



Análisis espacial y de supervivencia del cáncer de cuello uterino para el municipio de Santiago de Cali en el período 2006-2014.

Duver Arley Tulande Orozco

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2018

Análisis espacial y de supervivencia del cáncer de cuello uterino para el municipio de Santiago de Cali en el período 2006-2014.

Duver Arley Tulande Orozco

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadístico

Director(a):
Ph.D. Mercedes Andrade Bejarano
Codirector(a):
Ph.D. Luis Eduardo Bravo

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2018

“El futuro recompensa a los que siguen adelante. No tengo tiempo para sentir pena por mí mismo. No tengo tiempo para quejarme. Voy a seguir adelante”

Barack Obama.

“Todos nuestros sueños se pueden volver realidad si tenemos el coraje de perseguirlos”

Walt Disney.

“La educación es el arma más poderosa que puedes usar para cambiar el mundo”

Nelson Mandela.

“Un hombre que nunca ha ido a la escuela podría robar un vagón de carga, pero si tuviese una educación universitaria, podría robar el tren entero”

Theodore Roosevelt.

Agradecimientos

A mi madre Ana Milena Orozco, por ser la mujer de mi vida, mi mujer guerrera, mi mayor ejemplo de vida, que mediante tus sacrificios y superaciones me enseñaste que si vale la pena luchar por la felicidad de tu hogar, de mi hogar; por ayudarme a salir adelante cuando nadie mas lo hizo, cuando ya nadie creía en mi, cuando me desvié del camino. Gracias por tus consejos, porque de entre todos los que tenia presente, los tuyos siempre fueron los mejores y certeros. Tu gran humildad, fue lo que me inspiro a realizar este trabajo de grado, sin tu apoyo y amor incondicional, esto no fuera sido posible.

A mi abuela Luz Morales Grueso, por ser mi segunda madre, mi reina, la que me crió, tu fuiste la que me enseñaste a darlo todo por una oportunidad y a ver una oportunidad donde no la hay, gracias por prepararme para enfrentar la vida y terminar con lo que me proponga, el hombre que soy, es lo que tu formaste y lo que siempre quise ser. Estoy muy feliz de haberte conocido y a tu lado haber crecido.

A mi padre Sindulfo, por apoyarme para poder culminar con esta etapa tan importante en mi vida y ayudarme a superar. Te quiero padre y Gracias por llevarme presente en tus oraciones.

A mi directora de trabajo de grado Mercedes Andrade Bejarano por su completa asesoría, gracias por transmitirme esta linda metodología y enseñarme la importancia de poder comprender el proceso que hay dentro de esas llamadas cajas negras.

Al doctor Luis Eduardo Bravo, por haberme brindado la oportunidad de poder realizar esta investigación, gracias por todos los conocimientos y consejos tan valiosos. Ademas agradecerle a todo el equipo del Registro Poblacional de Cáncer de Cali, por el buen trato y atención brindada.

Al profesor Roberto Behar Gutiérrez, por el amor y esfuerzo en su forma de educar, sus clases cambiaron mi vida, gracias por transmitirme ese amor hacia la Estadística, usted es la esencia del programa.

A Derly Ruiz Alvarado, mi negra bella, que fue pieza fundamental para la finalización de este trabajo de grado, este amor me permitió reivindicarme y plantearme nuevas metas en mi vida, gracias por preocuparte por mí en cada momento y desear lo mejor para mi porvenir.

Duver Arley Tulande Orozco

Resumen

En Colombia el cáncer de cuello uterino (CaCu) es la segunda neoplasia más frecuente en mujeres y la primera causa de muerte por cáncer en mujeres. Objetivo: determinar el comportamiento espacial del cáncer de cuello uterino y estimar la supervivencia relativa (SR) a 5 años en Cali durante el periodo 2006-2014. Se incluyeron los casos nuevos de mujeres con CaCu residente en Cali, Colombia, registradas en la base de datos del Registro Poblacional de Cáncer de Cali durante el periodo 2006 - 2014. En los resultados, el 90,4 % de las mujeres pertenecían a los estratos 1, 2 y 3; la edad media al diagnóstico fue 53 años en el 2006 y 50 años en el 2014. Las pruebas de completa aleatoriedad espacial evidenciaron un patrón agregado para todos los años del estudio con altos niveles de intensidad en las comunas 3, 6, 13, 15 y 21. La SR a 5 años para las mujeres con CaCu fue de 60,1 % y se asoció de manera inversa con la edad. La SR a 5 años de las pacientes ubicadas dentro de los sitios de alto nivel de intensidad, fue menor en todos los grupos de edad, comparado con las pacientes que se encontraban por fuera. En conclusión, Se observó que los altos niveles de intensidad asociados a las comunas 6, 13, 15 y 21 (población vulnerable); se relacionan de manera directa con la edad y con la presencia de Asentamientos Humanos de Desarrollo Incompleto dentro de estas comunas.

Palabras clave: Cáncer de cuello uterino, análisis espacial, supervivencia.

Abstract

In Colombia, cervical cancer (CaCu) is the second most frequent neoplasm in women and the leading cause of cancer death in women. Objective: to determine the behavior of uterine cancer and estimate the relative survival (SR) at 5 years in Cali during the period 2006-2014. We included new cases of women with CaCu resident in Cali, Colombia, registered in the database of the Cali Population Registry of Cali during the period 2006 - 2014. In the results, 90.4 % of the women Belonging to strata 1, 2 and 3; the mean age at diagnosis was 53 years in 2006 and 50 years in 2014. The evidence of complete randomization shows an aggregate pattern for all years of the study with high levels of intensity in the communes 3, 6, 13, 15 and 21. The SR at 5 years for women with CaCu was 60.1 % and was inversely associated with age. The SR to 5-year-old patients was located within the sites of high intensity level, was lower in all age groups, compared with patients who were on the outside. In conclusion, populations 6, 13, 15 and 21 (vulnerable population); It is a direct relationship with age and with the presence of Human Settlements of Incomplete Development within these communities.

Keywords: Cervical cancer, spatial analysis, survival.

Contenido

Resumen	vii
1. Introducción	2
1.1. Planteamiento del Problema	2
1.2. Justificación	3
1.3. Objetivos	5
1.3.1. Objetivo General.	5
1.3.2. Objetivos Específicos.	5
1.4. Antecedentes	6
2. Marco Teórico	9
2.1. Marco Conceptual.	9
2.2. Marco Teórico Estadístico.	13
2.2.1. Estadística Espacial.	13
2.2.2. Función de supervivencia con enfoque de período.	24
2.2.3. Modelo Lineal generalizado.	24
2.2.4. GLM con estructura de error Poisson.	26
3. Metodología.	31
3.1. Base de datos.	31
3.2. Variables del estudio.	32
3.3. Zona de estudio.	33
3.4. Depuración de la información.	33
3.5. Geocodificación.	34
3.6. Análisis estadístico.	35
3.6.1. Análisis descriptivo.	35
3.6.2. Análisis espacial.	35
3.6.3. Modelación	42
3.6.4. Análisis de Supervivencia.	46
4. Resultados.	48
4.1. Análisis descriptivo.	48
4.1.1. Estrato socioeconómico.	48
4.1.2. Edad.	50

4.2. Análisis Espacial.	52
4.2.1. Distribución espacial	52
4.2.2. Pruebas de completa aleatoriedad.	53
4.2.3. Estimación de la intensidad	54
4.3. Análisis de Supervivencia	56
5. Conclusiones y recomendaciones	61
5.1. Conclusiones	61
5.2. Limitaciones del estudio	64
5.3. Recomendaciones	64
A. Anexo: Pruebas de Completa aleatoriedad	65
B. Anexo: Tablas de variables para modelo lineal generalizado.	69
C. Anexo: Curvas de supervivencia relativa de 5 años	73
D. Anexo: Comparación de mujeres ubicadas dentro y fuera de los puntos con altos niveles de intensidad.	75
Bibliografía	79

1. Introducción

1.1. Planteamiento del Problema

El cáncer es por definición : “Un conjunto de enfermedades en las cuáles las células del cuerpo crecen sin control”, los signos y síntomas de esta patología son variables, se presentan a cualquier edad y dependen del tamaño, localización y órganos o tejidos afectados (ACS 2014).

Según la Liga Contra el Cáncer (2017), en Colombia el cáncer de cuello uterino es la segunda neoplasia más frecuente en mujeres y la primera causa de muerte por cáncer entre mujeres de 30 a 59 años. Cada año se diagnostican aproximadamente 4.661 casos nuevos y mueren 1.986 mujeres por este tipo de cáncer, es decir una de cada 2 mujeres diagnosticadas con cáncer de cuello uterino muere a causa de esta enfermedad y una gran parte es debido a la detección tardía del mismo.

El cáncer cervicouterino es prevenible mediante las citologías y la vacunación de las niñas de entre los 9 y 15 años edad contra el virus del papiloma humano, que es una vacuna “profiláctica”, no terapéutica, es decir que evita la infección, más no la cura (Block et al. 2006); y además es tratable mediante el tamizaje y el tratamiento de las lesiones precancerosas en las mujeres, ya que estas lesiones preceden al cáncer de entre 10 a 20 años después. El cáncer cervicouterino es una de las formas más prevenibles y tratables de cáncer si se detecta precozmente y se atiende eficazmente (OMS 2016).

Aunque desde el 2013 hasta la actualidad se disponga de forma gratuita la vacuna contra el VPH (Ley 1626 del 30 de Abril de 2013, Congreso de Colombia), además de tener avances para la tamización, el diagnóstico y el tratamiento de la enfermedad para mejorar la calidad de vida ó mejor aún evitar la muerte del pacientes con cancer de cuello uterino, se siguen reportando un gran número de casos a nivel mundial para las mujeres de países en vía de desarrollo, como lo es actualmente Colombia (Bernard 2014).

Cuando una persona tiene una enfermedad en específico, no necesariamente muere por la misma, a veces pueden ocurrir otros sucesos fatales alternos a la enfermedad inicial (Esteve et al. 1994), además la probabilidad de sobrevivir a la enfermedad se ve afectada por el lugar donde se encuentre (ciudad, país, etc) y la edad en que se le fue diagnosticada la enfermedad, por lo tanto, se realizará un análisis de supervivencia para determinar

la probabilidad de sobrevivir al cáncer de cuello uterino con 5 años de seguimiento y se realizarán comparaciones dependiendo de la ubicación de la paciente con cáncer de cuello uterino en el perímetro de la ciudad de Santiago de Cali.

El Registro Poblacional de Cáncer de Cali (RPCC) cuenta con información de pacientes con cáncer de cuello uterino. Desde el año 1962, se tienen registros de la dirección o el domicilio de la mujer, pero en la actualidad no existen investigaciones donde se determine la distribución espacial e intensidad de los casos de este tipo de cáncer, en particular. Por esta razón en esta investigación se pretende además de analizar la supervivencia de las pacientes, identificar gráficamente la ubicación espacial de cada caso y determinar cuál es el comportamiento espacial de los casos de cáncer de cuello uterino de la ciudad de Santiago de Cali, con el fin de suministrar información a las entidades de salud que están relacionadas con este problema de salud pública y de control. Teniendo en cuenta lo expuesto anteriormente, surgen 2 preguntas de investigación: ¿Como es la distribución espacial de la ocurrencia de los casos de cáncer de cuello uterino del municipio de Santiago de Cali en el período 2006-2014? y ¿La edad y el año de seguimiento explican el comportamiento de la supervivencia relativa a 5 años del cáncer de cuello uterino en el municipio de Santiago de Cali en el período 2006-2014?

1.2. Justificación

El reporte global de la agencia internacional para la investigación en cáncer (IARC), Indica que por su incidencia a nivel mundial el cáncer de cuello uterino (CCU) ocupa el séptimo lugar entre todos los tipos de cáncer y es el cuarto cáncer más común en las mujeres. Se reportaron en promedio 528,000 nuevos casos en 2012, de las cuales 266,000 murieron a causa de este cáncer, alrededor del 85 % de estas ocurrencias son mujeres del así llamado “Tercer Mundo”, es decir, países en vía de desarrollo como lo es Colombia (OMS 2016).

El Instituto Nacional de Cancerología de Colombia (INC) estimó que en el período 2007-2011, se registraron aproximadamente 4,462 nuevos casos de cáncer de cuello uterino y se reportaron 1,861 muertes a causa de esta enfermedad (Ramos & Duarte 2015).

De las cifras publicadas por el Registro Poblacional de Cáncer de Cali (RPCC 2017), se observó que en el período 2008-2012 se reportaron 1019 casos de cáncer de cuello uterino y 478 mujeres murieron a causa de este cáncer en el período 2009-2013.

La tasa ajustada por edad (TAE) de mortalidad del CCU para Colombia fue de 7,6 por cada 100.000 mujeres, promedio anual. Para la mayoría de los departamentos de Colombia se observó una tendencia de descenso en la mortalidad del CCU en el período 1994-2013, específicamente una disminución del 2,2 %, promedio anual. Además la investigación reportó

que los sitios con altos riesgos de mortalidad se presentaron particularmente, en las zonas cercanas a los ríos navegables, las áreas de frontera, el golfo de Urabá y la serranía del Baudó (Pardo et al. 2017).

La metodología de la de la estadística espacial es apropiada para solucionar los objetivos de esta investigación porque permitirá identificar como es el comportamiento en conjunto de los casos de cáncer de cuello uterino a lo largo del perímetro del municipio de Santiago de Cali en el período 2006-2014 y también se obtendrán mapas de intensidad indicando los lugares donde hay más frecuencias de los casos de cáncer de cuello uterino.

La metodología del análisis de supervivencia relativa permite modelar los casos en donde en la información obtenida en los DCO (certificados de defunción), no se encuentra disponible la causa de muerte o no es confiable la información, a causa de errores de clasificación (Colditz 2015); además con esta metodología se puede cuantificar la probabilidad de sobrevivir al cáncer de cuello uterino, después de 5 años de haber diagnosticado la enfermedad.

1.3. Objetivos

1.3.1. Objetivo General.

Determinar la distribución espacial de la ocurrencia del cáncer de cuello uterino en la ciudad de Santiago de Cali y modelar la supervivencia relativa en el período 2006-2014.

1.3.2. Objetivos Específicos.

- Obtener los mapas de la distribución espacial de los casos de cáncer de cuello uterino en el municipio de Santiago de Cali, para el período 2006-2014 y analizar a nivel descriptivo las variables asociadas al fenómeno.
- Determinar la intensidad y el tipo de patrón espacial de los casos de cáncer de cuello uterino en el municipio de Santiago de Cali, para el período 2006-2014 y obtener mapas.
- Modelar la supervivencia relativa a 5 años con un enfoque de período , de pacientes con cáncer de cuello uterino en el municipio de Santiago de Cali, para el período 2006-2017.

1.4. Antecedentes

Alvarez et al. (2009)

Este estudio se realizó en Cuba y utilizaron datos del Registro Nacional de Cáncer (NCR) en el período 1999-2003, estos datos referenciaban que en promedio se reportaban anualmente 2.438 y 1451 casos nuevos de cáncer de mama y cuello uterino, respectivamente; este país está dividido en 169 municipios y en el 2001 contaba en promedio con una población de 33.200 mujeres por municipio. Una hipótesis que se planteó basados en análisis descriptivos anteriores en base al NCR indicó que los riesgos de cáncer de mama eran estadísticamente diferentes, a un nivel de significancia de $\alpha = 0,05$, entre la parte occidental y oriental del país. Para cada municipio, se registraron las cifras anuales de nuevos casos de cáncer de mama y de cuello uterino y también se registraron las cifras anuales de los recuentos de la población en grupos de edad de 5 años. Se halló la razón de incidencia estandarizada (SIR) para cada municipio y como se presentaban inconvenientes para el uso de los SIR en la presentación de mapas, se utilizó un enfoque bayesiano para poder proporcionar estimaciones suavizadas del riesgo relativo (RR).

Se utilizaron los modelos basados en los recuentos observados, los cuáles eran los modelos Poisson-Gamma, Poisson-Lognormal y autoregresivo condicional (CAR); Los 2 primeros modelos fueron independientes espacialmente y el CAR estaba espacialmente estructurado, este último se utilizó porque conocimientos anteriores probaron que las áreas geográficamente cercanas tienden a tener RR similares, entonces con este modelo los SIR se nivelan localmente hacia el riesgo promedio en el conjunto de áreas vecinas. Para el modelo CAR y Lognormal se incluyó la variable “nivel de urbanización” para cada municipio, que es el porcentaje de población que vive en cada uno. Como se querían comparar los modelos, se utilizó el Criterio de información de la desviación (DIC), que permite la evaluación simultánea de la bondad de ajuste y la complejidad del modelo (Spiegelhalter et al. 2002). Por último se calcularon las distribuciones posteriores para todos los parámetros en los modelos, mediante el muestreo de Gibbs. En los resultados se observó que el modelo CAR proporcionó el mejor ajuste para la distribución geográfica del cáncer, puesto que tenía el DIC más bajo y la inclusión de la covariable nivel de urbanización, mejoró el ajuste en los 2 modelos donde se incluyó; también se resalta que en los SIR crudos, no se observaba ningún patrón claro en la distribución del riesgo, en cambio los mapas de las probabilidades posteriores de los riesgos que excedían la unidad, proporcionaron áreas homogéneas bien delimitadas, en las que se pudieron distinguir diferentes formas del riesgo geográfico.

Manotas et al. (2014)

En el año 2006, se realizó este estudio para establecer una caracterización del cáncer de cuello uterino, se le aplicó un formulario a una muestra de 105 pacientes pertenecientes al departamento del Atlántico, Colombia. En esta investigación se desarrolló un modelo de regresión logística ordinal; se definió como variable dependiente el cáncer de cuello uterino (leve, moderado, agudo) y como variables independientes la edad, la cantidad de hijos nacidos vivos, la cantidad de hijos nacidos muertos, el tipo de parto, el tipo de embarazo y por último si portaba alguna enfermedad de transmisión sexual; luego se ajustó un modelo y seguidamente se calcularon los riesgos, riesgos relativos y razones ODDS. Al modelo se le aplicó una prueba de bondad de ajuste para observar la significancia de cada variable. Los resultados mostraron que había algunos predictores con significancia mayor a 0.05, por lo tanto se ajustó otro modelo excluyendo algunas variables, hasta que en el modelo ajustado todas sus variables fueran significativas. El modelo que mejor se ajustó era el que tenía como variables independientes la cantidad de hijos nacidos vivos, la cantidad de hijos nacidos muertos y la variable categórica que indicaba si el parto fue por cesárea o espontáneo (natural). Por último se concluyó que el cáncer de cuello uterino leve tiene un mayor riesgo para las mujeres en el que su único hijo nació muerto y cuando el parto fue espontáneo; en el cáncer de cuello uterino moderado y agudo se obtuvieron riesgos elevados para las mujeres que tuvieron partos con cesárea y mayor cantidad de hijos vivos; se concluyó de manera general que el riesgo de tener cáncer de cuello uterino es mayor cuando, aumenta el número de hijos en partos por cesárea y se ha perdido un hijo.

García-Soto et al. (2016)

Este estudio se realizó con el fin de describir los atributos sociodemográficos y de salud de las pacientes con cáncer de cuello uterino, atendidas en el Hospital Universitario San Vicente de Paúl (HUSVP) de Medellín, Colombia, entre los años 1997 y 2002. Las variables que se consideraron en este estudio fueron la edad, la raza, el estado civil, la escolaridad (ninguna, primaria, secundaria o universitaria), la procedencia, la edad de la menarquia, la edad de la menopausia y el estrato socioeconómico (0, 1, 2, 3 ó 4). La base de datos inicial constaba de 843 casos de pacientes con cáncer de cuello uterino. Después de la depuración de los datos se contó con 735 registros de pacientes con diagnóstico de cáncer de cérvix. Los resultados obtenidos mostraron que en el 64,8% de los registros no se encontró información con respecto al uso de métodos anticonceptivos y solo en el 15,1% de los casos, la información disponible, reportó el uso de condones. El 39,1% de las 435 pacientes (información disponible) eran fumadoras de cigarrillo, el 18,5% de las pacientes aceptaron que nunca se habían echo una citología y el 47,5% solo una citología. Por último se concluyó, que la gran presencia de mujeres menores de 30 años en el estudio, se debía al inicio temprano de la actividad sexual, además se encontró un promedio alto de embarazos

(factor de riesgos), y entre las mujeres mayores de edad, pobres y con bajo nivel educativo, se encontró que nunca se habían realizado una citología y que tenían un estado avanzado de la enfermedad.

Atalah et al. (2001)

Esta investigación tuvo un diseño de casos y controles de modalidad pareada; se realizó en la ciudad de Santiago de Chile, donde se seleccionaron 8 hospitales o centros oncológicos de la ciudad, el objetivo de este estudio se basó en analizar la asociación entre el cáncer de cuello uterino y el consumo de alimentos ricos en antioxidantes naturales, el tabaquismo y conductas reproductivas. Primero se definió una variable de control que fue la edad (esta fue agrupada para cada 5 años). Para observar el consumo de alimentos definidos se seleccionaron 58 alimentos con alto contenido de antioxidantes naturales (carotenos, vitamina C, y E), se incluyeron variables pertenecientes a la características físicas de la paciente como lo son la talla, el peso habitual y el índice de masa corporal (kg/m^2). Se analizó la historia reproductiva (paridad, edad de menarquia, edad al primer parto) y las conductas sexuales (edad a la primera relación, número de parejas) y por último se registraron antecedentes de cáncer en familiares directos (padre, madre, hermanos, tíos, abuelos), consumo de alcohol y hábito de fumar. En el análisis estadístico se utilizaron modelos logísticos multivariados a causa del procedimiento de adición de variables hacia adelante (forward step analysis), considerando como criterio de inclusión una probabilidad de 0,05 y de exclusión 0,2, asignándole un valor de uno (1) a la presencia del factor de riesgo (variable) y cero (0) a su ausencia. La muestra estaba conformada por 170 casos y 340 controles; se utilizó la prueba de Kruskal Wallis para observar diferencias significativas con un $\text{valor } p < 0,05$. En los resultados se observó que las mujeres con cáncer de cuello uterino tuvieron un mayor número de embarazos y el primer hijo a una edad más temprana, por otra parte, un alto riesgo de tener cáncer de cuello uterino estaba asociado al aumento de embarazos, una edad temprana en el primer parto, el hábito de fumar (actual o pasado) y prácticamente todas las variables alimentarias se asociaron también con el riesgo de esta patología (menor riesgo a mayor consumo). Por último se propuso un modelo, en el cual se referenciaba que un embarazo precoz, así como el hábito de fumar aumentan cerca de tres veces el riesgo de tener cáncer cervicouterino, e inversamente un mayor consumo de verduras y de vitamina E (a través de los alimentos) reducía el riesgo en un 40 %.

2. Marco Teórico

En este Capítulo se presenta el marco teórico de la investigación, la primera sección contiene el marco conceptual, donde se presentarán las definiciones de algunas terminologías referentes a la incidencia del cáncer de cuello uterino, el tipo de estudio a desarrollar y los sistemas de información geográfica; en la segunda sección se presenta el marco teórico estadístico donde se define la metodología de estadística espacial, características de sus áreas. Se profundizará en el área de patrones puntuales, definiendo la función de intensidad λ y algunas pruebas para determinar el tipo de patrón espacial, y en el final de esta sección se presentara en detalle que es el modelo lineal generalizado.

2.1. Marco Conceptual.

Estudios ecológicos

Estos estudios no utilizan la información del individuo de una forma aislada, sino que utilizan datos agregados de toda la población (Beaglehole et al. 1994). Estas unidades se obtienen por su ubicación espacial (países, regiones, comunas, colegios, etc) o por ubicación temporal (años, meses, semanas, días, etc) y su objetivo básico es establecer comparaciones entre las unidades de análisis respecto a ciertas condiciones de salud o enfermedad, la información frecuentemente utilizada son los registros de datos, anuarios estadísticos y otras fuentes que aportan información valiosa. Los estudios ecológicos pueden clasificarse: primero por el método de medición de la exposición (exploratorios y analíticos), y segundo el método de agrupamiento que se vaya a utilizar (diseños multi-grupo, tendencias temporales, diseños mixtos). Royo & Moreno (2009) definen los tipos de estudios ecológicos como:

- **Diseños multi-grupo:** En este diseño se encuentran los estudios exploratorios multi-grupo, que tiene como objetivo la comparación de tasas de enfermedad entre muchas regiones en un mismo período, o se compara la frecuencia de la enfermedad a través del tiempo en una misma región (Borja-Aburto 2000). Dentro de esta categoría se incluyen todos los estudios de patrones espaciales que sugieren una hipótesis enfocada comúnmente en etiologías ambientales, es decir, que la existencia de “patrones” apunta hacia determinantes ambientales de la enfermedad. Por otra parte, cuando a este tipo de estudio se le incluye una o varias variables explicativas en los modelos, se le conoce como estudio analítico multi-grupo.

- **Diseños de tendencias temporales:** Este diseño contiene 2 estudios, los estudios exploratorios de tendencias temporales que realizan comparaciones de las tasas en el tiempo, en una población definida geográficamente. Los análisis edad-período-cohorte, tan utilizados en diferentes ámbitos, constituyen un diseño de este tipo. Y la inclusión de variables explicativas en un diseño de tendencias temporales lo convertiría en un estudio analítico de tendencias temporales (Royo & Moreno 2009).
- **Diseños mixtos:** Este último diseño es la combinación de las componentes espacial y temporal, de los 2 diseños anteriormente descritos. Estos modelos tratan de mostrar si una relación que se observa en un estudio multigrupo se mantiene en el tiempo, o al revés, si un patrón de tendencia temporal se mantiene entre diferentes grupos (Royo & Moreno 2009).

