- Gómez-Rubio, V., López-Quílez, A. (2005). *Using GIS data with R.* Computers and Geosciences, 31:1000–1006.
- Haining R. (1990). *Spatial data analysis in the social and environmental sciences*. Cambridge University. Press Cambridge.
- Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge University Press Cambridge.
- Hall S., Kaufman J., Millikan R., Ricketts T., Herman D., Savitz D.(2005). *Urbanization and breast cancer incidence in North Carolina, 1995–1999.* Ann Epidemiol 15:796–803.
- Herrera M., Paz J., Cid J. (2012). *Introducción a la Econometría Espacial. Una Aplicación al Estudio de la Fecundidad en la Argentina usando R*. Universidad Nacional de Salta.
- Hox J (2002). *Multilevel analysis, techniques and applications.* New York: Lawrence Erlbaum Associates.
- Knowelden J., Oettle AG. (1962). *Cancer incidence of the African population of Kyadondo (Uganda)*. Lancet, ii, 328-330.
- Langford, I. (1995) *A log-linear multi-level model of childhood leukaemia mortality*, Journal of Health and Place, 1.2, 113-120.
- Langford, I., Bentham, G. (1996) *Regional variations in mortality rates in England and Wales: an analysis using multi-level modelling*. Social Science and Medicine, 42.6, 897-908.
- Langford, I., Leyland A., Rasbash J., Goldstein H. (1999). *Multilevel modeling of the geographical distributions of diseases*. Journal of Royal Statistical Society, Series C. Vol. 48, pps. 253-268.
- Larsen K., Petersen J., Budtz-Jorgensen E., Endahl L. (2000). *Interpreting parameters in the logistic regression model with random effects.* Biometrics 56: 909-914.
- Larsen K., Merlo J. (2005). *Appropriate assessment of neighborhood effects on individual health.* Americal Journal of Epidemiology 161: 81-88.
- Lawson A. (2001). *Statistical Methods in Spatial Epidemiology*. John Wiley & Sons, Ltd.
- Lawson A., Browne W., Rodeiro C. (2003). *Disease Mapping with WinBUGS and MLwiN.* Wiley Ed.
- Lee J., Wong D. (2001). *Statistical Analysis with ArcView GIS*. Wiley.
- Leyland A., McLeod A. (2000). *Mortality in England and Wales, 1979-1992: An introduction to multilevel modeling using MLwiN.* Occasional Paper nº1, MRC Social and Public Health Sciences Unit.
- Leyland A. (2001). *Spatial Analysis.* In Multilevel Modelling of Health Statistics, ed. A.H. Leyland and. H. Goldstein, 143-157. Chichester, UK:Wiley.

- Leyland A. y Goldstein H. (2001) *Multilevel modelling of health statistics*. New York: Wiley.
- Lin, G., Zhang, T. (2007). *Loglinear residual tests of Moran's I autocorrelation and their applications to Kentucky breast cancer data*. Geographical Analysis, 3:293–310.
- López-Abente G., Ibáñez C. (2001). *Aplicación de técnicas de análisis espacial a la mortalidad por cáncer en Madrid*. Documentos Técnicos de Salud Pública 66, Madrid: Dirección General de Salud Pública, Consejería de Sanidad, Comunidad de Madrid.
- López Hernández F., Palacios Sánchez M. (2000). *Distintos modelos de dependencia espacial. Análisis de autocorrelación*. Anales de Economía Aplicada. XIV Reunión ASEPELT-España. ISBN: 84-699-2357-9.
- Marshall, R. (1991). *Mapping disease and mortality rates using Empirical Bayes estimators.* Applied Statistics, 40:283–294.
- Moran, P. (1948). *The interpretation of statistical maps*. Journal of the Royal Statistical Society B, v.10, 243-251.
- Moreno R., Vayá E. (2000), *Técnicas econométricas para el tratamiento de datosespaciales: la econometría espacial*. Edicions Universitat de Barcelona, colecció UB 44, manuals.