Cáncer: “Nombre dado a las enfermedades en las que hay células anormales que se multiplican sin control y pueden invadir los tejidos cercanos. Las células de cáncer también se pueden diseminar hasta otras partes del cuerpo a través del torrente sanguíneo y el sistema linfático” (NCI 2018).

Supervivencia relativa: según Ederer (1961) es una medida objetiva de la probabilidad de supervivencia del cáncer, objetiva porque controla las distintas causas de muerte diferentes a la muerte por cáncer. Está definida como el cociente entre la Supervivencia Observada (SO) y la Supervivencia Esperada (SE):

- **SO:** es la supervivencia de la cohorte de pacientes con cáncer, teniendo en cuenta el tiempo de seguimiento del individuo y si este ha fallecido o no, independientemente de la causa de muerte, durante un intervalo específico (período de seguimiento).
- **SE:** para su cálculo se implementan tablas de vida esperadas o mortalidad, que representen la supervivencia de la población general. Los tres métodos más comunes para calcular la SE son: Ederer I (Ederer 1961), Ederer II (Ederer & Heise 1959b) y Hakulinen (Hakulinen 1982). La diferencia entre los 2 métodos de Ederer se encuentra en el número de pacientes que se considera en cada momento. Ederer I considera para cada intervalo a todos los pacientes, mientras que Ederer II sólo considera a los que están en riesgo durante ese intervalo (Ederer & Heise 1959a). Tanto el método de Ederer II como el de Hakulinen, son los más sugeridos, ya que minimizan el sesgo en la estimación de la SE, que se puede dar a causa de la heterogeneidad de las edades de los pacientes de la cohorte (Ederer 1961)(Hakulinen 1982).

Supervivencia relativa utilizando un enfoque de período :

Brenner & Gefeller (1996) sugirieron estimar la supervivencia basada en individuos con cáncer, utilizando un enfoque de período en lugar de un método de cohorte (completo). Con el análisis del período, cada probabilidad condicional se estima en función de la experiencia

de supervivencia de los pacientes recién diagnosticados (información más actualizada).

El análisis del período proporciona predicciones más precisas para el pronóstico de individuos recién diagnosticados y puede detectar tendencias temporales en la supervivencia de pacientes antes que el enfoque de cohorte tradicional (Brenner & Hakulinen 2009).

		Año de seguimiento									
		2006	2007	2008	2009	2010	2011	2012	2013	2014	
Año de diagnóstico	2006	---	---	---	---	---	---				Período
	2007		---	---	---	---	---	---			
	2008			---	---	---	---	---	---		
	2009				---	---	---	---	---	---	
	2010					1	1/2	2/3	3/4	4/5	
	2011						1	1/2	2/3	3/4	
	2012							1	1/2	2/3	
	2013								1	1/2	
	2014									1	
Año de diagnóstico	2006	1	1/2	2/3	3/4	4/5	5				Completo
	2007		1	1/2	2/3	3/4	4/5	5			
	2008			1	1/2	2/3	3/4	4/5	5		
	2009				1	1/2	2/3	3/4	4/5	5	
	2010					1	1/2	2/3	3/4	4/5	
	2011						1	1/2	2/3	3/4	
	2012							1	1/2	2/3	
	2013								1	1/2	
	2014									1	
Año de diagnóstico	2006	1	1/2	2/3	3/4	4/5	5				Cohorte
	2007		1	1/2	2/3	3/4	4/5	5			
	2008			1	1/2	2/3	3/4	4/5	5		
	2009				1	1/2	2/3	3/4	4/5	5	
	2010					1	1/2	2/3	3/4	4/5	
	2011						1	1/2	2/3	3/4	
	2012							1	1/2	2/3	
	2013								1	1/2	
	2014									1	

Figura 2-1.: Período de estudio para un enfoque de período (superior), completo (centro) y cohorte (inferior).

En la Figura 2-1 se puede observar los periodos de estudio para los distintos tipos de enfoque, que para un enfoque de período se trabaja con los tiempos de exposición de las pacientes diagnosticadas en el periodo 2010-2014. Para un enfoque completo se utilizó los tiempos de exposición de las mujeres reportadas en el período 2006-2014, con seguimiento hasta el final del año 2017. Por último, si en el enfoque de cohorte se implementara la información de las mujeres diagnosticadas entre los años 2006-2010, se debería tener

información de seguimiento de las pacientes hasta el año 2015, por lo tanto, no se podría aplicar este tipo de enfoque a la presente investigación.

Factor de riesgo: “Un factor de riesgo es cualquier rasgo, característica o exposición de un individuo que aumente su probabilidad de sufrir una enfermedad o lesión.” (OMS 2018).

El envejecimiento es el factor de riesgo más importante para la mayoría de cánceres. El principal factor de riesgo del cáncer de cuello uterino es la infección por el Virus del Papiloma Humano (NCI 2017); otros factores del riesgo de esta neoplasia son: el tabaquismo, inmunosupresión, infección con clamidia, una alimentación con pocas frutas y verduras, tener sobrepeso, uso prolongado de anticonceptivos orales (píldoras para el control de la natalidad), uso de un dispositivo intrauterino, tener muchos embarazos a término, tener menos de 17 años en el primer embarazo a término, Dietilestilbestrol (DES) y antecedente familiar de cáncer de cuello uterino (ACS 2014).

Exceso de mortalidad:

Se define como la diferencia entre en numero de muertes observada y el numero de muertes esperadas en ausencia de la enfermedad (Imaz et al. 2006).

Sistema de información geográfica (SIG):

Los SIG son un conjunto que mezcla hardware, software y datos geográficos, los cuales brindan una representación gráfica, este sistema permite relacionar cualquier tipo de dato con una localización geográfica, es decir, en un solo mapa el sistema muestra la distribución de recursos, edificios y/o poblaciones, sobre datos de los municipios, departamentos, regiones o todo un país; los SIG están diseñados para capturar, almacenar, manipular, analizar y desplegar la información de todas las formas posibles de manera lógica y coordinada (Monroy Pacheco & Vega Valero 2014).

Los SIG por sí solos son un instrumento de gran utilidad útil, pero cuando se une a las técnicas de análisis espacial muestran todas sus ventajas, es decir, las técnicas de análisis espacial son un complemento extraordinario cuando se apoyan en un SIG, dado que en la actualidad, la capacidad de análisis espacial de los SIG son muy limitadas, al menos, en el ámbito de la epidemiología (Ortega & Martí 2001).

En los Sistemas de Información Geográfica se utiliza comúnmente (Ortega & Martí 2001):

- **Geocodificación:** Es el proceso de asignar coordenadas geográficas (latitud – longitud), a los puntos de interés (dirección o nombre geográfico), en nuestro caso sería la localización de los casos de cáncer de cuello uterino en el mapa de la ciudad de Santiago de Cali, estas coordenadas geográficas asignadas pueden ser usadas para localizar un punto observado en el mapa.

- **Georreferenciación:** La georreferenciación se refiere a los procedimientos que permiten asociar entidades cartográficas o fenómenos (edificaciones, cultivos, etc) a una localización geográfica única, definida con precisión y teniendo en cuenta un sistema de referencia base.

2.2. Marco Teórico Estadístico.

2.2.1. Estadística Espacial.

“La estadística espacial es la reunión de un conjunto de metodologías apropiadas para el análisis de datos que corresponden a la medición de variables aleatorias en diversos sitios (puntos del espacio o agregaciones espaciales) de una región. De manera formal se puede decir, que la estadística espacial trata con el análisis de realizaciones de un proceso estocástico $\{Z(s) : s \in D \subset R^p\}$, en el que s es la ubicación en el espacio Euclidiano P -dimensional y $Z(s)$ es una variable aleatoria en la ubicación s ” (Giraldo 2011). La estadística espacial se subdivide en 3 grandes áreas, esta distinción se asocia a la característica del conjunto D de índices del proceso estocástico de interés, es decir, que las ubicaciones s pueden provenir de un conjunto D continuo o discreto y además si la ubicaciones son seleccionadas por el investigador (D fijo) o no (D aleatorio). A continuación describen las características D de dichas áreas.

Geoestadística

Las ubicaciones s provienen de un conjunto D continuo y estas ubicaciones son seleccionadas a conveniencia del investigador o bajo algún esquema de muestreo probabilístico, cuando se habla de un conjunto D continuo, se hace referencia a continuidad en el espacio, esto quiere decir, que la variable de interés puede ser medida en cualquier sitio de la zona de estudio. Es importante resaltar que en geoestadística el propósito esencial es la interpolación y si no hay continuidad espacial, pueden hacerse predicciones carentes de sentido (Giraldo 2011). La geoestadística es una herramienta que dispone de una forma de describir la continuidad espacial, que es una característica distintiva y esencial de muchos fenómenos naturales (Isaaks & Srivastava 1989).

Datos de Área (Lattices)

las ubicaciones s pertenecen a un conjunto D discreto y son seleccionadas por el investigador (D fijo). Estas pueden estar regular o irregularmente espaciadas, cuando se dice que el conjunto D es discreto y fijo, se dice que la variable no puede ser medida en cualquier sitio de la zona de estudio, pero el investigador decide donde se realizan estas mediciones, como por ejemplo la producción cafetera (en kilogramos) de las fincas en Colombia, para este caso, no todas las fincas de Colombia producen café, por esta instancia D es discreto, puesto que no hay continuidad de la medición de la variable de interés, pero el investigador

puede decidir en qué fincas hacer la medición (D fijo), es decir, solo realizar la medición en las fincas que produzcan café (Giraldo 2011).

Se considera los datos espaciales en un conjunto D , tal conjunto es una finita (o contable) colección de ubicaciones espaciales donde se hacen observaciones. La colección de puntos en D se conoce como un “lattice”. Los sitios espaciales en un lattice son generalmente identificados usando su longitud (x) y su latitud (y); Una característica importante de los datos de área es un “vecindario”, el vecindario de un área será el conjunto de áreas que están cerca, en términos de distancia, un vecindario también está definido como el conjunto de áreas que rodean un área en específico (Cressie 1993).

Patrones puntuales

Los patrones puntuales se caracterizan de los demás por que las ubicaciones que provienen del conjunto $D \subset R^p$, dicho conjunto puede discreto o continuo y los lugares donde puede ocurrir el evento de interés está dado por la naturaleza del mismo, por lo tanto los sitios donde se puede medir la variable de interés no depende del criterio del investigador; estos puntos se presentan aleatoriamente en el espacio, es decir, que no se sabe específicamente donde va a ocurrir un evento del proceso Algunos ejemplos de esta área son los siguientes: Localización de nidos de pájaros en una región dada, puntos de imperfectos dentro de una placa metálica, ubicación de los lugares de terremoto en Colombia, etc. En los anteriores ejemplos está claro el concepto de la aleatoriedad de los puntos espaciales, ya que estas ubicaciones no dependen del criterio del investigador (Giraldo 2002)

“Los patrones puntuales surgen cuando la variable importante a analizar es la ubicación de los eventos. Muy a menudo, la primera pregunta que se debe responder es si el patrón exhibe aleatoriedad espacial completa, agrupamiento o regularidad” (Cressie 1993).

Los modelos de patrones puntuales son en gran parte derivados de los patrones espaciales completamente aleatorios, estos patrones tiene como característica que el promedio de eventos por unidad de área, la intensidad $\lambda(s)$, es homogénea a lo largo D , el número de eventos en dos subregiones que no se solapan, A_1 y A_2 son independientes, y el número de eventos en cualquier subregión sigue una distribución de Poisson (Schabenberger & Gotway 2005).

Los patrones espaciales completamente aleatorios pueden ser independientes en subregiones que no se cruzan, pero el promedio de eventos por unidad de área $\lambda(s)$, no es homogénea en el conjunto D , es decir, que los puntos pueden presentarse en regiones con un grado de intensidad alto o bajo; y también se podría ver el caso donde para cada región la intensidad $\lambda(s) = \lambda$, es decir, que se presente una intensidad constante. Ahora podremos diferir entre patrones aleatorios, regulares, o agregados. En el caso de patrones uniformes o regulares, la

probabilidad de encontrar un punto a una longitud mínima de otro punto es menor de la que tendría un patrón aleatorio, mientras que en los patrones agrupados la probabilidad es mayor (De la Cruz 2006).

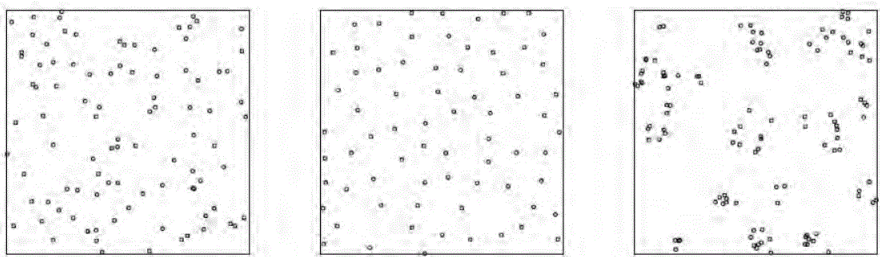


Figura 2-2.: Patrones de puntos: aleatorio (izquierda), regular (centro), agrupado (derecha).
Fuente: Cressie, N., 1993. Statistics for Spatial data

De la figura 2-2 podemos observar que los patrones de puntos aleatorios no tienen ninguna estructura: se podría decir, que las posiciones de los puntos son independientes entre sí, en cambio en los patrones de puntos regular notamos que la intensidad es constante en todo el espacio y los puntos se encuentran alejados entre sí; por último para los patrones de puntos agrupados se observa que la intensidad de los puntos es alta en algunas zonas y en otras es muy baja .

Intensidad

La intensidad es un factor muy importante en el análisis de las áreas de estudio, porque esta indica la cantidad de eventos ocurridos por unidad de área (Cressie 1993), es denotado como:

$$\lambda = \text{Número de eventos por unidad de área}$$

En general, si las áreas son muy pequeñas, se tiene que:

$$\lambda(s) = \lim_{|s| \rightarrow \infty} \frac{E[N(s)]}{\nu(s)}$$

Pero si el patrón espacial es completamente aleatorio entonces se tiene que:

$$\lambda(s) = \lambda = \frac{n}{\nu(D)}$$

Donde $N(s)$ denota el número de eventos en la región s , $\nu(s)$ representa el area de la región s y n como el número de regiones.

Estimación Kernel

La estimación Kernel es un método no paramétrico de estimación que no tiene en cuenta

los supuestos rígidos sobre los datos como ocurre en el planteamiento paramétrico. Las son un tipo de funciones utilizadas para estimar modelos de regresión, para estimar densidades de funciones no lineales, y para estimar intensidades (Silverman 1986).

Función Kernel

En el caso bi-dimensional, la función Kernel univariada requiere ser remplazada por una función que se ajuste a las coordenadas del plano. Un producto de funciones kernel puede ser extraído multiplicando dos funciones kernel univariadas. La elección de estas funciones se pueden hacer a juicio del investigador, lo que genera la ausencia de interacción entre las coordenadas (Giraldo 2011).

En un patrón completamente aleatorio la estimación de λ viene dada como:

$$\hat{\lambda} = \frac{n}{\nu(A)} = \frac{\text{número de eventos}}{\text{area de } A}$$

si las coordenadas x_i y y_i son las de localización del sitio s_i , entonces el producto kernel implica un estimador de la intensidad dado por:

$$\hat{\lambda} = \frac{1}{\nu(A)h_xh_y} \sum_{i=1}^n K\left(\frac{x-x_i}{h_x}\right)K\left(\frac{y-y_i}{h_y}\right) = \frac{1}{\nu(A)} \sum_{i=1}^n \frac{1}{h} K\left(\frac{s-s_i}{n}\right)$$

Donde h_x y h_y son los anchos de banda en las respectivas direcciones del sistema de coordenadas y K es una función denotada como función Kernel o núcleo. Por otra parte, si se presentan problema con el efecto de borde $\delta(S)$, la expresión anterior no tiene en cuenta los efectos de borde, que pueden ser significativos, por lo que Diggle (1985) sugirió una estimación de la intensidad de kernel corregida para un solo ancho de banda.

$$\hat{\lambda} = \frac{1}{\delta(S)} \sum_{i=1}^n \frac{1}{h^2} K\left(\frac{s-s_i}{n}\right) \quad , \text{ donde } \delta(S) = \int_A \frac{1}{h^2} K\left(\frac{s-u}{n}\right) du$$

La función Kernel $K(u)$ o núcleo, donde $u \in \mathbb{R}$, es una función simétrica en torno a cero (0) y se debe cumplir que:

1. $\int K(u) = 1$
2. $\int uK(u) = 0$

Entonces algunas funciones Kernel son las siguientes:

- Gaussianiana $\Rightarrow K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

$$\begin{aligned}
\blacksquare \text{ Cuadratica } \Rightarrow K(x) &= \begin{cases} 0.75(1-x^2) & |x| \leq 1 \\ 0 & \text{otro caso} \end{cases} \\
\blacksquare \text{ Mínima varianza } \Rightarrow K(x) &= \begin{cases} 0.375(3-5x^2) & |x| \leq 1 \\ 0 & \text{otro caso} \end{cases}
\end{aligned}$$

Las funciones Kernel son muy convenientes cuando el espacio donde suceden los eventos es continuo, pero como la ubicación de los casos de cáncer de cuello uterino suceden de forma aleatoria en el ciudad de Santiago de Cali, es decir, que pueden haber barrios e incluso cuadras donde no se reporten casos de CCU, por lo tanto se plantea calcular el ancho de banda óptimo para estimar y graficar la función de intensidad, esto a pesar de no solucionar el problema, reduce el sesgo en la estimación de la función de intensidad (Espinal & Aruneri 2014). Para la selección del ancho de banda de la intensidad Kernel se utilizó el criterio de validación cruzada y el criterio de validación cruzada de probabilidad.

Validación Cruzada.

La validación cruzada para la selección del ancho de banda del kernel (núcleo) de la intensidad, es un proceso el cual selecciona un ancho sigma (σ) de banda apropiado para el estimador del kernel de la intensidad del proceso puntual. σ se escoge para minimizar el criterio del error cuadrático medio definido por Diggle (1985) y en base a el método de Berman & Diggle (1989), se ejecuta el algoritmo para calcular el valor:

$$M(\sigma) = \frac{MSE(\sigma)}{\lambda^2} - g(0)$$

La seleccion del ancho de banda deseado, está en función del ancho de banda de σ , donde $MSE(\sigma)$ es el error cuadrático medio en el ancho de banda σ , λ es la intensidad promedio y $g(0)$ es la función de correlación de pares.

Verosimilitud de la validación cruzada.

Verosimilitud de la validación cruzada o la validación cruzada de probabilidad se implementa para seleccionar un ancho de banda suavizado para la estimación del kernel de la intensidad del proceso puntual, esta función selecciona el ancho de banda más apropiado σ y se elige para maximizar el criterio de validación cruzada de probabilidad del proceso puntual, denotado como:

$$LCV(\sigma) = \sum_i \log(\hat{\lambda}_i(x_i)) - \int_W \hat{\lambda}(u) du$$

Donde $\hat{\lambda}_i(x_i)$ es la estimación de suavizado del kernel de la intensidad de salida en x_i con ancho de banda de suavizado σ y $\hat{\lambda}(u)$ es la estimación del suavizado del kernel de la intensidad en la ubicación espacial u con el ancho de banda de suavizado σ (Loader 2006).

Test de completa aleatoriedad espacial

Son algunas pruebas para identificar el tipo de patrón espacial que tendrían los datos (aleatorio, agregado o regular), esto se hace con el fin de conocer el tipo de patrón espacial de los 2 años a estudiar en la presente investigación, de ahí observaremos si en los 2 períodos de tiempo la distribución de las ocurrencias de cáncer de cuello uterino son las mismas o han cambiado. Algunas de estas pruebas son las pruebas basadas en métodos de Monte Carlo, los test basados en cuadrantes y los test basados en distancias (Schabenberger & Gotway 2005).

Métodos basados en cuadrantes

Uno de los métodos para probar la aleatoriedad espacial de un patrón son basados en la división del dominio D en sub-regiones o cuadrantes, no traslapadas de igual tamaño, estas sub-regiones las definimos como A_i , con $i = 1, \dots, k$.

se debe cumplir que $\cup_{i=1}^k A_i = D$ donde A_i y A_j no se traslapen.

Test de bondad de ajuste X^2

Esta es una prueba de ajuste estadístico y tiene como hipótesis nula que los n puntos son distribuidos uniformemente independientes a lo largo de D , es decir, que los conteos de sitios por sub-región son variables Poisson independientes con media común. Es decir, la región de estudio es dividida en cuadrantes de igual tamaño.

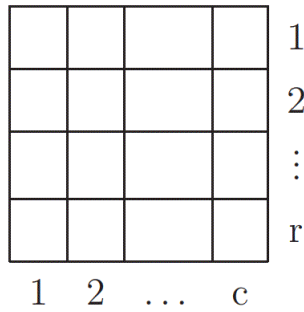


Figura 2-3.: Región de dominio D dividida en Sub-regiones

De la figura 2-3, se puede observar que no necesariamente la sub-división del espacio D tiene que ser cuadrada, pero cada uno de ellos deben de contener igual cantidad de área. Y se define n como el número total de eventos, y n_{ij} como el número total de eventos en la cuadrícula ij , con $i = 1, \dots, r$ y $j = 1, \dots, c$. El estadístico de prueba para este caso viene dado como:

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - \bar{n})^2}{\bar{n}}, \text{ donde } \bar{n} = \frac{n}{rc}$$

El **test de Montecarlo** es muy similar a la prueba de bondad de ajuste X^2 , ya que manejan el mismo estadístico de prueba, lo que cambia es que este tiene como requisito que en un cuadrante deben de existir por lo menos 5 eventos.

Métodos basados en distancias

Son pruebas estadísticas basadas en distancias entre eventos o entre puntos muestreados y eventos. Estos métodos basados son más robustos ya que eliminan las características de cómo va hacer la forma de elegir los cuadrantes y cuantas sub-divisiones se necesitan para llevar a cabo una decisión en el contraste de hipótesis de aleatoriedad. A continuación se presentarán algunas pruebas que se basan en dicha distancia (Cressie 1993):

Función G

Esta función tiene en cuenta la mínima distancia entre eventos, es decir, que tiene en cuenta la distancia al vecino más cercano $G(r)$, donde $G(r)$ es la probabilidad de que la distancia de un evento elegido al azar a su evento más cercano sea menor o igual a r (Cressie 1993). Su función de distribución empírica esta denotada como:

$$\hat{G}(R) = \frac{N^o\{r_i \leq r\}}{n}$$

donde el numerador es el número de elementos en el conjunto de distancias que son menores que o igual a r y n es el número total de puntos, r_i es la distancia mínima entre un evento y sus vecinos, R es la variable aleatoria distancia al lugar del evento más próximo y $G(R)$ es la función de distribución de R .

Distribución teórica $G_{theo}(r)$

$G(r) = P(R \leq r) = 1 - P(R > r) = 1 - P(N(A) = \pi r^2)$, donde $N(A)$ es aleatoriamente espaciado. Entonces si $N(A) \sim \text{Poisson}(\lambda \pi r^2)$, se tiene que la $P(N(A) = 0) = e^{-\lambda \pi r^2}$, donde λ representa el número de eventos por unidad de área, Entonces bajo aleatoriedad se tiene que:

$$G(r) = 1 - e^{-\lambda \pi r^2}$$

Entonces si $\hat{G}(R)$ crece rápidamente en distancias cortas, los eventos son agregados, por el contrario si crece lentamente hasta cierta distancia y después crece rápidamente los eventos son regularmente espaciados.

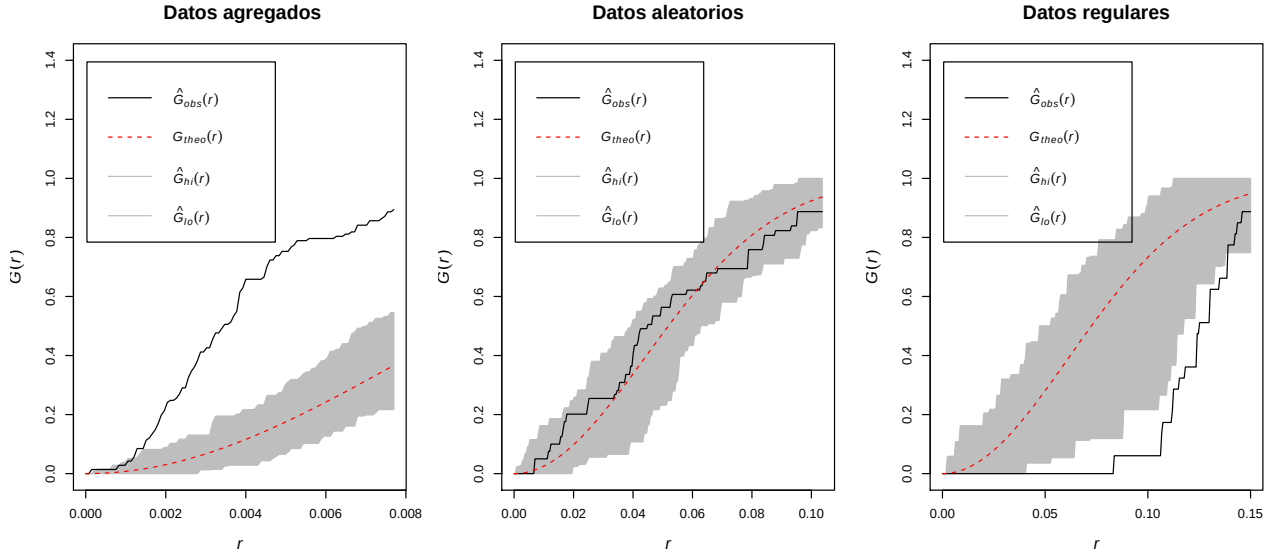


Figura 2-4.: Prueba de completa aleatoriedad G aplicada a datos simulados con diferente comportamiento espacial.

En la Figura 2-4 se puede observar el gráfico de la función $G(r)$ para tres (3) conjuntos de datos con patrón agregado (izquierda), aleatorio (centro) y regular (derecha). En los gráficos se puede apreciar que para un patrón agregado la curva empírica de los valores observados $\hat{G}_{obs}(r)$ además de crecer rápidamente en distancias cortas, también está por encima de la curva distribución teórica $G_{theo}(r)$; para un patrón aleatorio se puede ver que $\hat{G}_{obs}(r) = G_{theo}(r)$ y para un patrón regular la curva $\hat{G}_{obs}(r)$ está por debajo de la curva $G_{theo}(r)$; en donde r representa las distancias, $\hat{G}_{hi}$ y $\hat{G}_{lo}$ conforman las bandas de confianza superior e inferior, respectivamente.

Función F

La función del espacio vacío $F(r)$ también llamada la “distancia desde el punto más cercano al evento” tiene en cuenta la mínima distancia punto-evento, donde un punto es un evento escogido aleatoriamente dentro de la región de estudio (Cressie 1993). Para su elaboración se siguen los siguientes pasos:

1. Se selecciona aleatoriamente m puntos p_i , con $i = 1, \dots, m$.
2. Se calcula $r_i = r(p_i, s_i)$, que es la distancia de cada punto escogido al azar, al sitio del evento más cercano s_i . Ahora definimos R como la variable aleatoria de mínima distancia de un punto a un evento, entonces se tiene que:

$$\hat{F}(r) = \frac{N^{\circ}\{r_i \leq r\}}{m}, \text{ su función teórica es } F_{theo}(r) = 1 - e^{-\lambda\pi r^2}$$

Entonces se tiene que si $\hat{F}(r)$ crece lentamente en distancias cortas y rápidamente para distancia largas, el patrón espacial de los datos es agregado, pero Si $\hat{F}(r)$ crece rápidamente al comienzo o crece rápidamente para distancias cortas r , el patrón es considerado regular.

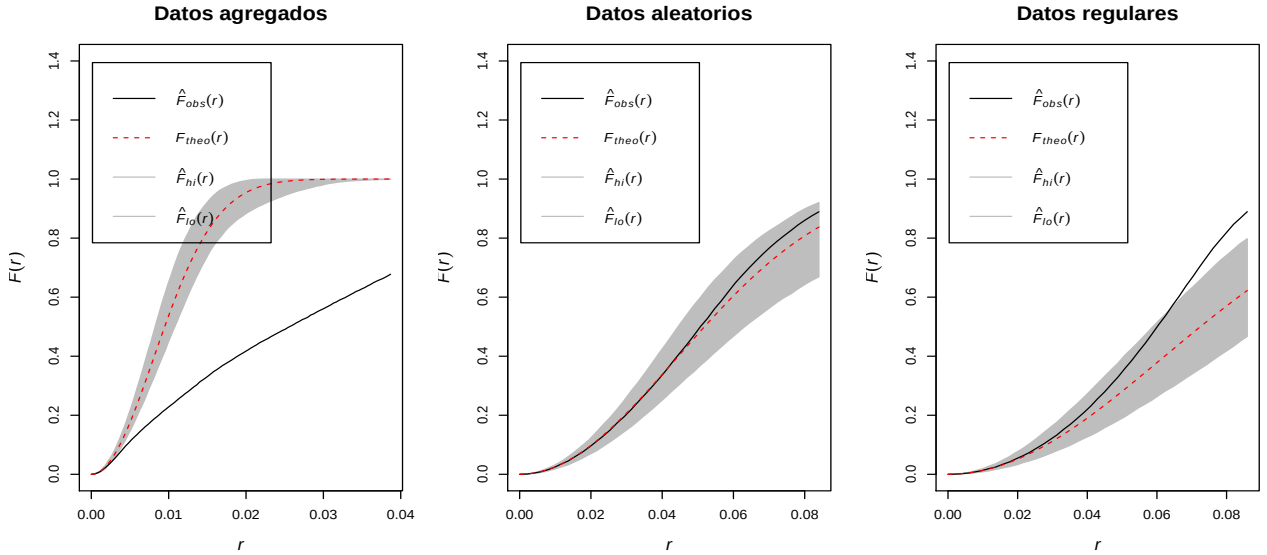


Figura 2-5.: Prueba de completa aleatoriedad $F(r)$ aplicada a datos simulados con diferente comportamiento espacial.

En la Figura 2-5 se pueden observar los gráficos de la prueba F aplicado a tres conjuntos de datos con patrón agregado (izquierda), aleatorio (centro) y regular (derecha). Se puede analizar que para un patrón agregado la curva de los valores observados $\hat{F}_{obs}(r)$ crece lentamente al comienzo y también que se encuentra por debajo de la curva teórica $G_{theo}(r)$; para un patrón aleatorio se puede ver que $\hat{F}_{obs}(r) = F_{theo}(r)$; para un patrón regular la curva $\hat{G}_{obs}(r)$ está un poco por encima de la curva $G_{theo}(r)$.

Función K de Ripley

La función K de Ripley (1976) es una función de λ_2 para procesos estacionarios e isotrópicos. También es conocida como la medida reducida del segundo momento o segunda función de momento reducido (Cressie 1993). Esta función se define como el número de individuos en un radio (r) alrededor de cualquier individuo. Esta función es un estimador de segundo orden, es decir, es como la varianza del proceso puntual, y se encuentra dada por:

$$K(r) = \frac{E[N^{\circ} \text{ de eventos a una distancia } r \text{ de un evento elegido aleatoriamente}]}{\lambda}$$

Entonces:

$$\hat{K}(r) = \frac{\sum \sum_{i \neq j} I_r(r_{ij})}{\hat{\lambda}} \quad , \text{ donde } \quad \hat{\lambda} = \frac{n}{A}$$

Donde n es el número de puntos del patrón, A es la superficie del área de estudio y $I(r_{ij} < r)$ es la función indicadora, que toma el valor 1 si la distancia entre los puntos i y j es menor que r , y en el caso contrario toma el valor 0. Bajo Aleatoriedad se tiene que:

$$K(r) = \frac{\lambda \pi r^2}{\lambda} = \pi r^2$$

Entonces si $\hat{K}(r) < \pi r^2$ los datos se distribuyen regularmente en el espacio, pero si $\hat{K}(r) > \pi r^2$ entonces se dice que los datos son agregados, es decir, que los datos se distribuyen a manera de grupos.

Función K de L

También llamada la función L , no es más que otra función de segundo orden implementada en el estudio de dependencia en patrones puntuales, basada en la función K y propuesta por Besag (1977), la función L está denotada como:

$$\hat{L}(r) = \sqrt{\frac{\hat{K}(r)}{\pi}}, \text{ y bajo aleatoriedad } L(r) = \sqrt{\frac{\pi r^2}{\pi}} = r$$

Como podemos observar que la función L bajo aleatoriedad es una función lineal, entonces si el patrón es aleatorio no debe haber desviación de la recta y $\hat{L}(r) - r = 0$, caso contrario se puede pensar que si $\hat{L}(r) - r > 0$ el proceso puntual es agregado y si $\hat{L}(r) - r < 0$ el proceso puntual es regular.

En la Figura 2-6 podemos apreciar la aplicación de la prueba K para los tres distintos conjuntos de datos con patrón agregado (izquierda), aleatorio (centro) y regular (derecha); y en la Figura 2-7 se puede ver el resultado de la prueba L efectuado en los distintos tipos de datos simulados. Para ambas pruebas se puede detallar que para un patrón espacialmente agregado la curva de distribución empírica $\hat{K}_{obs}(r)$ y $\hat{L}_{obs}(r)$ están sobre las curvas teóricas correspondientes $K_{theo}(r)$ y $L_{theo}(r)$; para un patrón aleatoriamente espaciado no se evidenciaron desviaciones entre la curva empírica y teórica; para un patrón regularmente espaciado la curva empírica presenta desviaciones por debajo de la curva teórica en distancias cortas.

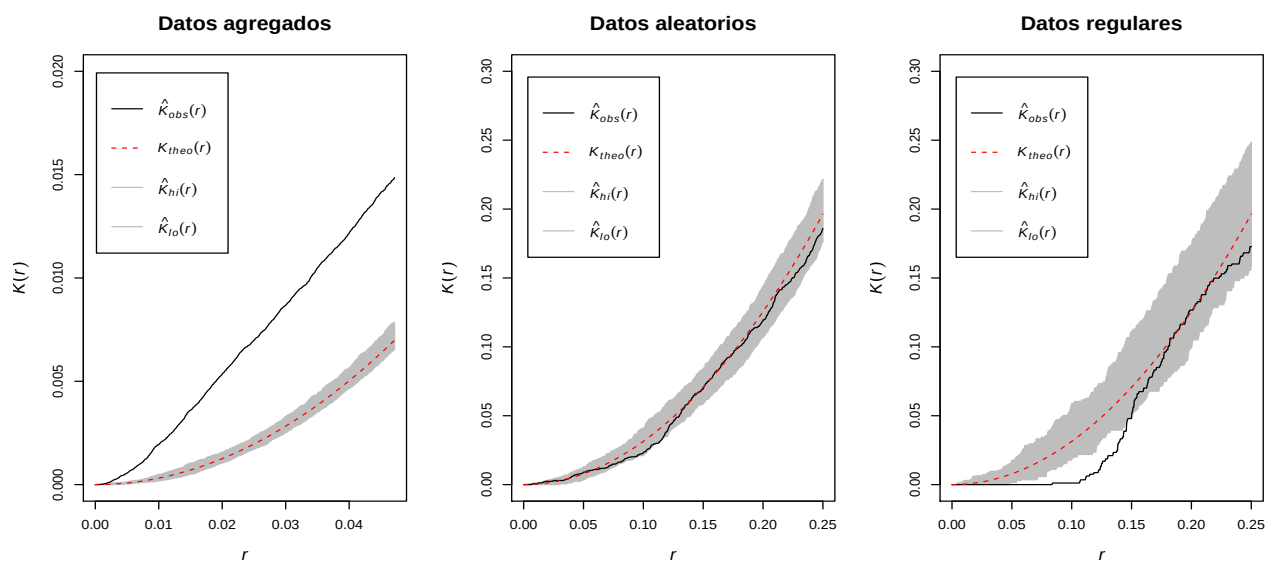


Figura 2-6.: Prueba de completa aleatoriedad K aplicada a datos simulados con diferente comportamiento espacial.

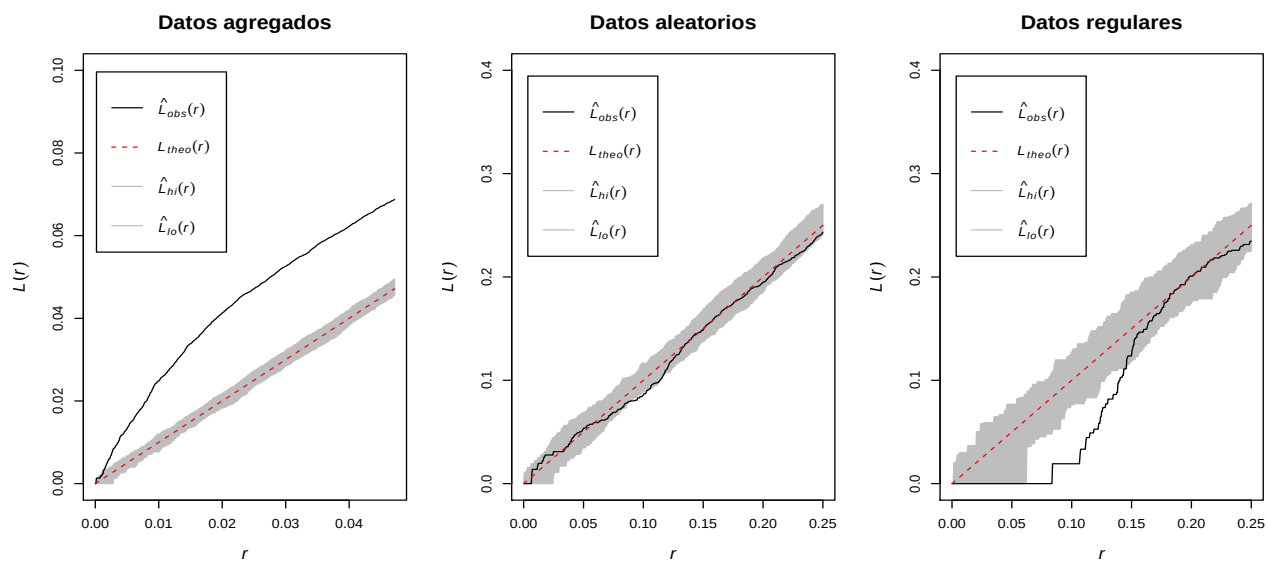


Figura 2-7.: Prueba de completa aleatoriedad K de L aplicada a datos simulados con diferente comportamiento espacial.

2.2.2. Función de supervivencia con enfoque de período.

Para el análisis de supervivencia con un enfoque de periodo se requiere de la siguiente información de los pacientes: fecha de diagnóstico, fecha de finalización del seguimiento y estado vital (vivo o muerto) en el fin del seguimiento (Brenner & Gefeller 1996). La estimación de la función de supervivencia al final del año de seguimiento k ($\hat{S}_k$) se escribe como:

$$\hat{S}_k = \prod_{x=1}^k p_x$$

Donde p_x se define como la probabilidad condicional de sobrevivir al año de seguimiento x (a condición de que un paciente esté vivo al comienzo del año de seguimiento) (Cutler & Ederer 1958). Estas probabilidades condicionales p_x , $1 \leq x \leq k$ se pueden estimar por $1 - \frac{d_x}{L_x}$, donde d_x representa el número de muertes y L_x simboliza la cantidad de pacientes en riesgo durante el año de seguimiento x . La supervivencia relativa se propone para evaluar la supervivencia del paciente (con una enfermedad en particular), corregida por las otras causas de muerte (Ederer 1961). Las Las probabilidades de supervivencia condicional relativa pueden estimarse por:

$$\frac{1 - d_x/L_x}{E_x} = \frac{L_x - d_x}{L_x \cdot E_x}$$

Donde E_x indica la probabilidad estimada de supervivencia y $(L_x \cdot E_x)$ indica el número estimado de sobrevivientes entre los pacientes en riesgo durante el año de seguimiento x en ausencia de cáncer. La información sobre E_x se obtiene generalmente de las tablas de vida para la población general.

2.2.3. Modelo Lineal generalizado.

Nelder & Baker (1972), originaron la idea de utilizar los GLM (Generalized Linear Models). El GLM se define en términos de un conjunto de variables aleatorias independientes $Y_1, Y_2, \dots, Y_N$ y de una componente sistemática, en la cual, la componente aleatoria proviene de las distribuciones pertenecientes a la familia exponencial y cumplen las siguientes propiedades (Dobson & Barnett 2008):

1. La distribución de cada Y_i 's tiene la forma canónica y depende de un solo parámetro θ .

$$f(y_i; \theta_i) = \exp[y_i b(\theta_i) + c_i(\theta_i) + d_i(y_i)]$$

2. Las distribuciones de todas las Y_i son de la misma forma (por ejemplo, todas son binomial), de modo que los subíndices de b , c y d no son requeridos. Por lo tanto, la función de densidad de probabilidad conjunta de $Y_1, Y_2, \dots, Y_n$ es definida como:

$$f(y_1, \dots, y_N; \theta_1, \dots, \theta_N) = \prod_{i=1}^N \exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)] = \exp\left[\sum_{i=1}^N y_i b(\theta_i) + \sum_{i=1}^N c(\theta_i) + \sum_{i=1}^N d(y_i)\right]$$

Los parámetros θ_i no son el principal interés (ya que se puede haber uno para cada observación). La especificación del modelo por lo general es de interés en un conjunto más pequeño de parámetros $\beta_1, \dots, \beta_p$, donde $p < n$. El segundo componente que es la parte sistemática $f(\cdot)$, también llamada función de enlace o vínculo, la cual se encarga de linealizar la relación entre la variable de respuesta y las variables independientes mediante la transformación de la variable de respuesta $E(Y_i) = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki}$ (Cayuela & Bolancé 2016).

Supongamos que $E(Y_i) = \mu_i$, donde μ_i es alguna función de θ_i . Para un modelo lineal generalizado se tiene una transformación para μ_i de la forma (Dobson & Barnett 2008):

$$g(\mu_i) = X_i^T \beta$$

Donde $g(\cdot)$ es una función monótona y diferenciable también llamada función de enlace; El vector de variables explicativas $\vec{X}_i$ (puede ser de covariables o niveles de factores) es de dimensión $px1$; y β es el vector de parámetros de dimensión $px1$.

El modelo lineal generalizado tiene tres componentes (McCullagh & Nelder 1989):

- **Componente aleatoria:** Corresponde a la variable de respuesta o dependiente, y esta proviene de una distribución de la familia exponencial.
- **Componente sistemática:** También es llamado como predictor lineal, se denota por g y corresponde a $g(\mu_i) = X_i^T \beta$,
y a esta ecuación pertenece la función de enlace.
- **Función de enlace:** $g(\cdot)$ relaciona la esperanza matemática de la variable de respuesta con el predictor lineal. Además, linealiza la relación entre la variable respuesta y la(s) variable(s) independiente(s) mediante la transformación de la variable de respuesta.

Como sabemos las variables aleatorias independientes $Y_1, Y_2, \dots, Y_N$ provienen de distribuciones pertenecientes a la familia exponencial, esta diversidad dependen del valor de la variable de respuesta, por ejemplo: datos de conteo (Distribución Poisson o Binomial Negativa), respuesta binaria (distribución Binomial), tiempos de espera (Distribución Binomial Negativa), respuestas continuas estrictamente positivas (distribución Gamma), etc. Por lo tanto cada distribución sugiere una función de enlace.

Tabla **2-1.**: Funciones de enlace comúnmente utilizadas en modelos lineales generalizados (Cayuela 2009).

Nombre	Función de enlace	Utilización
Identidad	μ	datos continuos con errores normales
Logaritmica	$\text{Log}(\mu)$	conteos con errores de tipo poisson
Logit	$\text{Log}\left(\frac{\mu}{n - \mu}\right)$	Proporciones(entre 0 y 1) con errores binomiales
Reciproca	$\frac{1}{\mu}$	datos continuos con errores gamma
Raiz cuadrada	$\sqrt{\mu}$	Conteos
Exponencial	μ^n	Función de potencia

En la Tabla **2-1** podemos observar la función de enlace, su nombre y su respectivo uso. Para la distribución normal se sugiere utilizar el enlace identidad; para la Binomial se sugiere hacer uso de la Logit, Probit, Log-Log o Log-Log Complementaria; para datos de conteo se sugiere emplear la transformación Logarítmica (Madsen & Thyregod 2010).

2.2.4. GLM con estructura de error Poisson.

En la presente investigación se incluyeron las categorías por edad de acuerdo con los Estándares Internacionales de Supervivencia del Cáncer (ICSS), el cual sigue los siguientes grupos de edad: 15-44, 45-54, 55-64, 65-74 y 75 años o más (Corazziari et al. 2004).

El modelo lineal generalizado propuesto (Holleczek & Brenner 2013) modela el exceso de mortalidad. El número observado de muertes $d_{año, edad}$ está definido para cada combinación de las variables independientes: grupo de edad ($0 < edad < 4$) y año de seguimiento ($1 < año < 5$).

$$d_{año, edad} = f(año, edad)$$

Como se asume, que la estructura de los errores son de la forma Poisson, entonces se utilizó la función de enlace logarítmica:

$$\text{Ln}(\mu_{año,gedad} - d_{año,gedad}^*)$$

Donde $\mu_{año,gedad}$ representa en el modelo el número de muertes observadas y $d_{año,gedad}^*$ el número de muertes esperadas (debido a causas distintas al cáncer de interés y estimadas a partir de las tasas de mortalidad de la población en general), definida como:

$$d_{año,gedad}^* = (I_{año,gedad} - \frac{d_{año,gedad}}{2}) \cdot \text{Ln}(\frac{I_{año,gedad} - e_{año,gedad}}{I_{año,gedad}})$$

Donde $I_{año,gedad}$ representa el número efectivo de personas en riesgo (incluyendo entradas tardías y el retiro como media persona) y $e_{año,gedad}$ el número esperado de muertes en ausencia del CCU, para cada posible combinación formada entre las 2 variables categóricas independientes. Por lo tanto el modelo queda denotado como:

$$\text{Ln}(\mu_{año,gedad} - d_{y,gedad}^*) = \text{Ln}(I_{año,gedad} - \frac{d_{año,gedad}}{2}) + \sum_{i=1}^5 \alpha_i año_i + \sum_{j=1}^4 \beta_j gedad_j$$

En el modelo anterior se excluye el intercepto; el $\text{Ln}(I_{año,gedad} - \frac{d_{año,gedad}}{2})$ representa el logaritmo de las personas en riesgo, que se introduce como compensación (offset) en el GLM; $\mathbf{X}$ simboliza el vector de covariables de las diferentes observaciones; α y β son los coeficientes del modelo para las 2 variables de estudio($año, gedad$). Si se desea obtener estimaciones de supervivencia absoluta, solo se tiene que $d_{año,gedad}^* = 0$.

Los coeficientes en el modelo calculado se interpretan como logaritmos del exceso de mortalidad relativa, entonces para el cálculo de las estimaciones de supervivencia relativa condicional para intervalos de 1 año después del diagnóstico, se utilizaron los coeficientes del modelo de la siguiente forma:

$$r_{año,gedad=0} = \exp(-\exp(\alpha_{año})) \quad \text{y} \quad r_{año,gedad=gedad} = \exp(-\exp(\alpha_{año} + \beta_{gedad}))$$

Para las estimaciones de la RS acumulada de 5 años se puede calcular como:

$$R_{gedad} = \prod_{año=1}^5 r_{año,gedad}$$

Estimación de parámetros

Teniendo en cuenta que en el modelo lineal generalizado la variable de respuesta tiene una distribución de tipo Poisson, se procede a estimar los $\hat{\beta}$ mediante los métodos de máxima verosimilitud utilizando procedimientos iterativos basados en el algoritmo de Newton Raphson, que se explicaran a continuación.

Máxima verosimilitud:

sean $Y_1, \dots, Y_n$ variables aleatorias independientes cada una con distribución Poisson, denotado como (Aldrich et al. 1997):

$$f(y_i; \lambda) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}, y_i = 0, 1, 2, 3, \dots$$

Donde λ es el parámetro de interés, entonces su distribución conjunta, o también conocida como función de verosimilitud $L(\lambda; y_1, \dots, y_n)$, viene dada como:

$$f(y_1, \dots, y_n; \lambda) = \prod_{i=1}^n f(y_i; \lambda) = \frac{\lambda^{\sum y_i} e^{-n\lambda}}{y_1! y_2! \dots y_n!}$$

Por facilidad se utiliza el logaritmo de la verosimilitud.

$$l(\lambda; y_1, \dots, y_n) = \log L(\lambda; y_1, \dots, y_n) = (\sum y_i) \ln \lambda - n\lambda - \sum (\ln y_i!)$$

Algoritmo de Newton Raphson.

El Algoritmo de Newton Raphson, es un proceso iterativo que a partir de una secuencia de estimaciones repetitivas se logra obtener que ciertos parámetros converjan y se estabilicen en un solo valor. La estimación de $\hat{\beta}$ por este método tiene la siguiente forma (Bianco 2010).

$$\beta^{(t+1)} = \beta^{(t)} - [\ell''(\beta^{(t)})]^{-1} \ell'(\beta^{(t)})$$

Donde $\ell(\beta) = \ell(\beta, y)$ es el logaritmo de la máxima verosimilitud y el cálculo de las derivadas de estas se expresa como:

$$\frac{\partial}{\partial \beta} \ell(\beta; y_i) = \sum_{i=1}^n \frac{y_i - \mu_i}{a(\phi) V(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ji}$$

$$\frac{\partial}{\partial \beta_j \partial \beta_k} \ell(\beta; y_i) = - \sum_{i=1}^n \frac{1}{a(\phi)} \left[\frac{1}{V(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 - (\mu_i - y_i) \frac{1}{V(\mu_i)^2} \left(\frac{\partial \mu}{\partial \eta} \right)_i \frac{\partial V(\mu_i)}{\partial \mu} - \frac{1}{V(\mu_i)} \left(\frac{\partial^2 \mu}{\partial \eta^2} \right)_i \right] x_{ji} x_{ki}$$

Con $i = 1, 2, 3, \dots, n$, $j = 1, 2, 3, \dots, k$. y $V(\mu)$ es la función de varianza estimada de μ .

Pruebas de Bondad de ajuste para GLM.

Razón de verosimilitud: La desviación del modelo contrasta el logaritmo de la verosimilitud del modelo ajustado con el logaritmo de la verosimilitud de un modelo saturado, que es un modelo que se ajusta perfectamente a los datos de la muestra y el mayor valor del logaritmo de la función de verosimilitud es cero (Montgomery et al. 2006).

$$\lambda(\vec{\beta}) = 2Ln(Modelosaturado) - 2Ln(\vec{\beta})$$

Si el tamaño de la muestra es grande entonces $\lambda(\vec{\beta}) \sim \chi^2_{(n-p)}$. Para valores grandes de $\lambda(\vec{\beta})$ indican que el modelo está bien ajustado, mientras que para valores pequeños implican que el modelo no está bien ajustado. El criterio de la prueba es el siguiente:

Si $\lambda(\vec{\beta}) > \chi^2_{(n-p)}$, el modelo está bien ajustado, por el contrario, si $\lambda(\vec{\beta}) < \chi^2_{(n-p)}$ entonces el modelo no es adecuado.