- Niclis C., Díaz M.P., La Vecchia C. (2010). *Breast cancer mortality trends and patterns in Córdoba, Argentina in the period 1986–2006.* European Journal of Cancer Prevention 2010, 19:94–99.
- Niclis C., Díaz M., Eynard A., Román M., La Vecchia C. (2011[a]). *Dietary Habits and Prostate Cancer Prevention: A Review of Observational Studies by Focusing on South America*. Nutrition and Cancer, 1–11, 2011.
- Niclis C., Pou S., Bengió R., Osella A., Díaz M. (2011[b]). *Tendencias en la mortalidad por cáncer de próstata en Argentina 1986-2006: análisis joinpoint y de edad-período-cohorte.* Cad. Saúde Pública, Rio de Janeiro, 27(1):123-130.
- Ogle MW. (1982) *Proposal for the establishment and international use of a standard population*. Bulletin de l'Institut International de Statistique, tome VI, livre 1:83-5. Rome.
- PAHO; Health Surveillance and Disease Management Area (2007). *Health Statistics and Analysis Unit*. PAHO Regional Mortality Database. Direct adjusted mortality rate using the World Population Prospects 2006 Revision.
- Peto R., Lopez A., Boreham J., Thum M., Heath Jr.C. (1994). *Mortality from smoking in developed countries. 1950-2000*. Oxford. Oxford University Press. Pp A.30.
- Pickle L., Mungiole M., Jones G., White A. (1999). *Exploring spatial patterns of mortality: the new atlas of United States mortality.* Statistics in Medicine. 18. 3211-3220.

- Popkin B. (2004). *The nutrition transition: an overview of world patterns of change*. Nutr Rev 62:S140–S143.
- Potthoff, R., Whittinghill, M. (1966). *Testing for homogeneity: II. The Poisson distribution*. Biometrika, 53:183–190.
- Pou S., Osella A., Eynard A., Niclis C., Diaz M. (2009). *Colorectal cancer mortality trends in Córdoba, Argentina.* Cancer Epidemiology 33 (2009) 406–412.
- Pou S. (2012). *Estudio comparativo de la relación cáncer- dieta en regiones sanitarias de la Provincia de Córdoba empleando la estrategia de modelos multinivel.* Trabajo de Tesis para optar al Título de Doctor en Ciencias de la Salud. Facultad de Ciencias Médicas. Universidad Nacional de Córdoba.
- Rabe-Hesketh S., Toulopoulou T., Murray R. (2001). *Multilevel modeling of congnitive function in schizophrenic patiens and their first degree relatives.* Multivariate Bahavioural Research 36, 279-298.
- Rabe-Hesketh S., Skrondal A., Pickles A. (2002). *Reliable estimation of generalized linear mixed models using adaptative quadrature.* . Stata Journal 2.
- Rabe-Hesketh S., Skrondal A., Pickles A. (2005). *Maximun likelihood estimation of limited and discrete dependent variable models with nested random effects.* . Journal of Econometrics 128.
- Rabe-Hesketh S. and Skrondal A. (2005). *Multilevel and longitudinal modelling using Stata*. College Station, TX: Stata Press.
- Rabe-Hesketh S., Skrondal A. (2012). *Multilevel and Longitudinal Modeling using Stata.* Third Edition. Stata Corp LP. College Station, Texas.
- Rasbash J., Browne W., Goldstein H., Yang M., Plewis I., Healy M., Woodhouse G., Draper D., Langford T., Lewis T. (2000). *A User's Guide to MLwiN.* London. Institute of Education.
- Raudenbush S., Yang M., Yosef M. (2000). *Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation.* Journal of Computational and Graphical Statistics 9, 141-157.
- Riboli E., Lambert R. (2002). *Nutrition and lifestyle: opportunities for cáncer prevention.* Intenational Agency for Research on Cancer. Lyon. France.