Tabla de análisis de Desvianza:

La prueba inicial de bondad de ajuste es un estadístico de una prueba correspondiente al modelo inicial, es frecuentemente representado en una tabla de análisis de desvianza, parecida a la tabla ANOVA de los modelos lineales Tabla **2-2**. En la tabla de la bondad del ajuste el estadístico de prueba correspondiente al modelo inicial es $G^2(H_1) = D(y; (\hat{\beta}))$ se muestra en la línea denominada "Error". El estadístico debe ser comparado con los percentiles de la distribución $\chi^2_{(n-k)}$. La tabla también muestra el estadístico de prueba para H_0 bajo el supuesto de que H_1 es cierto. Esta prueba investiga si al menos alguno de los parámetros es significativo.

Tabla **2-2**.: Evaluación inicial de la bondad del ajuste de un modelo H_1 , H_0 y $\hat{\mu}_0$ se representa a modelo mínimo, es decir, un modelo con todas las observaciones que tienen el mismo valor medio (Dobson & Barnett 2008).

Fuente	g.l ¹	Desvianza	Desvianza media	Interpretación de la prueba de bondad de ajuste
Modelo H_0	$k - 1$	$D(\mu(\hat{\beta}); \hat{\mu}_0)$	$\frac{D(\mu(\hat{\beta}); \hat{\mu}_0)}{k-1}$	$G^2(H_0 H_1)$
Residual (Error)	$n - k$	$D(y; \mu(\hat{\beta}))$	$\frac{D(y; \mu(\hat{\beta}))}{n-k}$	$G^2(H_1)$
Total	$n - 1$	$D(y; \hat{\mu}_0)$		$G^2(H_0)$

Nota. ¹ Grados de libertad.

Hipótesis sobre parámetros individuales:

Teniendo un modelo previamente seleccionado y ajustado, se utiliza el test de Wald para

evaluar la significancia de los parámetros individuales del modelo probabilístico (Wald 1944).

El contraste de hipótesis que maneja esta prueba es:

$$H_0 : \beta_j = 0 \text{ vs } H_0 : \beta_j \neq 0$$

El estadístico de la prueba es:

$$u_j = \frac{\hat{\beta} - \beta_j}{s.e(\beta)}$$

Bajo la hipótesis nula $H_0 : u_j = \frac{\hat{\beta}}{s.e(\beta)}$

Otro test equivalente a este se realiza utilizando $z_j = u_j^2$, se define un nivel de significancia α y se rechaza la hipótesis nula si $z_j > X_{1-\alpha}^2(1)$ (Madsen & Thyregod 2010).

Criterios de información:

En este estudio se utilizó el criterio de información de Akaike (AIC), que implementa la función de verosimilitud L y todos los parámetros p estimados. El AIC se define como (Montgomery et al. 2006):

$$AIC = 2Ln(L) + 2p$$

Coefficiente de determinación R^2 :

Es un estadístico estadística utilizado en los GLM y es análogo al R^2 de la regresión lineal múltiple, denotado como (Dobson & Barnett 2008):

$$R^2 = \frac{l(\bar{\pi}; y) - l(\hat{\pi}; y)}{l(\bar{\pi}; y)}$$

Donde $l(\bar{\pi}; y)$ es la desviación del modelo saturado, y $l(\hat{\pi}; y)$ es la desviación del modelo mínimo, que representa la mejora proporcional en la función de probabilidad logarítmica debido a los términos en el modelo de interés, en comparación con el modelo mínimo.

3. Metodología.

En el presente capítulo se presenta la metodología que se utilizó para realizar el estudio de análisis espacial y de supervivencia sobre los casos de cáncer de cuello uterino en el municipio de Santiago de Cali en los años 2006 y 2014. Inicialmente se detalla la base de datos que tiene el registro población de cáncer de Cali y también la base de datos otorgada para la investigación donde se define las variables y la zona de estudio, se hace un proceso de depuración de la base de datos y luego se realiza la geocodificación de las direcciones. La última sección corresponde al análisis estadístico el cual está conformado por el análisis descriptivo, la estimación de la intensidad, los test de completa aleatoriedad espacial y el modelo GLM que se implementa para realizar las estimaciones de supervivencia relativa con un enfoque de período.

3.1. Base de datos.

Esta investigación utilizó información suministrada por “El Registro Poblacional de Cáncer de Cali” (RPCC), el registro es uno de los más importantes del mundo y el único de base poblacional de tan larga trascendencia en Colombia (Muñoz et al. 2014). La base de datos del RPCC comprende datos demográficos, de tumor y de base diagnóstica de más de 100.000 casos nuevos de cáncer en el área urbana de Cali, este resultado se debe a la búsqueda activa y permanente de datos en todas las fuentes de información, protegiendo siempre la información personal obtenida, para que no vaya a ser divulgada, sin consentimiento de la persona. El RPCC cumple con los requerimientos de calidad del Centro Internacional de Investigaciones sobre el Cáncer (IARC) y es miembro activo de la Asociación Internacional de Registros de Cáncer (IACR)

la base de datos tiene información poblacional desde el año 1962, el registro ha sido utilizado para la planeación de actividades de prevención y control de cáncer en la región y en el país. En la base de datos filtrada solo se incluyó la información de casos nuevos de mujeres con cáncer de cuello uterino en el municipio de Santiago de Cali para el período 2006 - 2014, la cual cuenta con 1.945 observaciones.

3.2. Variables del estudio.

A continuación, se definirán las variables que se utilizaron para llevar a cabo el estudio y además la proporción de datos faltantes encontrados.

- **Edad:** en esta variable se encuentra la edad (en años cumplidos) en la cual se le fue diagnosticado el cáncer de cuello uterino. Se encontró un 0.15 % de información faltante.
- **Estrato socioeconómico:** nos indica en que clasificación se encuentra o encontraba la vivienda de la mujer en el momento que se le diagnóstico el cáncer de cuello uterino. Las clasificaciones son las siguientes (DANE 2018):
 1. Bajo-bajo
 2. Bajo
 3. Medio-bajo
 4. Medio
 5. Medio-alto
 6. Alto

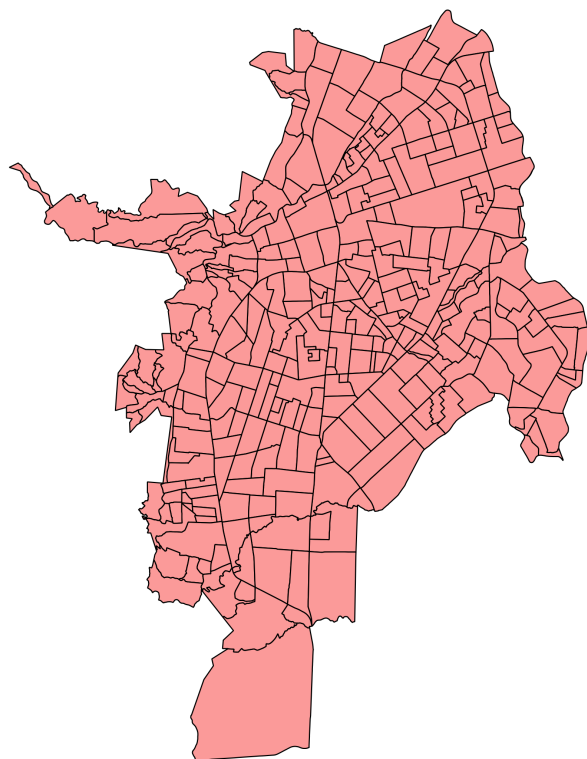
El 38.71 % de las observaciones no tenía información de esta variable, pero mediante la geocodificación de las direcciones se pudo encontrar el estrato socioeconómico asociado a esa ubicación y de esta manera reducir el porcentaje de datos faltantes.

- **Dirección:** En esta variable se encuentra la dirección del predio de cada caso de cáncer de cuello uterino. La variable es un identificador alfanumérico único asignado a un predio. Además se observó que todas las direcciones se encontraban estandarizadas, pero con abreviaturas diferentes, por lo que se procedió a homogeneizar las abreviaturas de todas las direcciones. La variable contaba con un 18.35 % de datos faltantes.
- **Fecha de diagnóstico:** Esta variable nos referencia el día, mes y año, en el cual se le fue encontrado o diagnosticado el cáncer de cuello uterino a la paciente.
- **Fecha de finalización del seguimiento :** Esta variable indica el día, mes y año, en que el paciente terminó su tratamiento o dejó de asistir al mismo.
- **Estado vital:** Esta variable representa el estado vital del paciente al final del seguimiento, es una variable dicotómica, porque solo toma dos valores “vivo” o “muerto”.

3.3. Zona de estudio.

El análisis espacial se realizó en el municipio de Santiago de Cali, capital del departamento del Valle del Cauca, Colombia. El municipio está subdividido en 22 comunas en la zona urbana y por 249 barrios en el perímetro de la ciudad.

Barrios de Cali.



Comunas de Cali.

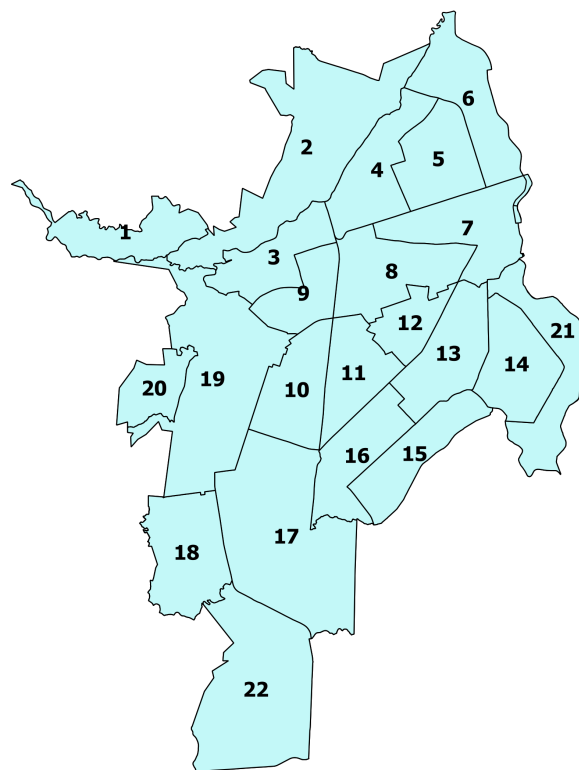


Figura 3-1.: Municipio de Santiago de Cali dividido por barrios

Figura 3-2.: Municipio de Santiago de Cali dividido por comunas.

En la Figura 3-1 se puede observar el mapa del municipio de Santiago de Cali subdividido por barrios y en la Figura 3-2 podemos observar la ciudad pero subdividida por sus comunas.

3.4. Depuración de la información.

Una vez se obtuvo el acceso a la base de datos, se procedió a identificar los individuos con información incompleta que no iban a entrar en el estudio, ya que en su información no se encontraba la variable edad, estrato socioeconómico y dirección, esta última variable es de gran importancia, porque es la que se debió transformar en coordenadas geográficas (longitud/latitud) para poder realizar nuestro estudio de análisis espacial.

Tabla **3-1.**: Distribución de frecuencias del cáncer de cuello uterino para las mujeres de Cali, período 2006-2014.

Año	Número de casos	Datos disponibles	
		No	Si
2006	249	58	191
2007	226	37	189
2008	214	21	193
2009	219	53	166
2010	213	35	178
2011	222	42	180
2012	181	26	155
2013	201	32	169
2014	220	53	167
Total	1945	357	1 588

En la Tabla **3-1** se puede observar para cada año del período 2006-2014 la cantidad de datos que habían originalmente (columna 2), la cantidad de observaciones sin la variable dirección registrada (columna 3) y la cantidad de casos de cáncer de cuello uterino que disponían de la dirección del individuo.

Por otra parte, en el modelamiento se excluyeron todos los casos de cáncer de cuello uterino que no contenían información de la edad, fecha de diagnóstico o última fecha de seguimiento. Se encontraron pocos individuos que no contenían este tipo de información.

Ante la presencia de muchos datos faltantes en la variable estrato socioeconómico (38,71 %), se realizó una estimación a partir del domicilio de la persona; mediante el programa de libre acceso QGIS se pudo hacer una extracción de la información de la capa “Estrato social por lado de manzana”, a través de los puntos o localización geográfica de los individuos con CCU, es decir, que a la paciente se le asignaba el estrato socioeconómico correspondiente al punto geográfico donde se encontrara.

3.5. Geocodificación.

Mediante el software de libre acceso QGIS (versión 2.18.19 - “Las Palmas”) y el complemento MMQGIS que sirve para manipular capas de mapas vectoriales, se procedió a realizar la geocodificación de las direcciones de los predios de cada observación. Este proceso usa como base, los servicios de WFS o servicio de fenómenos en la web, que es un estándar del Open Geospatial Consortium (OGC), que nos permite interactuar con los servicios de mapas en la web (WMS), este nos proporciona mapas de datos referenciados espacialmente, asociados al

sistema de coordenadas cartesianas MAGNA - SIRGAS origen Cali, por lo tanto se descargó la capa “Nomenclatura de ejes viales” y sobre esta se buscaron las direcciones, obteniendo como resultado su representación gráfica (archivo shapefile) y una tabla de atributos, con la cual se calcularon las coordenadas geográficas (longitud/latitud).

3.6. Análisis estadístico.

En esta sección se presenta el análisis estadístico que se llevó a cabo para realizar esta investigación, primero se muestra el análisis descriptivo de las variables, luego los test de completa aleatoriedad espacial, la estimación de la intensidad y por último la modelación estadística de supervivencia relativa de los casos de cáncer de cuello uterino en la ciudad de Santiago de Cali entre los años 2006 y 2014.

3.6.1. Análisis descriptivo.

Una vez definida las observaciones con los cuales se llevó a cabo el estudio, se procedió a realizar un análisis descriptivo para cada una de las variables. Para las variables cualitativas como lo es el estrato socioeconómico se obtuvieron tablas de frecuencia y gráficos de barras, y para las variables cuantitativas como la edad, tablas con estadísticas básicas conformadas por el mínimo, máximo, media, mediana, desviación estándar, 1er y 3er cuartil.

3.6.2. Análisis espacial.

Para observar el tipo de patrón espacial (aleatorio, regular o agregado) de los casos identificados en el período 2006-2014, se implementan los test de completa aleatoriedad espacial, alguna de estas pruebas se basan en crear cuadrantes en la zona de estudio, como el test de bondad de ajuste X^2 y otros test están basados en las distancias de entre los distintos puntos observados, Función G , F , K de Ripley y K de L . A continuación, se ilustra el uso de esta metodología con los datos de CCU del año 2007 y se comparará con los conjuntos de datos simulados con comportamiento espacial agregado, aleatorio y regular

Pruebas basadas en cuadrantes.

Para las pruebas basadas en cuadrantes, para el test de bondad de ajuste X^2 se dividió el dominio en 4 sub-divisiones (2 filas x 2 columnas) ya que esta prueba tiene como requisito tener por lo menos 5 eventos dentro de cada partición, por otro lado, como en el test de Montecarlo se realizó una sub-división del dominio en 100 partes (10 filas x 10 columnas) ya que esta prueba es más robusta ante la ausencia de eventos dentro de cada partición. Para las dos pruebas se tiene en cuenta las siguientes hipótesis:

$$H_0 : \text{El patrón es aleatorio} \quad H_a : \text{El patrón es agrupado}$$

Para poder realizar el contraste de hipótesis, se trabajó con un nivel de significancia del 5 %, lo cual implica que si el valor p del estadístico asociado a la prueba es menor que 0.05, se rechaza la hipótesis nula H_0 , concluyendo que los datos se distribuyen de forma agrupada. Como en la prueba chi-cuadrado se tiene como requisito que en cada división de la zona de estudio se debe tener por lo menos cinco (5) observaciones, en el caso de no cumplirse se utilizó el test de Monte Carlo, el cual resuelve el problema anterior. Frecuentemente las pruebas basadas en cuadrantes tienen problemas al realizar estas particiones, ya que se pierde el concepto de perímetro urbano de la ciudad, creando sesgo en la estimación de las pruebas (Cressie, 2015).

Función G.

Para la función de distancias del vecino más cercano o también llamada la distribución entre eventos, se debe tener en cuenta que si la función empírica $\hat{G}(r)$ crece rápidamente en distancias cortas, nos referencia que los eventos son regularmente espaciados, en cambio si crece lentamente en distancias cortas y después crece rápidamente en distancias largas, nos indica que los eventos presentan un patrón agregado; además los eventos aleatoriamente espaciados siguen el comportamiento de la distribución teórica $G(r) = 1 - e^{-\lambda\pi r^2}$.

A continuación se ilustra la metodología de la prueba con uno de los años de estudio, específicamente el año 2007. Se obtiene el siguiente gráfico:

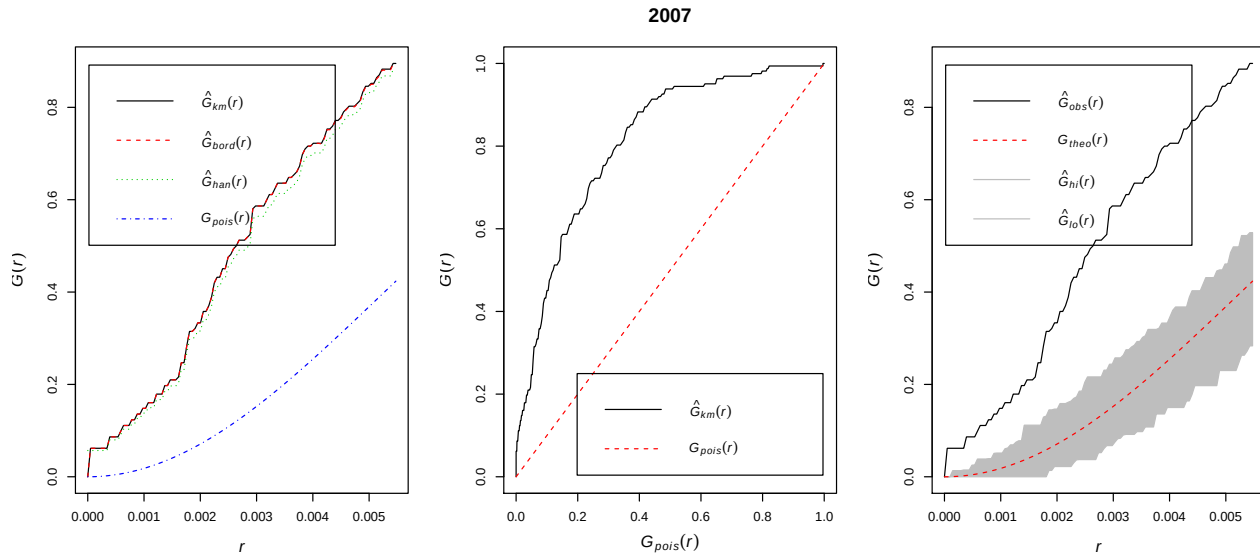


Figura 3-3.: Prueba de completa aleatoriedad G aplicada a los casos de cáncer de cuello uterino de la ciudad de Cali, 2007.

En la Figura 3-3 se puede detallar que la curva de distribución de los datos observados

$\widehat{G}_{obs}(r)$ se encuentra por encima de la curva de la distribución teórica $G_{theo}(r)$, además es similar al gráfico de los datos con patrón agregado de la Figura **2-4**, por lo tanto se puede concluir que las ubicaciones de los casos de cáncer de cuello uterino en la ciudad de Santiago de Cali para el año 2007 presentaron un patrón espacial agregado.

A continuación se especifica paso a paso como se obtuvieron los valores de la función teórica $G_{theo}(r)$ y empírica $\widehat{G}_{obs}(r)$, además se realizaran sus respectivas comparaciones. Estas funciones estan definidas como:

$$G_{theo}(r) = 1 - e^{-\lambda_{2007}\pi r^2}, \quad \widehat{G}_{obs}(r) = \frac{N^o\{r_i \leq d\}}{n_{2007}}$$

En donde r es la distancia entre los eventos, $\lambda_{2007} = 5572,3736$ y $n_{2007} = 162$. A continuación se observara los resultados en las primeras 10 unidades de distancia r .

Tabla **3-2.**: Resultados de función teórica y empírica de la Prueba G .

r	$\widehat{G}_{obs}(r)$	$G_{theo}(r)$
0.00000	0.05691	0.00000
0.00005	0.05691	0.00004
0.00010	0.05691	0.00018
0.00015	0.05691	0.00040
0.00020	0.05691	0.00070
0.00024	0.05691	0.00110
0.00029	0.05691	0.00158
0.00034	0.05691	0.00215
0.00039	0.07989	0.00281
0.00044	0.07989	0.00355

En la Tabla **3-2** se puede observar que la función empírica $\widehat{G}_{obs}(r)$ es mayor que la teórica $G_{theo}(r)$, evidenciando que la función empírica se encuentra por encima de función teórica y concluyendo que los eventos presentan un comportamiento espacial agregado.

Función F.

La función de espacio vacío $F(r)$, se tiene presente la mínima distancia entre punto-evento, donde el punto es un evento escogido aleatoriamente dentro de la región de estudio y consecutivamente se le mide la distancia con los demás eventos. Para esta prueba si la función empírica $\hat{F}(r)$ crece rápidamente en distancias cortas el patrón de los datos es regular, en cambio si crece lentamente en distancias cortas y después crece rápidamente, se dice que los datos presentan un patrón agregado.

Ahora se pasara a observar los resultados de la prueba aplicada a los datos de estudios en el año 2007:

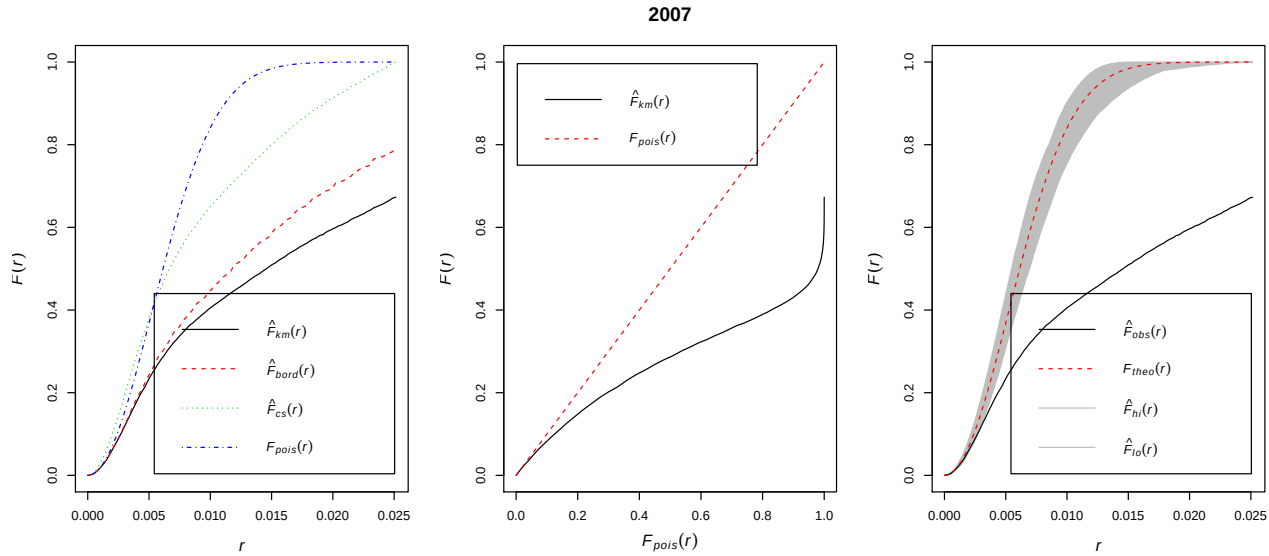


Figura 3-4.: Prueba de completa aleatoriedad F aplicada a los casos de cáncer de cuello uterino de la ciudad de Cali, 2007.

En la Figura 3-4 se puede apreciar que la curva de distribución de los datos observados $\hat{F}_{obs}(r)$ se encuentra por debajo de la curva de la distribución teórica $F_{theo}(r)$ y si vemos la Figura 2-5 se puede apreciar que es similar al gráfico aplicado a los datos con comportamiento espacial agregado, por lo tanto se puede concluir que las ubicaciones de los casos de cáncer de cuello uterino en la ciudad de Santiago de Cali para el año 2007 presentaron un patrón espacial agregado.

A continuación se detallara paso a paso como se obtuvieron los valores de la función teórica $F_{theo}(r)$ y empírica $\hat{F}_{obs}(r)$. Estas funciones estan denotadas como:

$$F_{theo}(r) = 1 - e^{-\lambda_{2007}\pi r^2}, \quad \hat{F}_{obs}(r) = \frac{N^o\{r_i \leq r\}}{m}$$

En donde r es la distancia entre los eventos, $\lambda_{2007} = 5572,3736$ y m son los puntos elegidos aleatoriamente. En la Tabla **3-3** se observan los resultados en las primeras 10 unidades de distancia r .

Tabla **3-3.**: Resultados de función teórica y empírica de la Prueba F .

r	$\hat{F}_{obs}(r)$	$F_{theo}(r)$
0.00000	0.00000	0.00000
0.00026	0.00222	0.00123
0.00052	0.00782	0.00493
0.00078	0.01765	0.01105
0.00104	0.03063	0.01956
0.00648	0.48851	0.53774
0.00674	0.50369	0.56595
0.00700	0.51969	0.59344
0.00726	0.53193	0.62013
0.00752	0.54670	0.64594

En la Tabla **3-3** se puede apreciar que la función empírica $\hat{F}_{obs}(r)$ es mayor que la teórica $F_{theo}(r)$ en las primeras distancias evaluadas, pero en las demás distancias $\hat{F}_{obs}(r) < F_{theo}(r)$, evidenciando que la función empírica se encuentra por debajo de la función teórica, por lo tanto, se puede concluir que los eventos presentan un comportamiento espacial agregado.

Función K de Ripley y Función K de L .

Función K , también conocida como función K de Ripley o función reducida del segundo momento (Cressie 1993), en la estimación de $\hat{K}_{obs}(r) = \frac{\sum \sum_{i \neq j} I_r(r_{ij})}{\hat{\lambda}_{2007}}$, donde $\hat{\lambda}_{2007} = \frac{n_{2007}}{A}$, se define $n_{2007}=162$ y $A= 619km^2$, por lo tanto $\hat{\lambda}_{2007} = 0,2617km^2$, comúnmente se compara con el valor verdadero de K para un patrón completamente aleatorio (Poisson), denotado como $K_{theo}(r) = \pi r^2$. Por otra lado, la Función L es una prueba basada en la Función K y propuesta por Besag (1977), la función empírica es denotada como $\hat{L}_{obs}(r) = \sqrt{\frac{K(r)}{\pi}}$ y bajo aleatoriedad se tiene que $L_{theo}(r) = r$. La raíz cuadrada tiene el efecto de estabilizar la varianza del estimador, por tanto, es el más adecuado para utilizarse en simulaciones y pruebas de hipótesis. Desviaciones entre la curva de la función empírica y la teórica, indican una agrupación espacial o regularidad de los eventos.

Ahora aplicando las pruebas a los datos de cáncer de cuello uterino en el año 2007 se obtiene lo siguiente:

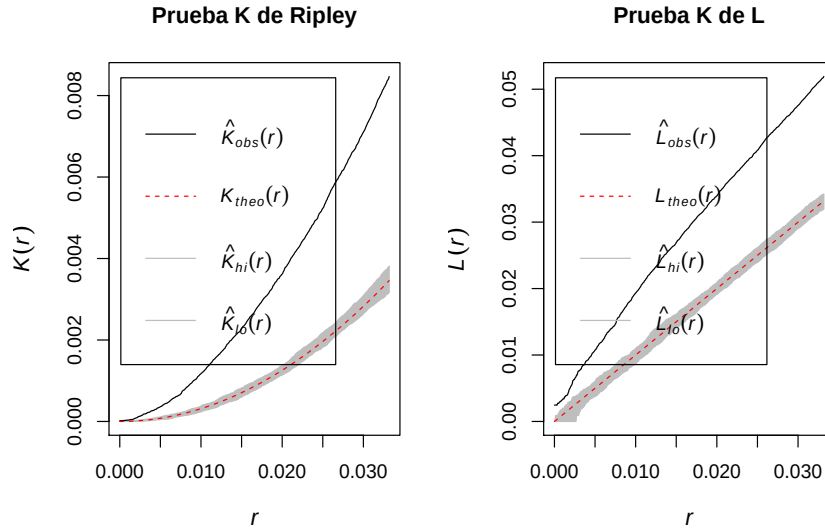


Figura 3-5.: Prueba de completa aleatoriedad K de Ripley (izquierda) y K de L (derecha) aplicada a los casos de cáncer de cuello uterino de la ciudad de Cali, 2007.

En la Figura 3-5 se puede ver que las 2 curvas de distribución de los datos observados $\hat{K}_{obs}(r)$ y $\hat{L}_{obs}(r)$ están sobre las curvas teóricas correspondientes $K_{theo}(r)$ y $L_{theo}(r)$, además si se observan las Figuras 2-6 y 2-7 se puede validar un comportamiento espacial agrupado para los casos de cáncer de cuello uterino en la ciudad de Cali en el año 2007.

En la Tabla 3-4 se muestran los valores obtenidos de la función empírica y teórica, de las pruebas K de Ripley y K de L .

Tabla 3-4.: Resultados de la función teórica y empírica, de la prueba K de Ripley y K de L .

r	$\hat{K}_{obs}(r)$	$K_{theo}(r)$	$\hat{L}_{obs}(r)$	$L_{theo}(r)$
0.00000	0.00002	0.00000	0.00246	0.00000
0.00006	0.00002	0.00000	0.00246	0.00006
0.00013	0.00002	0.00000	0.00246	0.00013
0.00019	0.00002	0.00000	0.00246	0.00019
0.00026	0.00002	0.00000	0.00246	0.00026
0.00227	0.00012	0.00002	0.00613	0.00227
0.00233	0.00012	0.00002	0.00619	0.00233
0.00240	0.00012	0.00002	0.00630	0.00240
0.00246	0.00012	0.00002	0.00630	0.00246
0.00253	0.00013	0.00002	0.00651	0.00253

En la Tabla 3-4 se puede apreciar que los valores de las funciones empíricas (columnas 2 y

4), presenta valores mayores ante las funciones teóricas (columnas 3 y 5) correspondientes, es decir que $\widehat{K}_{obs}(r_i) > K_{theo}(r_i)$ y $\widehat{L}_{obs}(r_i) > L_{theo}(r_i)$, para todas las distancias r_i evaluadas, reafirmando la existencia de una distribución espacial agregada para los casos de CCU en la ciudad de Cali para el año 2007.

Estimación de intensidad

Luego se procedió a estimar la función de intensidad para poder evaluar cuantos casos se reportan por unidad de área y con ello observar las zonas que presentan altas tasas de incidencia de cáncer de cuello uterino en el municipio de Santiago de Cali, en el período 2006-2014. Para realizar la estimación de la función intensidad, se llevó el conjunto de datos de patrones de puntos a un plano bidimensional o también conocido como superficie lisa en el espacio (Xie & Yan 2008), además, para realizar dicha estimación se utilizó la densidad Kernel, implementando una función kernel Gausiana, de la forma:

$$\widehat{\lambda} = \frac{1}{\nu(A)} \sum_{i=1}^n \frac{1}{h} K\left(\frac{s - s_i}{h}\right), \text{ donde } K = \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}}$$

Para la estimación de la función de intensidad se utilizó la optimización de los anchos de banda de la densidad kernel con el fin de que la intensidad estuviera lo más próxima posible a los casos observados de cáncer de cuello uterino en el municipio de Santiago de Cali. Para seleccionar el ancho de banda σ se hizo uso de las funciones “bw.diggle” (validación cruzada) y “bw.ppl” (validación cruzada de probabilidad) incluidas en el paquete estadístico “spatstat”. La función de la Validación Cruzada $M(\sigma)$ y la Verosimilitud de la validación cruzada $LCV(\sigma)$ están denotadas como:

$$M(\sigma) = \frac{MSE(\sigma)}{\lambda^2} - g(0) \quad ; \quad LCV(\sigma) = \sum_i \log(\widehat{\lambda}_i(x_i)) - \int_W \widehat{\lambda}(u) du$$

Para el $M(\sigma)$ podemos apreciar que su función dependen de la intensidad λ y a medida que λ empieza tomar valores altos, el valor del estimador $M(\sigma)$ se va reduciendo, cumpliendo con su objetivo de reducir el error cuadrático medio. Por otra lado, para el $LCV(\sigma)$ se puede observar que también depende de la intensidad λ , pero para valores altos de λ se obtiene valores altos para el estimador de verosimilitud de validación cruzada $LCV(\sigma)$, cumpliendo con el objetivo de maximizar este valor. A continuación, se presentarán los valores σ del ancho de banda para los 2 métodos.

Tabla 3-5.: Cálculo de ancho de banda (σ) para estimar el núcleo de la intensidad del proceso puntual de los casos de cáncer de cuello uterino de Cali, 2006-2014.

Año	Validación cruzada	Validación cruzada de probabilidad
2006	0,0042	0,0051
2007	0,0057	0,0055
2008	$3,24e^{-5}$	0,0065
2009	0,0057	0,0077
2010	$3,24e^{-5}$	0,0087
2011	$3,24e^{-5}$	0,0066
2012	0,0027	0,0059
2013	0,0039	0,0073
2014	0,0064	0,0079

En la Tabla 3-5 se puede observar los valores σ del ancho de banda utilizando el criterio de validación cruzada (columna izquierda) y validación cruzada de probabilidades (columna derecha). Para el criterio de validación cruzada se puede evidenciar que para los años 2008, 2010 y 2011 se presentaron valores muy bajos para el ancho de banda σ , por lo que no se generó en el mapa zonas de alta intensidad, el resultado fue un mapa con puntos, donde cada punto representaba el lugar en específico en donde sucedió el evento. Por otro lado, la selección del ancho de banda mediante el criterio de la verosimilitud de la validación cruzada resultó ser el más adecuado para el estudio, ya que para todos los años no se presentaron errores a la hora de la estimación de la intensidad y además los lugares señalados con alta intensidad no parecían ser irrazonables.

3.6.3. Modelación

En el modelo que se planteó (ver Sección 2.2.3.) los errores presentan una estructura del tipo Poisson, entonces recordando la función probabilidad de una distribución Poisson, denotada como:

$$P(Y = y) = f(y, \lambda) = e^{-\lambda} \frac{\lambda^y}{y!} = e^{y \log(\lambda) - \lambda - \log(y!)}$$

En donde se puede identificar las componentes con las que se verifica que se trata de un modelo perteneciente a la familia exponencial, por lo tanto $\theta = \log(\lambda)$, $b(\theta) = e^\theta$, $\phi = 1$, $a(\phi) = 1$ y $c(y, \theta) = -\log(y!)$

El logaritmo de la función de máxima verosimilitud esta representada como:

$$\ell(\lambda, y_i) = \sum_{i=1}^n \{y_i \lambda - \lambda - \ln(y_i!)\}$$

Entonces implementando una función de enlace logarítmica $\ln(\lambda)$ para el modelo de regresión lineal, dado k variables independientes X se obtiene el siguiente modelo lineal:

$$\ln(\lambda|X_1 = x_1, \dots, X_k = x_k) = \theta = \theta(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

Por lo tanto para los datos de esta investigación se tiene:

$$\ln(\lambda) = c + \alpha_1(año_1) + \alpha_2(año_2) + \alpha_3(año_3) + \alpha_4(año_4) + \alpha_5(año_5) + \beta_1(gedad_1) + \beta_2(gedad_2) + \beta_3(gedad_3) + \beta_4(gedad_4)$$

Simplificando un poco el modelo se obtiene:

$$\ln(\lambda) = c + \sum_{i=1}^5 \alpha_i año_i + \sum_{j=1}^4 \beta_j edad_j$$

Donde $\lambda = \mu_{año, edad} - d_{año, edad}^*$, $c = \ln(I_{año, edad} - \frac{d_{año, edad}}{2})$ representa el logaritmo de las personas en riesgo, que se introduce como compensación en el modelo, las primeras 5 variables representan los años de seguimiento ($año_1, año_2, año_3, año_4$ y $año_5$) y el resto representa los grupos de edad definidos ($edad_1, edad_2, edad_3$ y $edad_4$), donde el grupo de edad de referencia son los que tienen entre 15-44 años de edad ($edad_0$).

Entonces si $\mathbf{X} = (1, año_1, \dots, año_5, edad_1, \dots, edad_4)$ y $\beta = [c, \alpha_1, \dots, \alpha_5, \beta_1, \dots, \beta_4]^T$, se obtiene lo siguiente:

$$\ln(\lambda) = \mathbf{X}'\beta$$

Si despejamos la función de respuesta λ de nuestro modelo se obtiene:

$$\lambda = e^{\mathbf{X}'\beta}$$

Si hacemos $\eta = \mathbf{X}'\beta$, se obtiene $\lambda = e^\eta$. Ahora parametrizando el logaritmo de la función de verosimilitud en términos de λ se obtiene:

$$\ell(\beta, y_i) = \sum_{i=1}^n \{y_i(\mathbf{X}'_i \beta) - e^{\mathbf{X}'_i \beta} - \log(y_i!)\}$$

Estimación de parámetros y bondad de ajuste del modelo.

Ahora pasaremos a estimar el vector de parámetros $\hat{\beta}$, mediante el método iterativo de Newton Raphson:

$$\beta^{(t+1)} = \beta^{(t)} - [\ell''(\beta^{(t)})]^{-1} \ell'(\beta^{(t)})$$

Calculando las derivadas del logaritmo de la función de máxima verosimilitud, se obtiene:

$$\ell'(\beta) = \frac{\partial}{\partial \beta} \ell(\beta; y_i) = \sum_{i=1}^n (y_i - e^{x_i \beta}) x_i$$

$$\ell''(\beta) = \frac{\partial}{\partial \beta_j \partial \beta_k} \ell(\beta; y_i) = \left[- \sum_{i=1}^n (e^{x' \beta}) x_i x_j^{-1} \right]$$

Reemplazando los resultados de las derivadas encontradas obtenemos:

$$\beta^{(t+1)} = \beta^{(t)} - \left[\left[- \sum_{i=1}^n (e^{x' \beta^{(t)}}) x_i x_j \right]^{-1} \right]^{-1} \sum_{i=1}^n (y_i - e^{x_i \beta^{(t)}}) x_i$$

Realizando algunos cálculos se obtiene el algoritmo de Newton-Raphson para estimar los parámetros de un modelo con estructura de error Poisson:

$$\beta^{(t+1)} = \beta^{(t)} + \sum_{i=1}^n (e^{x' \beta^{(t)}}) x_i x_j \cdot \sum_{i=1}^n (y_i - e^{x_i \beta^{(t)}}) x_i$$

Pruebas de Bondad de ajuste para GLM.

- **Criterios de información:** recordemos que el AIC se define como (Montgomery et al. 2006):

$$AIC = 2Ln(L) + 2p$$

En donde $p = 9$, dado que representa la cantidad de parámetros en el modelo ajustado.

- **Test de Wald:** Recordemos que su estadístico de prueba bajo la hipótesis nula $H_0 : \beta_j = 0$, esta denotado como (Wald 1944):

$$u_j = \frac{\hat{\beta} - \beta_j}{s.e(\hat{\beta})}$$

De acuerdo a Madsen & Thyregod (2010) si $z_j = u_j^2 > X_{1-\alpha}^2(1)$, se rechaza H_0 , entonces si se define una confianza del 95 % ($\alpha = 0,05$), se rechaza la hipótesis nula si $z_j > 3,84$.

Para el análisis espacial se hizo uso del software gratis y de libre acceso RStudio (Versión 1.1.447). los paquetes que se utilizaron fueron “spatstat”, “raster”, “rasterVis” y “rgdal”, para obtener los tests de completa a aleatoriedad, graficar sus respectivas pruebas y observar los las bandas de ancho propuestas por los criterios de validación cruzada y validación cruzada de probabilidades. Para el análisis de supervivencia también se utilizó el mismo software pero además se implementó el paquete “periodR” versión 1.0-6 (Registry 2011), el cual ha sido utilizado para análisis de período basados en modelos para obtener estimaciones de supervivencia relativa y absoluta.

La función *period* del paquete periodR, requiere de una base de datos de probabilidades de mortalidad específica por edad y año de calendario, para ambos sexos de una población estándar, por lo que se utilizaron los datos del repositorio de datos del observatorio de la salud global (OMS, 2018) y se seleccionaron las tablas de los países de Alemania, Colombia y Jordania, para poder realizar comparaciones con respecto a la tasa de mortalidad general de cada país, es decir que Alemania en el 2017 presentó una tasa de mortalidad (muertes/1000 habitantes) de 11,7, para Colombia de 5,5 y para Jordania de 3,4 (CIA 2017).

Como las tablas de mortalidad descargadas se encontraban especificadas en grupos quinquenales de edad, se hizo uso del software de las Naciones Unidas para la medición de la mortalidad MORTPAK versión 4.3 (UN 2013), para desagregar estos grupos y poder obtener las probabilidades de mortalidad específica por edad individualmente.

3.6.4. Análisis de Supervivencia.

Lo primero que se debe hacer antes de utilizar la función “*period*” es adecuar la base de datos de tal forma que tenga la estructura requerida por la misma.

Tabla 3-6.: Estructura de base de datos de cáncer de cuello uterino de Cali para análisis de supervivencia utilizando la función *period* en el software R, período 2006-2014.

sex	diagage	dm	dy	fm	fy	vitstat	agr
2	19	7	2011	9	2017	1	0
2	20	12	2013	5	2018	1	0
2	21	56	2007	9	2017	1	2
2	22	12	2006	9	2017	1	0
2	22	48	2008	9	2008	2	1
2	22	50	2007	9	2017	1	1
...	...	...	...	...	...	...	...
2	66	12	2013	12	2013	2	3
2	96	8	2013	8	2013	2	4
2	97	6	2013	11	2015	2	4
2	98	8	2012	3	2013	1	4

Nota. en la tabla la variable *sex* representa el sexo del individuo (1:hombre, 2:mujer), *diagage* simboliza la edad al momento del diagnóstico, *dm* y *dy* representan el mes y año de diagnóstico del cáncer, *fm* y *fy* son el mes y año de finalización del seguimiento, *vitstat* el estado vital del individuo al final del seguimiento (1: vivo, 2: muerto) y *agr* el grupo de edad al que pertenece el individuo(0: 15-44, 1: 45-54, 2: 55-64, 3: 65-74, 4: 75 o más.)

Una vez estructurada la base de datos, se realizó un análisis de período convencional del período 2013-2017, para poder obtener las variables del GLM con estructura de error poisson (ver anexos), en el análisis se utilizó el método de Ederer II (Ederer & Heise 1959b) para calcular el número de muertes esperadas.

Luego se procedió a calcular la supervivencia relativa de las pacientes con cáncer de cuello uterino de la ciudad de Cali del período 2006-2014, basados en el grupo de edad y el año de seguimiento y se comparó con la supervivencia de una población de referencia, para este estudio se empleo la población de Colombia, también se utilizó la supervivencia de la población de Jordania y Alemania para realizar comparaciones.

Otra comparación que se realizó fue de supervivencia relativa a 5 años de las pacientes que se encontraban dentro y fuera de los puntos estimados con altos niveles de intensidad, se definió el cuantil 90 como regla de decisión, es decir, que se si un individuo se encontraba

en un punto de intensidad superior al cuantil 90 se dice que el individuo SI pertenece a un punto con alta intensidad estimada, en caso contrario, se dice que el individuo se encuentra fuera de los puntos de alta intensidad estimada.

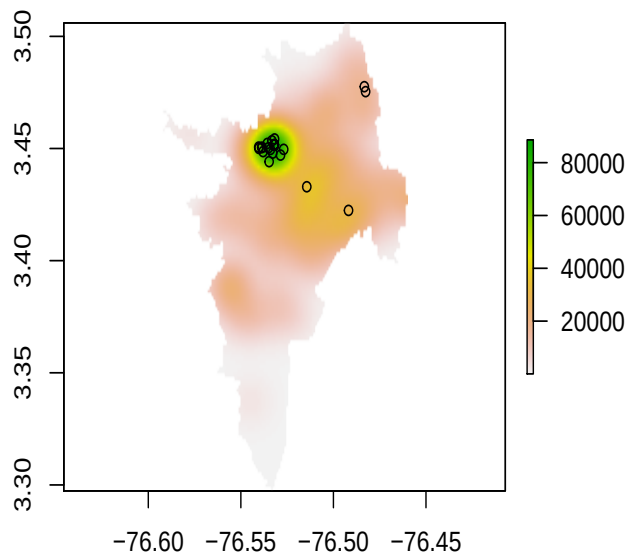
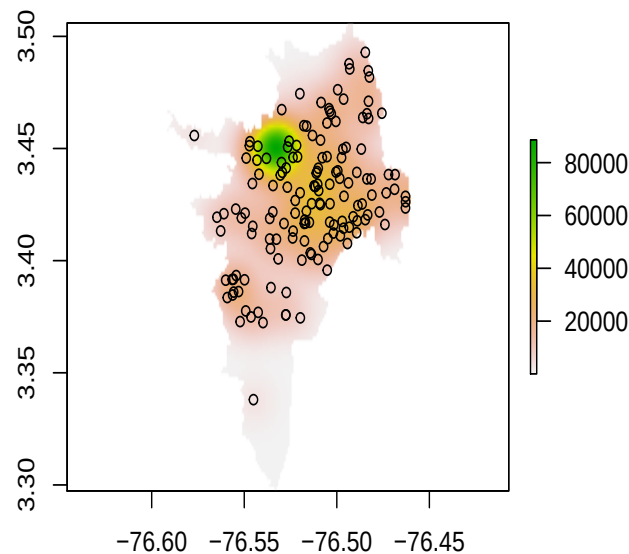
Grafico de intensidad de casos de CCU, 2008**Grafico de intensidad de casos de CCU, 2008**

Figura 3-6.: Comparación de la ubicación de las mujeres que se encuentra dentro y fuera de los puntos con altos niveles de intensidad estimada

En la Figura 3-6 se puede observar para el año 2008 los gráficos de la ubicaciones de las pacientes que fueron categorizadas como si estuvieran ubicadas en puntos de altos niveles de intensidad (izquierda) y las que se definieron como si se encontraran ubicadas por fuera de los puntos de altos niveles de intensidad(derecha).

4. Resultados.

En este capítulo se presentan los resultados de la investigación haciendo usos de los datos de cáncer de cuello uterino en el municipio de Santiago de Cali para el período 2006 - 2014. Primero se presenta el análisis descriptivo de las variables haciendo uso de tablas de frecuencia y gráficos de barras. En la segunda sección observaremos los resultados del análisis espacial que contiene una serie de pruebas de completa aleatoriedad espacial, la estimación del ancho de banda de la intensidad kernel y gráficas. Por último, se presentarán los resultados del modelo lineal generalizado propuesto, para modelar la supervivencia relativa con un enfoque de período.

4.1. Análisis descriptivo.

4.1.1. Estrato socioeconómico.

Para la variable Estrato socioeconómico se construyeron gráficos de barras para cada año de la investigación y además se obtuvo su respectiva tabla de frecuencias.

En el gráfico de barras del estrato socioeconómico de los casos de CCU (Figura 4-1) se puede observar la frecuencia absoluta de las categorías de la variable estrato socioeconómico para el período 2006-2014 en la ciudad de Santiago de Cali. Se puede apreciar que para todos los años de estudio se presentó una alta frecuencia en los niveles socioeconómicos 1, 2 y 3, sugiriendo que, el CCU es más habitual o común en personas de estrato 1, 2 y 3. De igual manera se puede observar que para los años 2007, 2011, 2012 y 2014 no se registraron mujeres con un nivel socioeconómico igual a 6. Por lo tanto, los resultados sugieren que los ingresos económicos en una mujer están relacionados con la adquisición de la enfermedad (CCU).

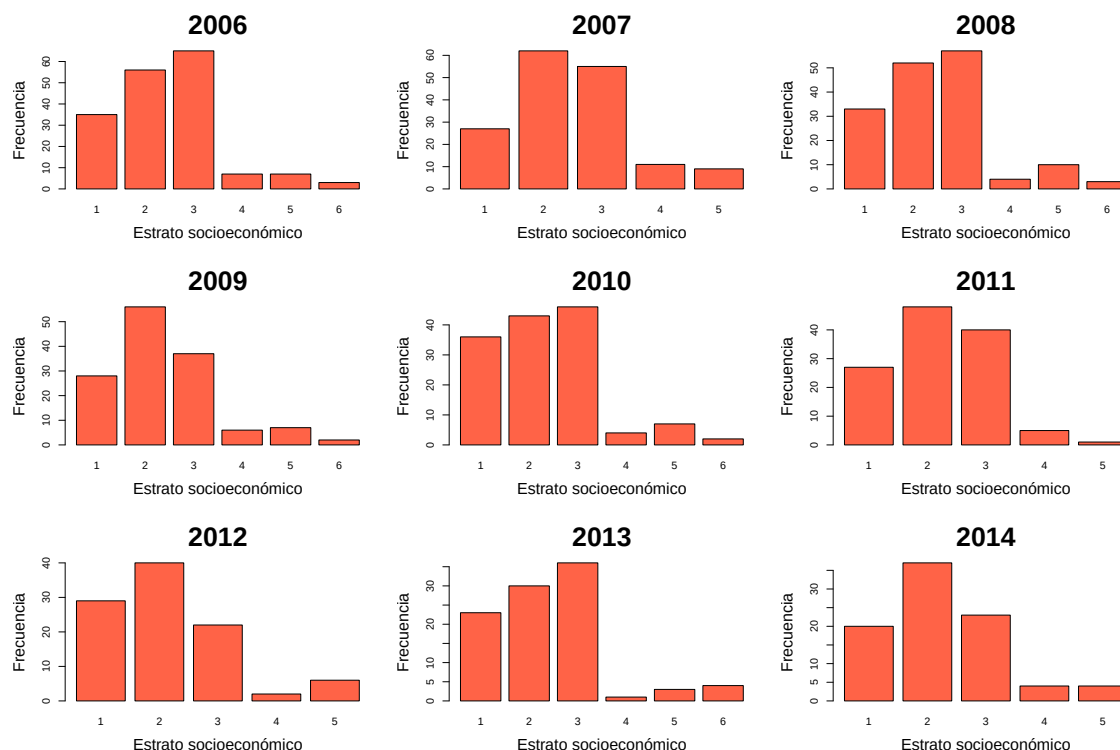


Figura 4-1.: Diagrama de barras del estrato socioeconómico de las mujeres con cáncer de cuello uterino de Cali, período 2006-2014

Tabla 4-1.: Distribución de frecuencias del estrato socioeconómico de las mujeres con cáncer de cuello uterino de Cali, período 2006-2014.