- Richardson S. (2003). *Spatial models in epidemiological applications.* In P. Green, N. Hjort, y S. Richardson (Eds.). Hoghly Structured Stochatic Sustems. London. Oxford University Press.
- Rothman K., Greenland S., Lash T. (2008). *Modern Epidemiology.* Thrird edition. Lippincott Williams & Wilkins. USA.
- Ryan T. (1997) *Modern reggression methods*. Wiley.

- Sáez M., Saurina Canals C. (2007). *Estadística y Epidemiología espacial*. Grup de Recerca en Estadística, Economia Aplicada i Salut (GRECS). Universitat de Girona.
- Sánchez-Cantalejo E., Ocaña-Riola R. (1999). *Los modelos multinivel o la importancia de la jerarquía*. 13: 391-8.
- Schöen R. (1970). *The geometric mean of the age-specific death rate as a summary index of mortality.* 7:317-24.
- Seeman T., Crimmins E. (2001). *Social Environment Effects on Health and Aging Integrating Epidemiologic and Demographic Approaches and Perspectives*. School of Medicine, University of California, Los Angeles, California, USA.
- Segi M. (1960). *Cancer mortality for selected sites in 24 countries (1950-57).* Department of Public Health, Tohoku University of Medicine, Sendai, Japan.
- Sokal R., Oden N., Thomson B. (1998[b]). *Local spatial autocorrelation in a biological model*. Geogr. Anal 30:331-354.
- Stryhn H., Sanchez J., Morley P., Booker C., Dohoo R. (2006). *Interpretation of variance parameters in multinivel Poisson regression models.* In Proceedings of the 11th symposium of the International Society for Veterinary Epidemiology and Economics, 702-704. Cairns, Australia.
- Tango, T. (1995). *A class of tests for detecting general and focused clustering of rare diseases*. Statistics in Medicine, 14:2323–2334.
- Tiefelsdorf M., Griffith D. (2007). *Semiparametric Filtering of Spatial Autocorrelation: The Eigenvector Approach*. Environment and Planning A, 39(5), pp. 1193-1221.
- Thomas R. (1990). *Spatial epidemiology*. London: Pion.
- Tumas N., Niclis C., Osella A., Díaz M., Carbonetti A. (2015). *Tendencias de mortalidad por cáncer de mama en Córdoba, Argentina, 1986–2011: algunas interpretaciones sociohistóricas*. Rev. Panam. Salud Pública. 37(4/5):330–6.
- Wakefield, J., Kelsall, J., Morris, S. (2000). *Clustering, cluster detection and spatial variation in risk*. In Elliott P., Wakefield J., Best N. and Briggs, D., editors. Spatial Epidemiology: Methods and Applications. Oxford University Press, Oxford, pp 128–152.
- Wall, M. (2004). *A close look at the spatial structure implied by the CAR and SAR models.* Journal of Statistical Planning and Inference, 121:311–324.
- Waller L., Carlin B., Xia H., Gelfand, A. (1997). *Hierarchical spatio-temporal mapping of disease rates*. Journal of the American Statistical Association, 92, 607-617.
- Waller L., Gotway, C. (2004). *Applied Spatial Statistics for Public Health Data*. Wiley. Hoboken, NJ.
- Walter S., Birmie S. (1991). *Mapping mortality and morbidity patterns: an international comparison*. International Journal of Epidemiology, 20:678-689.

- Walter S. (2000). *Disease mapping: a historical perspective*. En: Elliott P, Wakefield J, Best N, Briggs DJ (eds). Spatial epidemiology: methods and applications. Oxford: Oxford University Press, 223-252.
- Wolfenden H. (1923). *On the methods the mortalities of two o more communities and the standardization of death rates.* Journal of the Royal Statisitical Society. 86:399-411.
- Yerushalmy J. (1951). *A mortality index for use in place of the age adjusted death rate.* Public Health. 41:907-22.