Nivel de estrato	2006	2007	2008	2009	2010	2011	2012	2013	2014	Total
1	35	27	33	28	36	27	29	23	20	258
2	56	62	52	56	43	48	40	30	37	424
3	65	56	57	37	46	40	22	36	23	382
4	7	11	4	7	4	5	2	1	4	45
5	7	9	10	7	7	1	6	3	4	54
6	3	0	3	2	2	0	0	4	0	14
Total	173	165	159	137	138	121	99	97	88	1177

Nota. Los niveles de estrato presentados están basados en la clasificación definida por el Departamento Administrativo Nacional de Estadística (DANE), donde el estrato 1 es el más bajo y estrato 6 el más alto.

En la Tabla 4-1 se presenta la distribución de los casos de CCU en la ciudad de Cali, por estrato socioeconómico y por año de análisis. Se puede apreciar que el cáncer de

cuello uterino se presentó más frecuentemente en las mujeres pertenecientes al estrato 1, 2 y 3, con un porcentaje de mujeres que viven en el estrato 1, 2 y 3 entre el 71,09 % y 91,75 %.

Para total de la población de estudio, el 36 % eran mujeres de estrato 2, el 32,5 % mujeres de estrato 3, el 21,9 % mujeres de estrato 1 y para los demás estratos socioeconómicos se presentó una frecuencia menor al 4,5 %. Se puede evidenciar que la cantidad de mujeres registradas con CCU fue disminuyendo a medida que avanzaron los años (Figura 4-2).

4.1.2. Edad.

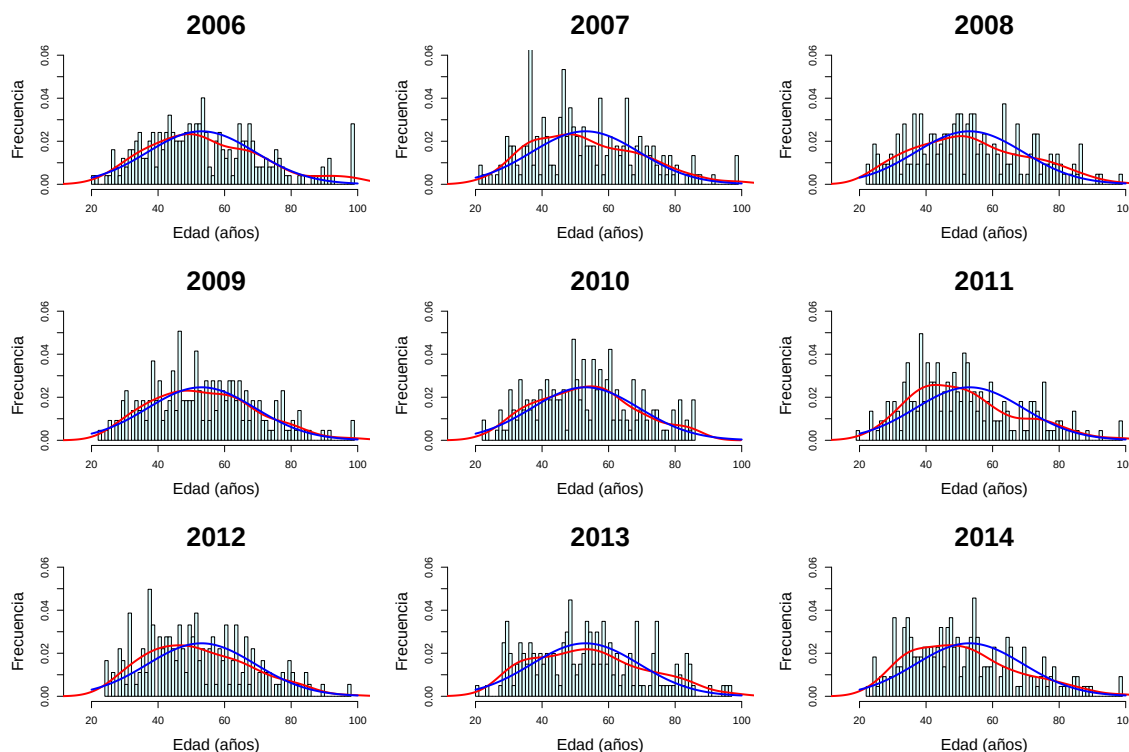


Figura 4-2.: Gráfico de barras y función de densidad de los casos de cáncer de cuello uterino en Cali, período 2006-2014

En la Figura 4-2 podemos observar el gráfico de barras de la variable edad para cada uno de los años del período 2006-2014; Se añadió en todos los gráficos la función de densidad normal de la edad sin tener en cuenta el año de diagnóstico (Línea Azul) y la función de densidad observada para cada año (Línea Roja), esto se hizo con el fin de poder comparar la variable con los distintos años analizados en este estudio. Se puede observar un desplazamiento hacia la izquierda de la curva de la función de densidad observada, a medida que van transcurriendo los años, esto quiere decir que con el pasar de los años fue disminuyendo la edad promedio en que se le diagnosticaba a una mujer, cáncer de cuello uterino.

Tabla 4-2.: Estadísticas básicas de la edad de las mujeres con cáncer de cuello uterino de Cali, período 2006-2014.

Año	Mínimo	Q1	Mediana	Media	Q3	Máximo	Desv.	CV
2006	20	42,00	52	53,30	65,00	93	15,82	0,30
2007	21	39,25	50	52,60	65,00	92	15,53	0,30
2008	22	40,00	51	53,20	65,00	93	16,68	0,31
2009	22	41,00	52	53,28	63,25	92	14,97	0,28
2010	22	42,00	53	53,38	63,00	86	14,90	0,28
2011	19	39,00	49	50,85	60,00	93	15,22	0,30
2012	24	41,00	51	52,17	63,00	90	14,88	0,28
2013	20	41,00	54	54,52	66,75	97	16,95	0,31
2014	22	38,00	48	50,65	60,75	90	15,53	0,31

En la Tabla 4-2 se presentan algunas estadísticas básicas de la variable edad en los diferentes años del período 2006-2014. El caso de CCU con la menor edad fue reportado en el año 2011, específicamente con una edad de 19 años; se pudo ver una reducción en el promedio de la edad, a medida que avanzaba el tiempo, aunque para el 2013 este promedio aumenta. La edad máxima reportada fue de 97 años en el año 2013.

Se puede observar que para el año 2006, el 50 % de la población que tenía Cáncer de Cuello uterino se encontraba en una edad entre los 42 y 65 años; Para el año 2007 la mitad de la población con la enfermedad se encontraban entre los 39,25 y 65 años; En el año 2011 la mitad de las mujeres con cáncer de cuellos uterino se encontraban en un rango de edad de entre los 39 y 60 años.

En la Figura 4-3 se muestra la gráfica de la distribución espacial de los casos de Cáncer de Cuello Uterino (puntos rojos) en la ciudad de Santiago de Cali, para los diferentes años del período 2006-2014. Se evidencia que en la parte sur de la ciudad se registraron muy pocos casos de CCU durante el período de estudio, los puntos parecen estar más concentrados en la parte central y oriental de Cali.

4.2. Análisis Espacial.

4.2.1. Distribución espacial

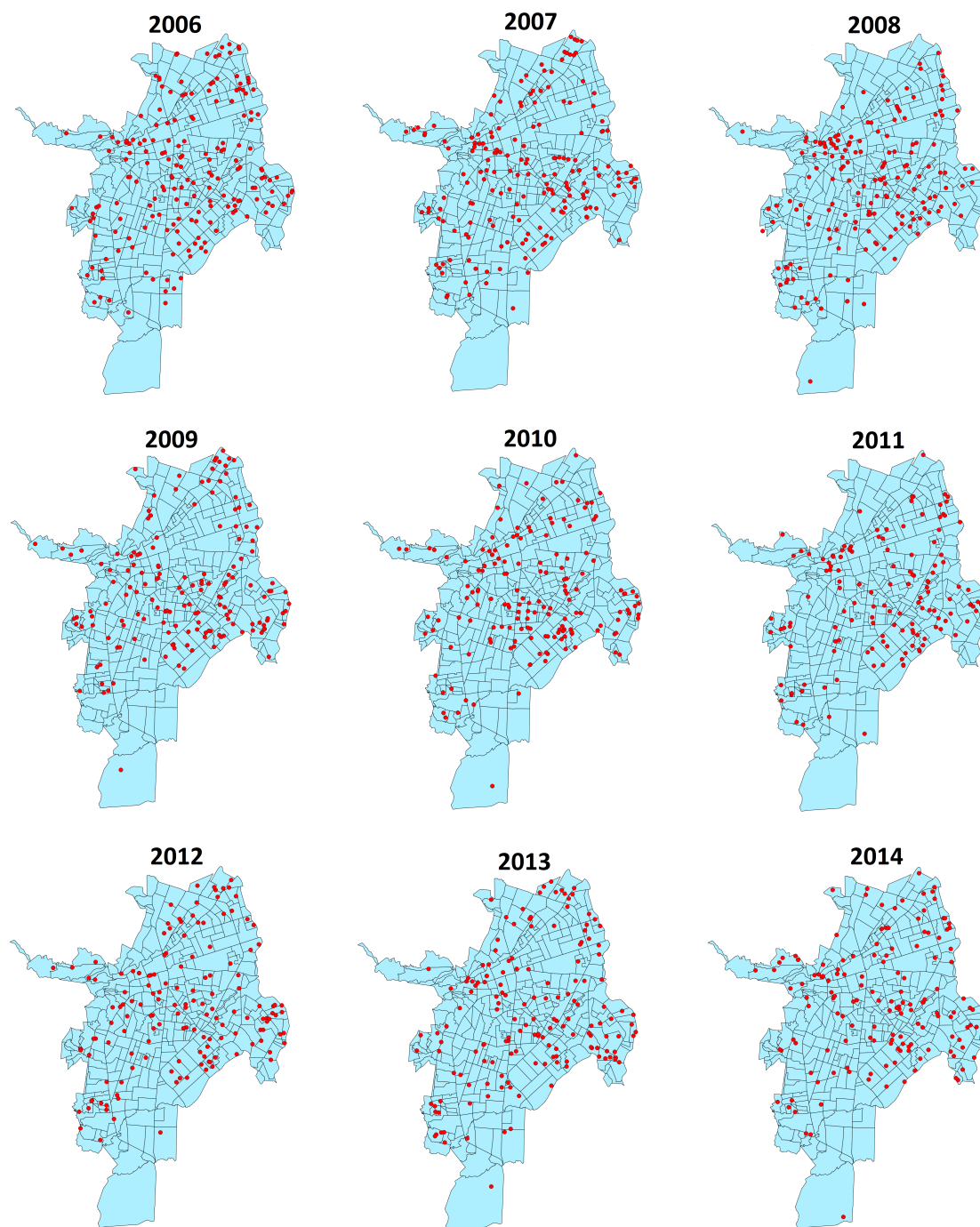


Figura 4-3.: Distribución espacial de las mujeres con cáncer de cuello uterino en Cali, período 2006-2014.

Uno de los problemas con el gráfico de puntos es la superposición y el tamaño de los puntos, la superposición de puntos puede ocurrir por varias razones, una de ellas puede ser que en un mismo sitio donde habitan muchas mujeres se reporten múltiples casos de CCU entonces puede haber varias mujeres con la enfermedad con la misma localización geográfica, por consiguiente se registran con el mismo domicilio. Entonces si tenemos estos puntos superpuestos, no podremos tener una idea clara de dónde están ocurriendo estos casos realmente o peor aún, no sabremos cuales son los sitios más frecuentes de la enfermedad, por esta razón se recurre a los gráficos de Intensidad, para obtener un comportamiento real de la enfermedad (CCU).

4.2.2. Pruebas de completa aleatoriedad.

A continuación, se presentan los resultados de las pruebas de completa aleatoriedad espacial aplicados a los datos de Cáncer de Cuello Uterino de la ciudad de Cali en el período 2006-2014, para cumplir con el objetivo de determinar el tipo de patrón espacial presente en cada año.

Para las pruebas basadas en cuadrantes se calculó el valor del estadístico y valor p asociado al contraste de hipótesis, donde en la H_0 se plantea que los n puntos son distribuidos uniformemente e independientes en todo el dominio de D . Para las pruebas basadas en distancias se comparó la función teórica de aleatoriedad espacial con sus respectivas bandas de confianza, contra la función empírica, los resultados se pueden observar en el Anexo A.

En la Tabla 4-3 se puede observar el resumen de los resultados obtenidos aplicando las pruebas de completa aleatoriedad para los diferentes años de estudio, estos resultados fueron constantes e idénticos, indicando la existencia de un patrón agregado en la distribución de los puntos, es decir, que los eventos se encuentran de forma agrupada, por lo que se podrían identificar zonas con alta frecuencia de casos de CCU en la ciudad de Cali, para cada año del período 2006-2014.

Tabla 4-3.: Resultado de pruebas de completa aleatoriedad aplicado a los datos de cáncer de cuello uterino de Cali, período 2006-2014.

Método	Prueba	Resultado	Valor p
Basado en cuadrantes	Test X^2	Agregado	$2,2e^{-16}$
	Test de Montecarlo	Agregado	0,001
Basado en distancias	Función G	Agregado	—
	Función F	Agregado	—
	Función K de Ripley	Agregado	—
	Función K de L	Agregado	—

4.2.3. Estimación de la intensidad

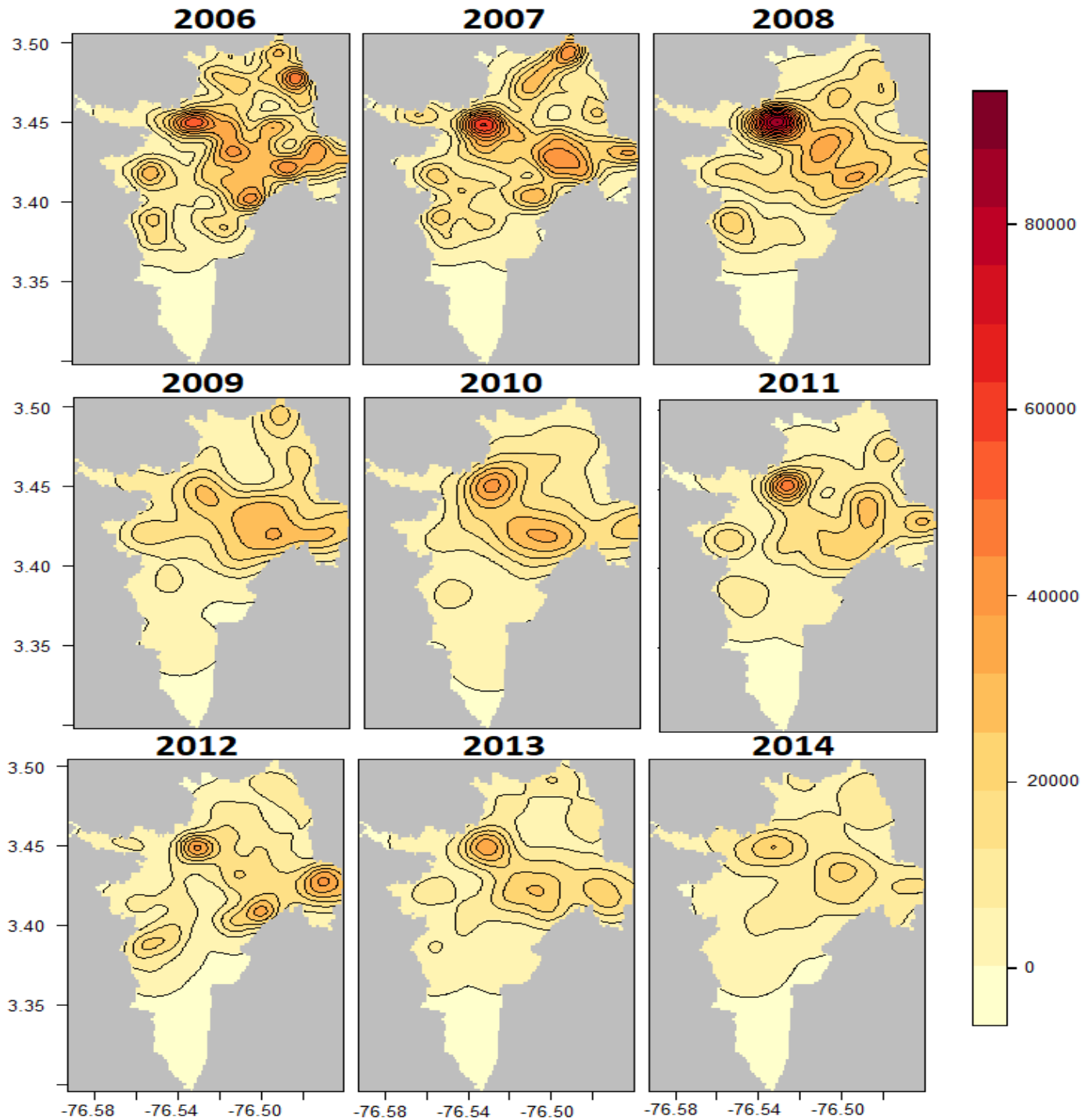


Figura 4-4.: Gráfico intensidad de los casos de cáncer de cuello uterino en Santiago de Cali, 2006-2014.

En la Figura 4-4 se puede observar los gráficos de intensidad estimada bajo el Criterio de la Verosimilitud de la Validación Cruzada para los casos de CCU en el municipio de Santiago de Cali para cada uno de los años del período 2006-2011.

Para el Año **2006** se presentaron altos niveles de intensidad en las comunas 3, 6, 8, 13, 15, y 20, para ser más concretos en los barrios: San Pedro, Ciudad Córdoba, poblado II, Petecuy, la Rivera I, el Peñón, Siloé y Belisario Caicedo; En el año **2007** se presentaron zonas altas de alta intensidad en la comuna 3, 6, 13, 15, 18, 20 y 21, específicamente den los barrios de San Pedro, el Calvario, San Juan Basco, San Cayetano, Santa Rosa, los Guaduales, Sector Puente del Comercio, Ciudadela Floraría, Eduardo Santos, Doce de Octubre, San Pedro Claver, Julio Rincon, Pondaje, Comuneros II, Poblado I, Vergel, LLeras Restrepo, Ciudad Talanga, Compartir, Brisas de Mayo, Belisario Caicedo, Lourdes, Prados del Sur y Mario Correa Rengifo; para el Año **2008** en las comunas 3, 5, 11, 13 ,14, 15 y 18, siendo más puntuales en los barrios San Pedro, Alfonso Barberena, 20 de Julio, Maracaibo, La Independencia, el Sena, Lourdes y Alto Napoles; para el año **2009** en las comunas 3, 6, 13, 14 y 15, particularmente en los barrios de San Pedro, el Vergel, Comuneros I, Manuela Beltrán, Asturias y Ciudadela Floralia; Para el Año **2010** las comunas 3, 13, 15 y 16, en los barrios de San Pedro, Santa Rosa, el Calvario, Vergel, Antonio Nariño, Poblado I, Comuneros I y el Retiro; para el año **2011** las comunas 3, 6, 13, 15, 20 y 21, singularmente en los barrios de San Pedro, San Nicolás, el Hoyo, Ulpiano Lloreda, Villa del Lago, Villa Blanca, Poblado I, Comuneros I, Promociones Populares B, Ciudadela del Rio - CVC y Jorge Eliécer Gaitán; para el año **2012** las comunas 3, 14, 15, 18 y 21, justamente en los barrios de San Pedro, Puertas del Sol, Promociones Populares B, Ciudad Talanga,el Retiro, el Vallado, Lourdes, Los Portales - Nuevo Rey y Alferez Real; para el año **2013** en las comunas 3, 11 y 21, en los barrios de San Pedro, León XIII, los Conquistadores, Primavera y Camilo Decepez; por último, el año **2014** en las comunas 3, 12, respectivamente en los barrios San Pedro, Asturias y el Rodeo.

Se puede apreciar que para la comuna 3 en todo el período 2006 - 2014 se presentaron zonas con altos niveles de intensidad de los casos de cáncer de cuello uterino; También fue muy frecuente el reporte de los casos de cáncer de cuello uterino en las comunas 6,13,15 y 21.

4.3. Análisis de Supervivencia

En la Tabla **B-1** se puede observar las variables que se utilizaron para crear el GLM con distribución de los errores Poisson, para Colombia, en donde la variable *año* representa el año de seguimiento, *gedad* el grupo de edad al que pertenece el individuo, *dstar* es denotado como $d_{y,agr}^*$ y *at_risk* es el logaritmo del número de pacientes efectivamente enfermos (con CCU), que es el término que se introduce como compensación (ver en Anexos).

$$\ln(\lambda) = c + (-1,79)año_1 + (-2,20)año_2 + (-2,72)año_3 + (-3,06)año_4 + (-4,44)año_5 + (0,50)gedad_1 + (0,39)gedad_2 + (0,54)gedad_3 + (0,68)gedad_4$$

Una vez estimado los parámetros del modelo lineal generalizado a través de los métodos iterativos de Newton Raphson, donde en 6 interacciones se obtuvo la convergencia, se realizaron las pruebas de Bondad de Ajuste del modelo ajustado.

Tabla **4-4.**: Análisis de Desvianza para modelo lineal generalizado, datos de pacientes con cáncer de cuello uterino de Cali, período 2006-2014.

	g.l ¹	Desvianza	g.l ¹ residuales	Desvianza residual	$Pr(>)$	
Modelo bajo H_0	—	—	25	3227.01	—	
Año de seguimiento (año)	5	3193.40	20	33.61	0.0000	***
Grupo de edad (gedad)	4	21.59	16	12.02	0.0002	***

Nota. ¹ Grados de libertad. $AIC = 132,29$

En la Tabla **4-4** se puede verificar el ajuste del modelo con todas las variables independientes incluidas; la diferencia entre la desvianza del modelo bajo H_0 y la desvianza residual, presenta el comportamiento de una distribución chi-cuadrado con 9 grados de libertad, entonces como $1 - Pr(\chi_{(9)}^2 \leq 3214,99) = 0$ se rechaza la hipótesis nula H_0 con un nivel de confianza del 95 %, indicando un buen ajuste del modelo.

Tabla 4-5.: Prueba Wald para modelo lineal generalizado, datos de pacientes con cáncer de cuello uterino de Cali, período 2006-2014.

	Estimado	Std. Error	Valor z	$Pr(> z)$	
$año_1$	-1,79	0,15	-12,28	0,0000	***
$año_2$	-2,20	0,16	-13,94	0,0000	***
$año_3$	-2,72	0,18	-14,94	0,0000	***
$año_4$	-3,06	0,21	-14,62	0,0000	***
$año_5$	-4,44	0,43	-10,18	0,0000	***
$edad_1$	0,50	0,17	3,01	0,0026	**
$edad_2$	0,39	0,19	2,05	0,0404	*
$edad_3$	0,54	0,23	2,36	0,0182	*
$edad_4$	0,68	0,26	2,67	0,0076	**

Notas: *** $p < 0$. ** $p < 0,001$. * $p < 0,01$.

En la Tabla 4-5 podemos observar la estimación de los coeficientes para cada año de seguimiento y grupo de edad, la respectiva desviación estándar de los errores, el valor del estadístico del test Wald y la probabilidad de obtener un valor como el observado o peor; se puede apreciar que todos los coeficientes de las variables son significativas para el modelo. Como los coeficientes del modelo pueden interpretarse como logaritmos de exceso de mortalidad relativa y lo que deseamos obtener son las estimaciones de período modeladas de la supervivencia relativa de seguimiento específico para cada grupo de edad, se pasó a calcular los coeficientes de la siguiente forma:

- Para la categoría de referencia definida como el grupo de menor edad (15-44) se tiene:

$$r_{y,agr=0} = \exp(-\exp(\alpha_y))$$

- para las demás categorías o grupos de edad se tiene:

$$r_{y,agr=agr} = \exp(-\exp(\alpha_y + \beta_{agr}))$$

Donde α_y es un vector que contiene los coeficientes de las variables $y1$, $y1$, $y3$, $y4$ y $y5$; y β_{agr} las variables $agr1$, $agr2$, $agr3$ y $agr4$

Luego se pasó a calcular las estimaciones de la supervivencia relativa con seguimiento acumulado a 5 años, pero además de utilizar las tablas de probabilidad de mortalidad de Colombia para realizar los cálculos de la supervivencia relativa, se evaluarán las tablas de los países de Alemania y Jordania.

Tabla 4-6.: Estimación de la supervivencia relativa a 5 años(%), de pacientes con cáncer de cuello uterino de Cali, período 2006-2014.

	Jordania	Colombia	Alemania
TGM ¹	3,4	5,5	11,7
Grupos de edad			
15-44	71,83	69,72	71,41
45-54	58,80	55,65	58,14
55-64	58,21	56,05	56,36
65-74	49,49	47,09	45,42
75+	51,38	40,86	42,74
General ²	61,58	60,09	59,19

Nota. ¹ Tasa general de mortalidad por cada 1.000 habitantes para el año 2017.

²Supervivencia relativa a 5 años, sin tener en cuenta los grupos de edad.

En la Tabla 4-6 se muestra la tasa general de mortalidad y la estimación de la supervivencia relativa con seguimiento acumulado a 5 años para los diferentes grupos de edad, en los diferentes países a estudiar. Podemos observar para Colombia que la supervivencia relativa de 5 años para el cáncer de cuello uterino fue de 67.01 % para las mujeres que se encontraban en la edad de 15-44 años y de 37.62 % para los mayores de 75 años; la supervivencia relativa de 5 años para el cáncer de cuello uterino basado en la supervivencia esperada de la población de Jordania es de 66.95 % y 45.19 % para los grupos de edad 15-44 y 75 años o más, respectivamente; y por último, utilizando la supervivencia esperada de la población de Alemana, se encontró que la supervivencia relativa de 5 años en el menor grupo de edad fue mayor que en los otros países, específicamente del 66.78 %, pero para las mujeres mayores de 75 años fue de 36.61 %, siendo menor entre los demás países.

Si observamos la tasa general de mortalidad para Jordania es de 3,4 (muertes/1000 habitantes), siendo la menor ante los demás países, por lo tanto se esperaba que la probabilidad de supervivencia relativa a 5 años del CCU fuera mayor y se evidenciara en los 4 grupos de edad definidos.

En la Figura 4-5 podemos observar las curvas de supervivencia relativa de 5 años de Colombia para los diferentes grupos de edad, la curva del grupo de menor edad (15-44 años) se encuentra por encima de las demás, esto significa que los pacientes con cáncer de cuello uterino en un seguimiento de 5 años tiene más probabilidades de sobrevivir, ya que su diagnóstico es más precoz, es decir que el carcinoma es detectado a una temprana edad. Por otro lado, se pudo apreciar que los pacientes con una edad entre los 55-64 años presentaron mayores porcentajes de supervivencia relativa en comparación con pacientes entre los 45-54

años (ver Anexo C para las curvas de supervivencia relativa para Alemania y Jordania).

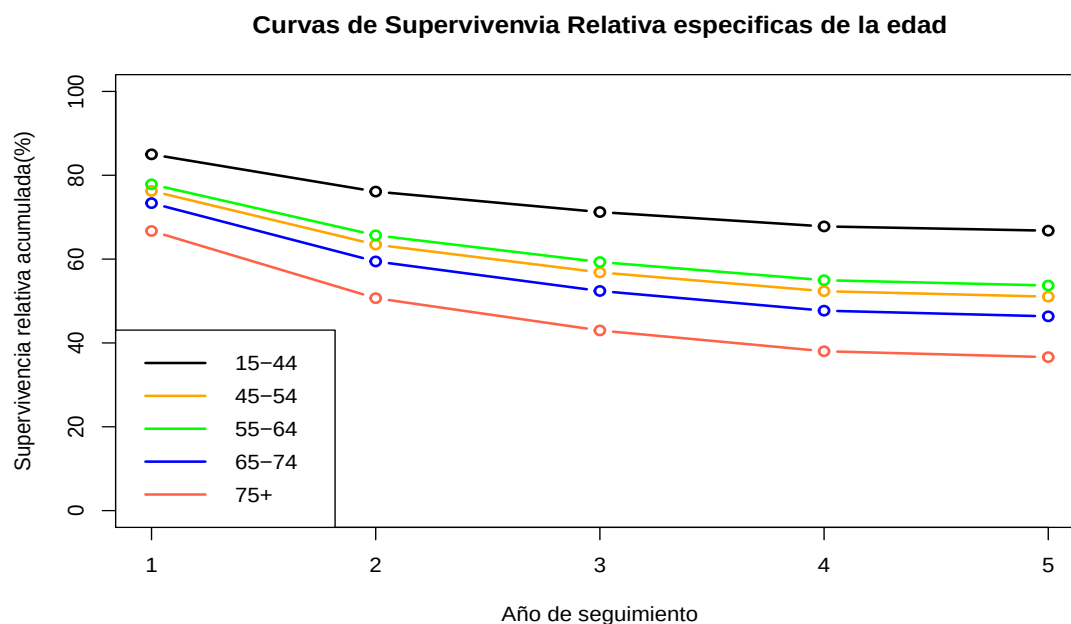


Figura 4-5.: Curvas de supervivencia relativa de 5 años, por grupos de edad, Colombia.

A continuación se procede a comparar la supervivencia relativa a 5 años (enfoque de período) de las mujeres que se encuentran ubicadas dentro de los puntos con altos niveles de intensidad y las que se encuentran fuera de estos puntos. Una vez se identificó los pacientes dentro y fuera de los puntos de alta intensidad, se procedió a ajustar el modelo y estimar respectivos coeficientes, para los 2 grupos definidos (ver en Anexos D).

Posteriormente se procedió a estimar la supervivencia relativa con seguimiento acumulado a 5 años implementado los coeficientes estimados y se obtuvo lo siguiente:

Tabla 4-7.: Estimación de la supervivencia relativa a 5 años de pacientes ubicados dentro y fuera de los puntos de alta intensidad.

Grupos de edad	Mujeres en altos puntos de intensidad	
	SI(%)	NO(%)
15-44	57,17	66,33
45-54	38,86	53,55
55-64	25,25	54,83
65-74	19,77	42,26
75+	16,49	38,45

En la Tabla 4-7 se observa que la probabilidad de supervivencia relativa con seguimiento acumulado de las mujeres que se encuentran dentro de los puntos con altos niveles de intensidad (columna 1, SI) es menor con respecto a las que se encuentran por fuera de estos puntos de alta intensidad, en todos los grupos de edad definidos. Además la probabilidad de supervivencia relativa para las mujeres que se encontraban por fuera de estos puntos con altos niveles de intensidad resultaron tener un comportamiento igual que en la probabilidad hallada para toda la población de estudio, su diferencia por grupos varía entre 5,2 y 7,29 unidades de porcentaje.

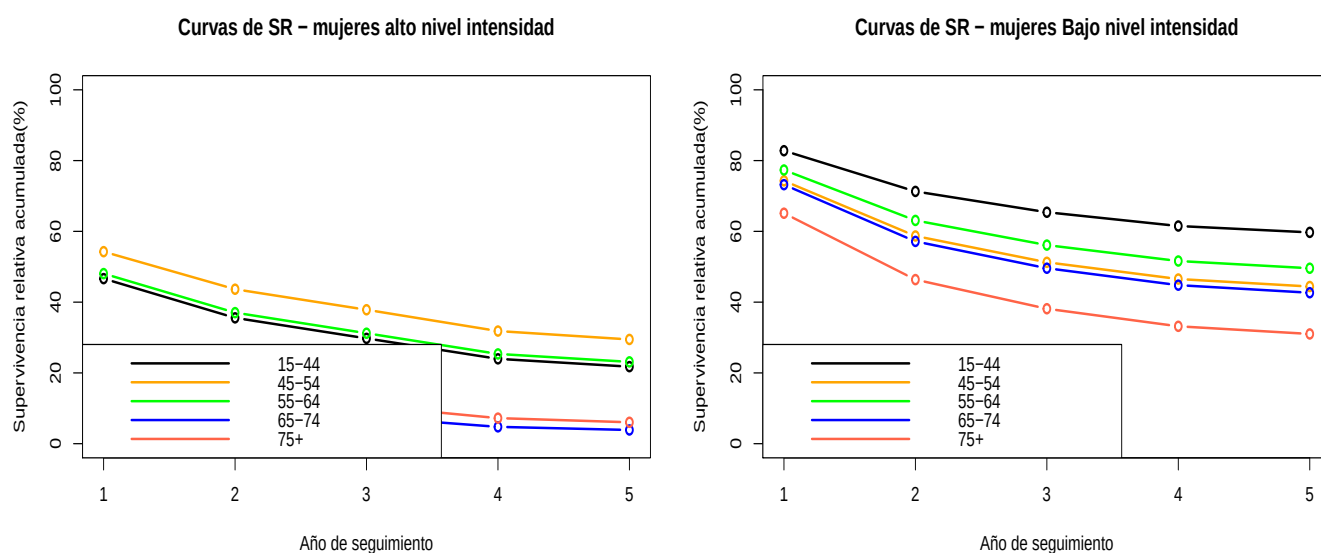


Figura 4-6.: Comparación de curvas de supervivencia relativa de 5 años fuera y dentro de los puntos de alta intensidad, grupos de edad, Cali-Colombia.

En la Figura 4-6 podemos observar las curvas de supervivencia relativas con seguimiento acumulado por grupos de edad para las mujeres que se encuentran dentro (izquierda) y fuera (derecha) de los puntos con altos niveles de intensidad. Para las mujeres ubicada dentro de los puntos de alta intensidad se puede apreciar que la curva que representa las mujeres con edad entre los 45-54 años presentan la mayor probabilidad de supervivencia ante los demás grupos de edad y para los grupos de edad de 66-74 y 75+ años se observar comportamientos similares, y sus curvas se encuentran muy distantes de las otras, se podría decir a partir de la edad de los 66 años la probabilidad de supervivencia es menor de 6.05 %. Para las curvas de supervivencia especificadas por grupos de edad para las mujeres que se encontraban fuera de los puntos de alta intensidad se puede apreciar un comportamiento similar al de la Figura 4-5, donde se calculó la supervivencia de toda la población de estudio.

5. Conclusiones y recomendaciones

5.1. Conclusiones

Esta investigación utilizó los datos del Registro Poblacional de Cáncer de Cali (RPCC) en el período 2006 - 2014, disponiendo de 1.588 datos para el análisis espacial y 1.148 datos para realizar el modelamiento estadístico. En donde el 90,4 % de las mujeres del estudio pertenecían a los estratos 1, 2 y 3; la edad mínima que se reportó en cada uno de los años de estudio fue entre los 19 y 24 años.

Desde el punto de vista espacial, se encontró un patrón agregado, es decir, la formación espacial de grupos de individuos, para cada uno de los años, entonces mediante la aplicación de las pruebas de completa aleatoriedad, los resultados fueron equivalentes, validando un comportamiento espacial agregado. Posteriormente se detallaron las zonas en donde realmente se encuentran concentrados los casos de cáncer de cuello uterino.

Pardo et al. (2017) en el “Atlas de mortalidad por cáncer en Colombia” habla de que las zonas con altos riesgo de presentar la enfermedad de CCU, era frecuente en lugares cercanos a ríos navegable, por lo que con los resultados obtenidos en la investigación se pudo dar veracidad de lo predicho. En la Figura 5-1 podemos observar mapa del Plan de Ordenamiento Territorial (POT:2014) para la protección civil ante el riesgo de inundaciones por lluvias (Zonas rojas), es decir que estas zonas son muy frecuentes que se inundan cuando llueve fuertemente y en la Figura 5-2 podemos observar el POT para el riesgo de inundaciones por los ríos, las zonas rojas representan posibles lugares a donde puede llegar el desbordamiento del río.

El Humedal de Charco Azul y el Humedal del Pondaje están ubicados en la comuna 13, estos humedales fueron creados de manera natural y modificados por la CVC para proteger la población de continuas inundaciones. En la actualidad alrededor de los Humedales se han establecido muchos asentamientos ilegales, debido a la escasez de recursos económicos, desplazamientos forzados y otras causas. Por otra parte, estas personas también están siendo expuestas a contaminantes ambientales, siendo estos sectores vulnerables. Según Cali en Cifras para la comuna 13 el estrato con mayor frecuencia fue el estrato 2.

En la Ciudadela de Floraríá ubicado en la comuna 6, se presenta una zona con alta de

intensidad de casos de CCU, específicamente en la parte oriente de la comuna, en el Jarillón de Florarí, que es un dique que protege a la población en caso de una posible creciente del río Cauca, en la actualidad todo el Jarillon ha sido invadido por Asentamientos Humanos de Desarrollo Incompleto (AHDI) (Vásquez 2001). Estos hogares se caracterizan por carencia de servicios públicos, baja vinculación de empleo formal, situaciones de pobreza crítica y alto riesgo de catástrofe (Mostacilla & Elena 2018). Según Cali en Cifras para el 2015 la comuna 6 el estrato con mayor frecuencia era el estrato 2.

En la Comuna 21 se presentó altos niveles de intensidad para los casos de CCU, la comuna no posee hospitales ni clínicas, cuenta con un puesto y un centro de salud, es decir, que posee aproximadamente dos establecimientos que prestan atención médica, por cada 100.000 habitantes ,por lo que genera complicaciones en la calidad de vida y la salud de los habitantes; además la comuna tiene como moda el estrato 1; es importante resaltar que la comuna se encuentra en el borde del Río Cauca. Para la comuna 15 con respecto a la estratificación, se tiene que la comuna no presenta lados de manzana de estrato 4, 5 y 6, por lo que se concentra en los estratos 1, 2 y 3, siendo el estrato 1 el mas común; es una de las comunas mas densamente pobladas de la ciudad y cuenta con 11 asentamientos.

Inundación Pluvial en Cali.

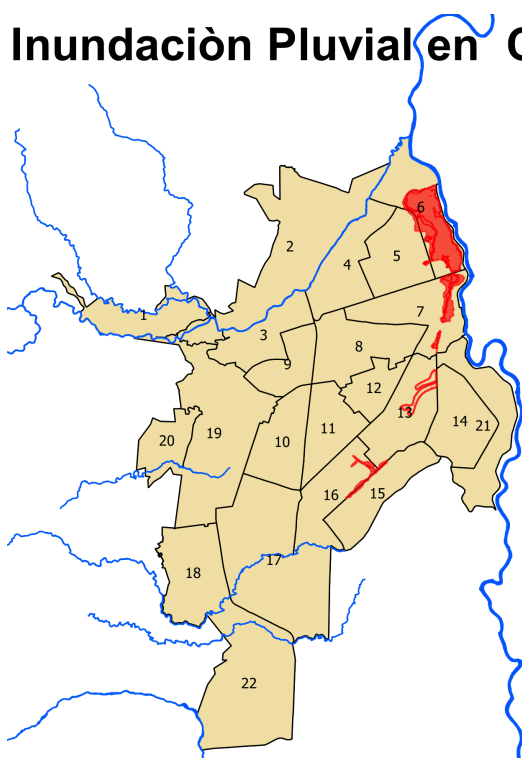


Figura 5-1.: Inundacion Pluvial (Lluvia) Cali, POT-2014.

Inundación Fluvial en Cali.

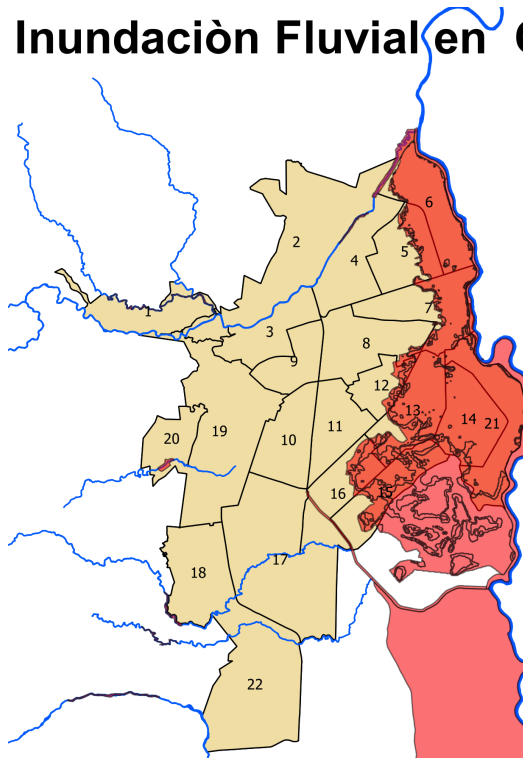


Figura 5-2.: Inundacion Fluvial (Rio) Cali, POT-2014.

En la comuna 3 fue el lugar donde se presentó una mayor intensidad de los casos de

cáncer de cuello uterino, según Cali en cifras el estrato mas frecuente es el estrato 3, pero de acuerdo con la información contemplada por la alcaldía de Cali (2008), se puede observar que la pirámide poblacional de esta comuna tiene una base estrecha, esto quiere decir, que la proporción de personas menores a 15 años es relativamente menor que la proporción de individuos entre 15 y 30 años, este proceso de “envejecimiento” relativo en la población podría ser la causa de tan alta intensidad de los casos de CCU en la comuna 3.

Para en análisis de supervivencia se utilizó la supervivencia relativa para poder controlar todas las diferentes causas muertas que no fuera causada por el cáncer de cuello uterino, es decir que tenía en cuenta el tiempo desde que a la persona se le diagnosticaba CCU hasta que finalizara el seguimiento, es decir el tiempo hasta que la paciente muere, se cura o deja de asistir al control y no se tiene más registro de ella. Adecuando la base de datos fue posible aplicar la función *period* del paquete “*periodR*”, para extraer los vectores de valores que fueron utilizados para la creación del GLM.

Utilizando una función de enlace no estándar implementada por Holleczeck y Brenner (2014) en el artículo llamado “*Model based period analysis of absolute and relative survival with R: Data preparation, model fitting and derivation of survival estimates*”, se obtuvieron estimaciones significativas de los coeficientes del modelo, los cuales representan el logaritmo de exceso de mortalidad y mediante una transformación de estos, se pudo realizar adecuadas interpretaciones de la supervivencia relativa a 5 años con un enfoque de período. Se puede apreciar que la supervivencia relativa a 5 años de las mujeres con cáncer de cuello uterino de Cali, presentó las probabilidades esperadas, con respecto al orden de la tasa general de mortalidad de los países de estudio, es decir, que entre más alta fuera la tasa general de mortalidad, menor iba a ser la probabilidad de supervivencia relativa estimada.

Mediante el test de Wald se verifico la significancia individual de los coeficientes estimados en el modelo ajustado y el análisis de devianza permitió comprobar la importancia de todas las variables independientes para la explicar el comportamiento de la variable de respuesta, el resultado de las 2 pruebas evidenciaron el buen ajuste del modelo.

Por ultimo en la comparación de probabilidades de las pacientes que se encontraban dentro y fuera de los puntos con altos niveles de intensidad, se pudo apreciar que las mujeres además de pertenecer a un punto de alta intensidad, también presentaron una menor probabilidad de supervivencia, comparado con las mujeres que se encontraban en lugares definidos como puntos de baja intensidad.

5.2. Limitaciones del estudio

Para el análisis espacial se presentaron limitaciones en el período de tiempo a estudiar, ya que la dirección de los individuos con cáncer de cuello uterino en el RPCC se encontraban en un proceso de tabulación o actualización de la información, cabe resaltar, que las direcciones se encontraban estandarizadas, por lo tanto, solo se escogió la información confiable de los pacientes con cáncer de cuello uterino hasta el año 2014, aun así se excluyó de la base de datos el 18.35 % (357) de los individuos, por ausencia del domicilio.

5.3. Recomendaciones

Para el análisis de supervivencia se sigue realizarse una modelación de la supervivencia relativa a 5 años para las diferentes categorías de estrato socioeconómico, ya que no se han hecho para la ciudad de Cali. Además para futuros trabajos se podría plantear calcular la supervivencia neta.

Para entrar más a fondo en el comportamiento espacial del cáncer de cuello uterino, se debería indagar en las zonas señaladas con alta intensidad de los eventos, para poder buscar aspectos ambientales, sociodemográficos o físicos de las personas, que se encuentren relacionadas con el fenómeno (cáncer de cuello uterino), por ejemplo la comuna 3, la cual está rodeada por iglesias y edificaciones públicas y/o privadas, como Teatro Jorge Isaacs, Palacio Nacional, iglesia la Ermita, etc.

Para obtener el Error Estándar (EE) de las estimaciones de la supervivencia relativa a 5 años se sugiere utilizar el método Delta (Chang 1968), su cálculo se realizaría con el objetivo de presentar un intervalo de confianza, dentro del cual (con un nivel de significancia α definido) se va a encontrar el parámetro poblacional, en vez de un solo valor estimado.

Ajustar un modelo que incluya la correlación espacial de los errores, porque aunque todos los casos de las pacientes están relacionados, puede existir, que las mujeres que tengan ubicaciones cercanas, se encuentren más relacionadas que las que están distantes.

A. Anexo: Pruebas de Completa aleatoriedad

■ Función G

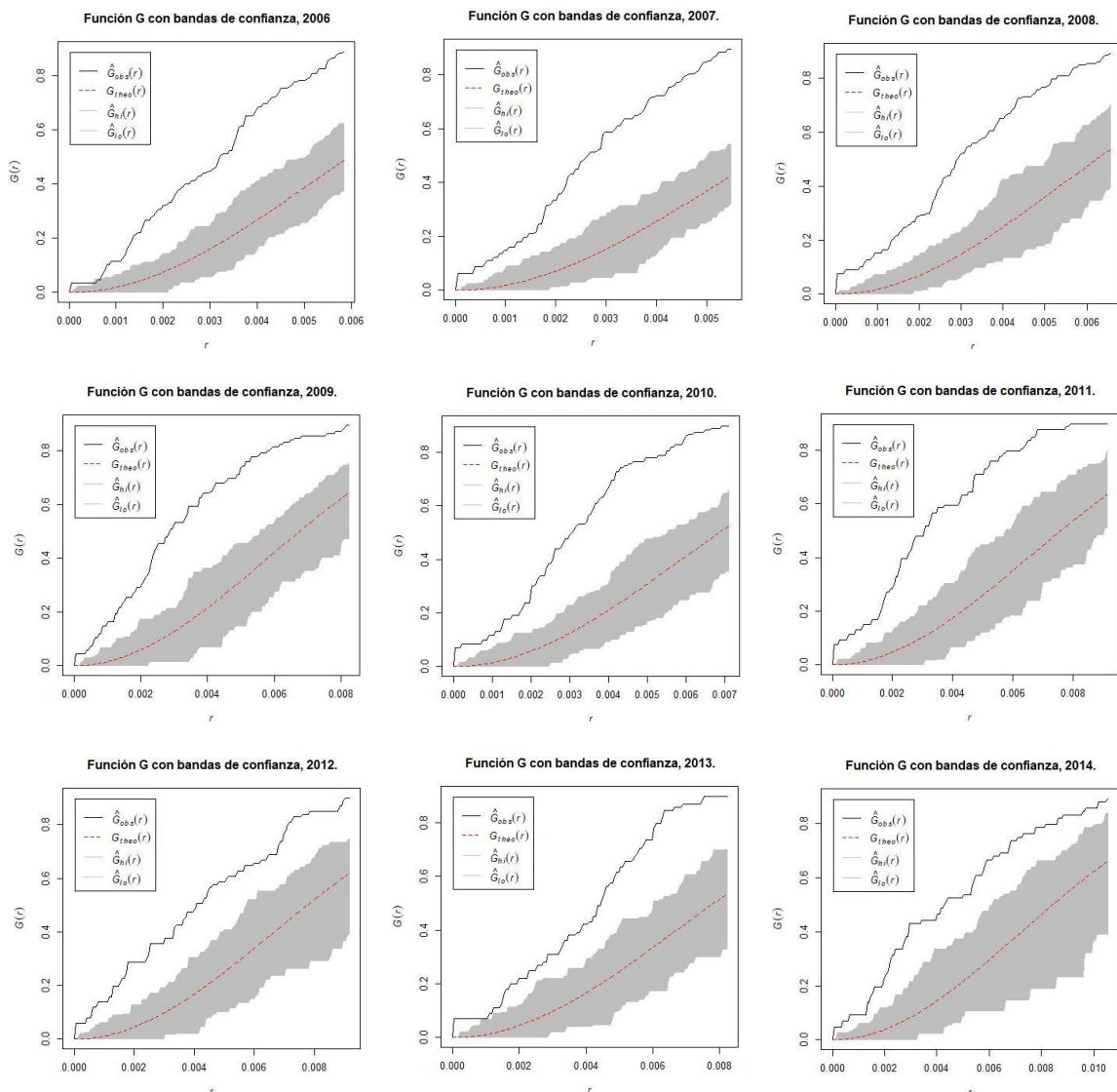


Figura A-1.: Función G para los casos de CCU en Santiago de Cali, 2006-2014.

■ Función F .

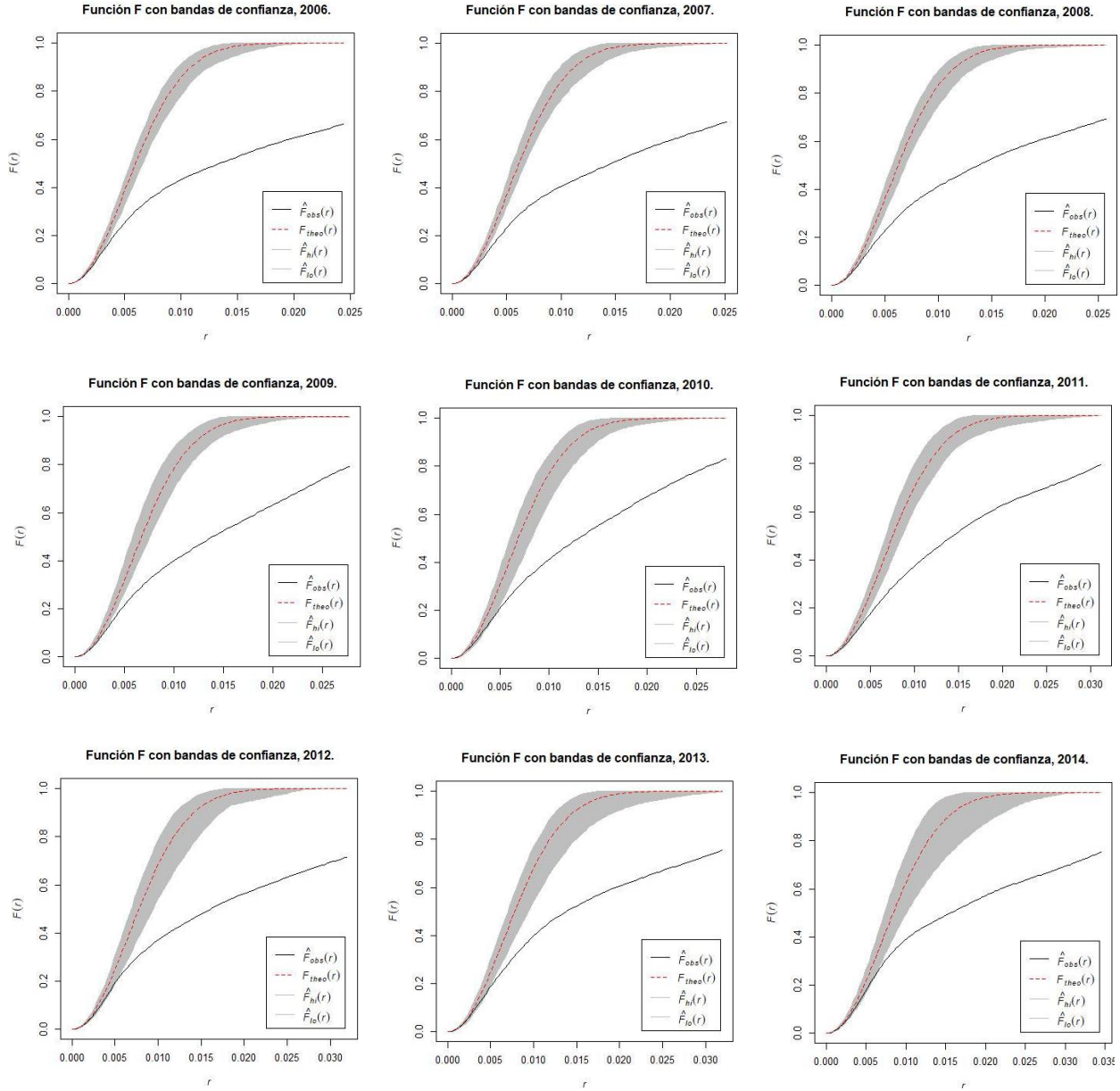


Figura A-2.: Función F para los casos de CCU en Santiago de Cali, 2006-2014.

■ Función K .

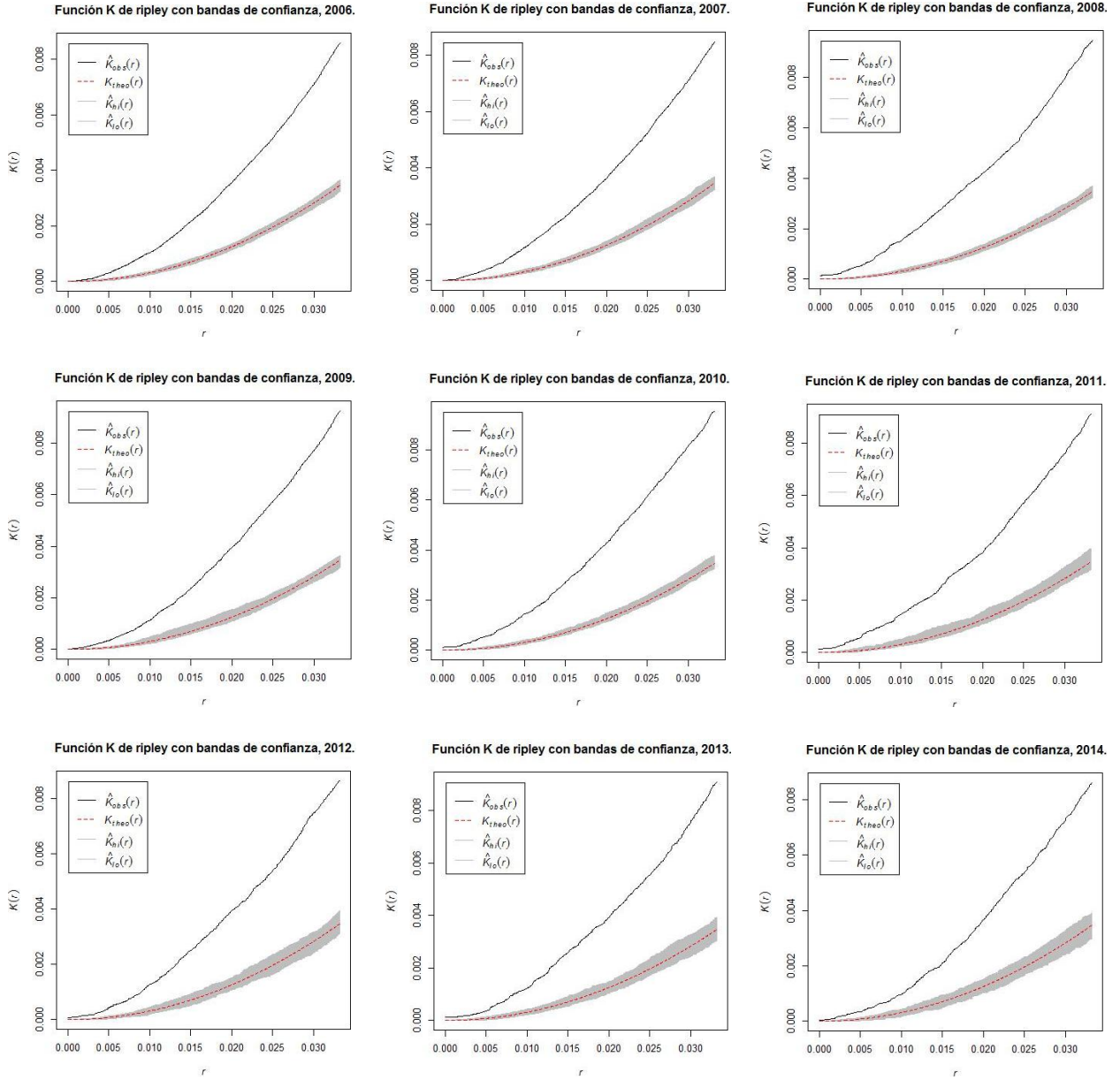


Figura A-3.: Función K para los casos de CCU en Santiago de Cali, 2006-2014.

■ Función K de L .

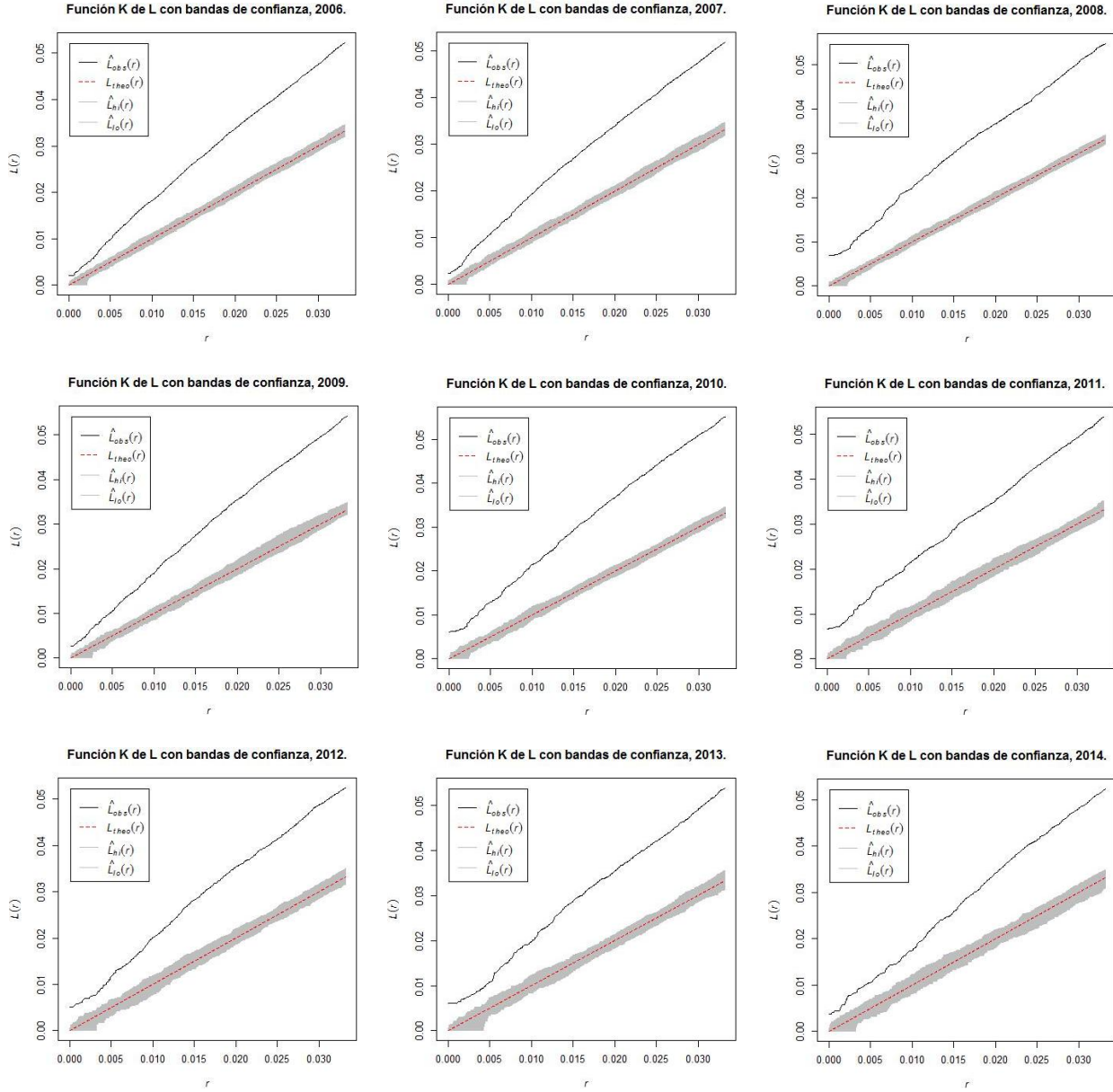


Figura A-4.: Función K de L para los casos de CCU en Santiago de Cali, 2006-2014.

B. Anexo: Tablas de variables para modelo lineal generalizado.

Tabla B-1.: Variables de modelo lineal generalizado para análisis de supervivencia de los casos de cáncer de cuello uterino de Cali-Colombia, período 2006-2014.

año	gedad	obs	dstar	at_risk
1	0	23	0,22	5,06
2	0	23	0,29	5,23
3	0	19	0,36	5,37
4	0	8	0,38	5,37
5	0	2	0,40	5,36
1	1	26	0,36	4,58
2	1	20	0,44	4,72
3	1	15	0,55	4,85
4	1	11	0,61	4,86
5	1	5	0,67	4,86
1	2	21	0,61	4,30
2	2	11	0,71	4,36
3	2	8	0,95	4,56
4	2	12	1,07	4,60
5	2	2	1,18	4,60
1	3	18	0,96	3,82
2	3	10	0,99	3,75
3	3	4	1,15	3,78
4	3	3	1,37	3,87
5	3	3	1,49	3,86
1	4	20	3,61	3,84
2	4	14	2,97	3,61
3	4	7	2,99	3,50
4	4	8	2,37	3,42
5	4	5	2,25	3,35

Tabla **B-2.**: Variables para modelo lineal generalizado para Jordania

año	gedad	obs	dstar	at_risk	
1	1	0	23	0,22	5,06
2	2	0	23	0,28	5,23
3	3	0	19	0,35	5,37
4	4	0	8	0,37	5,37
5	5	0	2	0,39	5,36
6	1	1	26	0,37	4,58
7	2	1	20	0,46	4,72
8	3	1	15	0,59	4,85
9	4	1	11	0,66	4,86
10	5	1	5	0,74	4,86
11	1	2	21	0,70	4,30
12	2	2	11	0,83	4,36
13	3	2	8	1,13	4,56
14	4	2	12	1,29	4,60
15	5	2	2	1,45	4,60
16	1	3	18	1,27	3,82
17	2	3	10	1,34	3,75
18	3	3	4	1,58	3,78
19	4	3	3	1,90	3,87
20	5	3	3	2,10	3,86
21	1	4	20	5,71	3,84
22	2	4	14	4,54	3,61
23	3	4	7	4,44	3,50
24	4	4	8	3,74	3,42
25	5	4	5	3,59	3,35

Tabla **B-3.**: Variables para modelo lineal generalizado para Alemania

año	gedad	obs	dstar	at_risk	
1	1	0	23	0,10	5,06
2	2	0	23	0,14	5,23
3	3	0	19	0,18	5,37
4	4	0	8	0,20	5,37
5	5	0	2	0,22	5,36
6	1	1	26	0,23	4,58
7	2	1	20	0,29	4,72
8	3	1	15	0,36	4,85
9	4	1	11	0,40	4,86
10	5	1	5	0,44	4,86
11	1	2	21	0,37	4,30
12	2	2	11	0,42	4,36
13	3	2	8	0,56	4,56
14	4	2	12	0,63	4,60
15	5	2	2	0,68	4,60
16	1	3	18	0,56	3,82
17	2	3	10	0,60	3,75
18	3	3	4	0,72	3,78
19	4	3	3	0,89	3,87
20	5	3	3	1,00	3,86
21	1	4	20	3,45	3,84
22	2	4	14	2,74	3,61
23	3	4	7	2,62	3,50
24	4	4	8	2,31	3,42
25	5	4	5	2,20	3,35

C. Anexo: Curvas de supervivencia relativa de 5 años

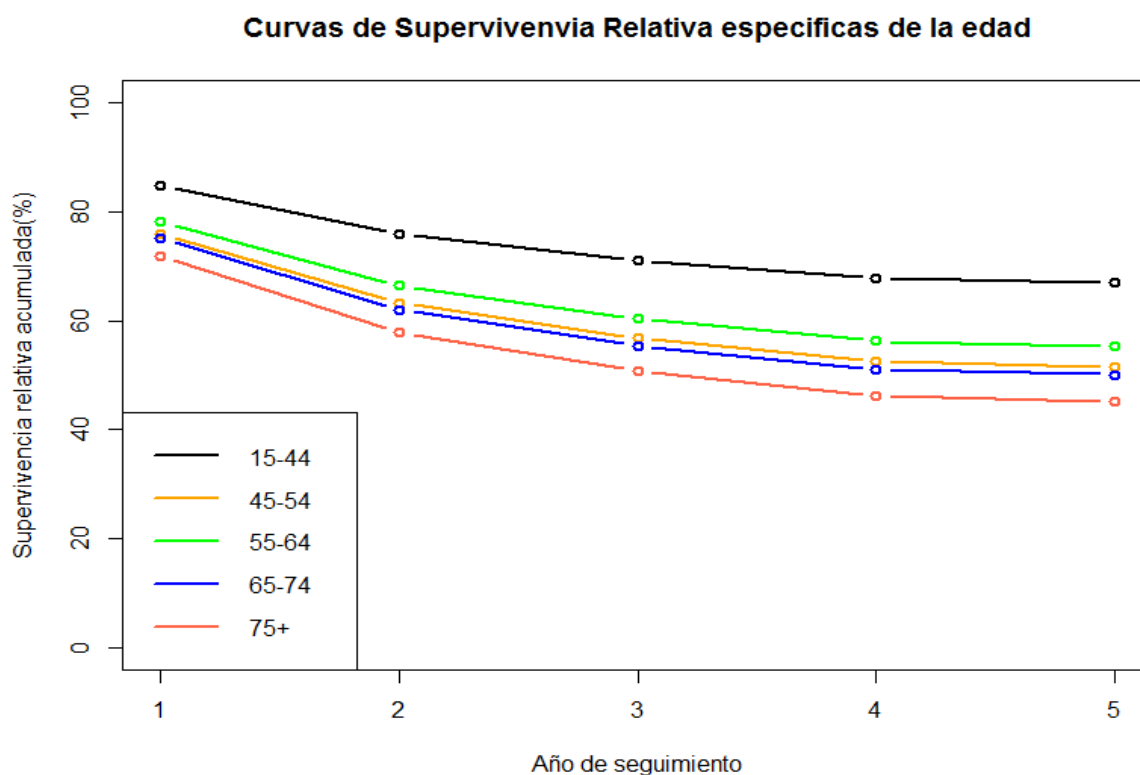


Figura C-1.: Curvas de supervivencia relativa de 5 años, por grupos de edad, Jordania.

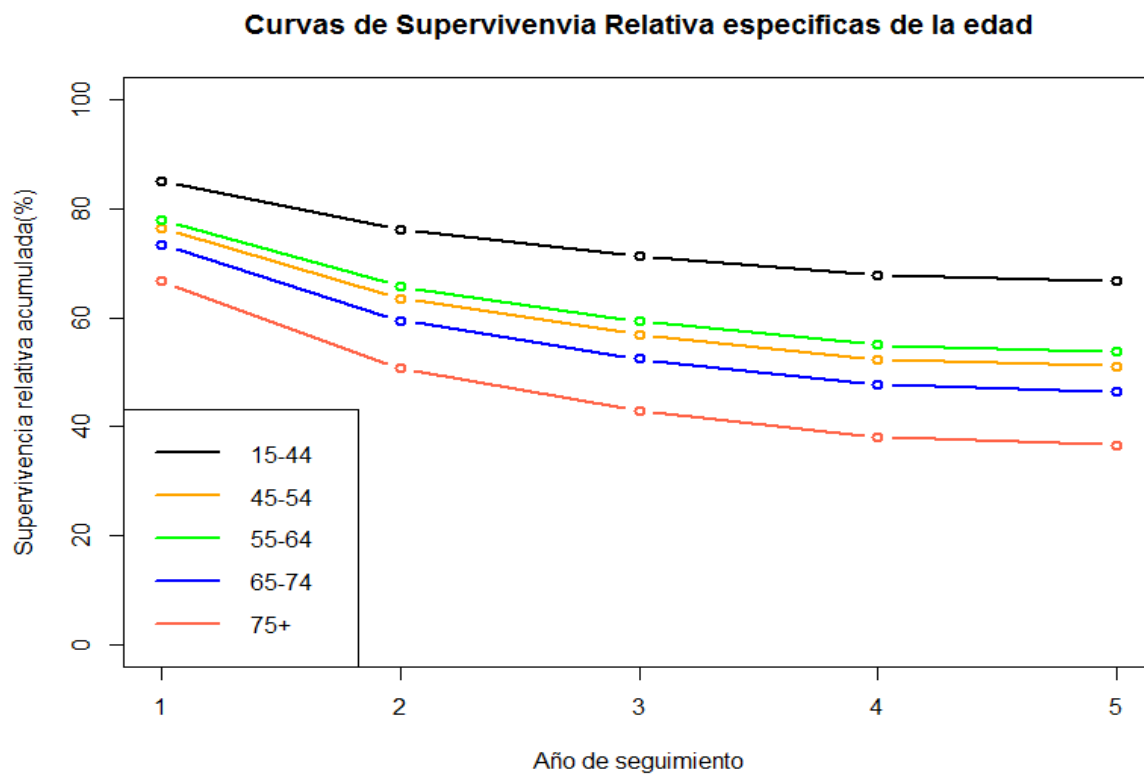


Figura C-2.: Curvas de supervivencia relativa de 5 años, por grupos de edad, Alemania.

D. Anexo: Comparación de mujeres ubicadas dentro y fuera de los puntos con altos niveles de intensidad.

Tabla D-1.: variable para modelo lineal generalizado de mujeres ubicadas en puntos de alta de intensidad.

año	gedad	obs	dstar	at_risk
1	0	5	0,008	1,872
2	0	3	0,010	1,946
3	0	2	0,017	2,351
4	0	2	0,023	2,526
5	0	1	0,030	2,741
1	1	2	0,014	1,504
2	1	1	0,022	1,792
3	1	1	0,038	2,197
4	1	3	0,045	2,303
5	1	1	0,049	2,251
1	2	5	0,045	1,792
2	2	1	0,054	1,872
3	2	1	0,071	2,079
4	2	2	0,076	2,015
5	2	1	0,083	1,946
1	3	4	0,058	0,916
2	3	1	0,036	0,000
3	3	1	0,064	0,693
4	3	1	0,095	1,099
5	3	1	0,129	1,386
1	4	4	0,183	0,916
2	4	1	0,224	0,916
3	4	2	0,285	1,253
4	4	1	0,280	1,099
5	4	1	0,326	1,253

Tabla **D-2.**: variable para modelo lineal generalizado de mujeres ubicadas en un bajo nivel de intensidad.

año	gedad	obs	dstar	at_risk
1	0	16	0,133	4,533
2	0	20	0,171	4,691
3	0	14	0,210	4,824
4	0	5	0,221	4,832
5	0	2	0,241	4,868
1	1	21	0,262	4,263
2	1	17	0,302	4,363
3	1	12	0,347	4,407
4	1	9	0,375	4,388
5	1	5	0,407	4,376
1	2	16	0,498	4,086
2	2	11	0,538	4,086
3	2	7	0,676	4,227
4	2	11	0,739	4,227
5	2	2	0,773	4,174
1	3	15	0,752	3,570
2	3	9	0,800	3,541
3	3	5	0,904	3,555
4	3	1	1,004	3,555
5	3	4	1,029	3,466
1	4	16	2,530	3,541
2	4	12	2,292	3,332
3	4	5	2,240	3,157
4	4	6	1,645	3,114
5	4	6	1,595	3,045

Tabla **D-3.**: Resumen de modelo lineal generalizado para análisis de supervivencia relativa (5 años) de pacientes con cáncer de cuello uterino de Cali, dentro de los puntos de alta intensidad, 2006-2014.

	Estimado	Std. Error	Valor z	$Pr(> z)$
y1	-0,271	0,337	-0,804	0,422
y2	-1,304	0,461	-2,831	0,005
y3	-1,726	0,469	-3,681	0,000
y4	-1,533	0,417	-3,679	0,000
y5	-2,338	0,550	-4,249	0,000
agr1	-0,221	0,457	-0,484	0,629
agr2	-0,040	0,432	-0,093	0,926
agr3	0,757	0,468	1,619	0,105
agr4	0,610	0,479	1,273	0,203

Tabla **D-4.**: Resumen de modelo lineal generalizado para análisis de supervivencia relativa (5 años) de pacientes con cáncer de cuello uterino de Cali, fuera de los puntos de alta intensidad, 2006-2014.

	Estimado	Std. Error	Valor z	$Pr(> z)$
y1	-1,667	0,167	-9,972	0,000
y2	-1,900	0,173	-11,015	0,000
y3	-2,453	0,201	-12,225	0,000
y4	-2,790	0,229	-12,187	0,000
y5	-3,519	0,324	-10,851	0,000
agr1	0,454	0,186	2,437	0,015
agr2	0,308	0,207	1,492	0,136
agr3	0,502	0,238	2,110	0,035
agr4	0,820	0,239	3,428	0,001

Bibliografía

ACS (2014), *American Cancer Society*. Accedió: 07/09/2017.

URL: <http://www.cancer.org/es/>

Aldrich, J. et al. (1997), 'Ra fisher and the making of maximum likelihood 1912-1922', *Statistical science* **12**(3), 162–176.

Alonso, J. C., Solano, M., Vera, R. & Gallego, A. I. (2007), 'Una mirada descriptiva a las comunas de Cali', *Santiago de Cali: Centro de Investigaciones en Economía y Finanzas (Cienfi)–Universidad Icesi*.

Alvarez, P. L.-L., Guerra-Yi, M. E., Faes, C., Alvarez, Y. G. & Molenberghs, G. (2009), 'Spatial analysis of breast and cervical cancer incidence in small geographical areas in cuba, 1999–2003', *European Journal of Cancer Prevention* **18**(5), 395–403.

Atalah, E., Urteaga, C., Rebolledo, A. & A Villegas, R. (2001), 'Alimentación, tabaquismo e historia reproductiva como factores de riesgo del cáncer de cuello del útero', *Revista médica de Chile* **129**(6), 597–603.

Beaglehole, R., Bonita, R. & Kjellström, T. (1994), *Epidemiología básica*, Organizacion Panamericana de la Salud. USA.

Berman, M. & Diggle, P. (1989), 'Estimating weighted integrals of the second-order intensity of a spatial point process', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 81–92.

Bernard, W. (2014), 'Stewart and christopher p. wild.: World cancer report, isbn 978-92-832-0429-9'.

Berrino, F., Gatta, G., Sant, M. & Capocaccia, R. (2001), 'The eurocare study of survival of cancer patients in europe: aims, current status, strengths and weaknesses', *European Journal of Cancer* **37**(6), 673–677.

Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattice systems (with discussion)', *J. Roy. Statist. Soc. Ser. B* **36** pp. 192—236.

Besag, J. (1977), 'Contribution to the discussion on Dr. Ripley's paper', *JR Stat. Soc. Ser* **39**, 193–195.

- Besag, J., York, J. & Mollié, A. (1991), 'Bayesian image restoration, with two applications in spatial statistics', *Annals of the institute of statistical mathematics* **43**(1), 1–20.
- Bianco, A. M. (2010), 'Modelo lineal generalizado', *FCEN* p. 142. Notas de Clase.
- Block, S. L., Nolan, T., Sattler, C., Barr, E., Giacoletti, K. E., Marchant, C. D., Castellsagué, X., Rusche, S. A., Lukac, S., Bryan, J. T. et al. (2006), 'Comparison of the immunogenicity and reactogenicity of a prophylactic quadrivalent human papillomavirus (types 6, 11, 16, and 18) 11 virus-like particle vaccine in male and female adolescents and young adult women', *Pediatrics* **118**(5), 2135–2145.
- Borja-Aburto, V. H. (2000), 'Estudios ecológicos', *salud pública de méxico* **42**(6), 533–538.
- Bravo, L. E., García, L. S. & Collazos, P. A. (2014), 'Cancer survival in Cali, Colombia: A population-based study, 1995-2004', *Colombia Médica* **45**(3), 110–116.
- Brenner, H. & Gefeller, O. (1996), 'An alternative approach to monitoring cancer patient survival', *Cancer: Interdisciplinary International Journal of the American Cancer Society* **78**(9), 2004–2010.
- Brenner, H. & Hakulinen, T. (2009), 'Up-to-date cancer survival: Period analysis and beyond', *International journal of cancer* **124**(6), 1384–1390.
- Cayuela, L. (2009), 'Modelos lineales generalizados (glm)', *Materiales de un curso del R del IREC*. EcoLab, Centro Andaluz de Medio Ambiente, Universidad de Granada–Junta de Andalucía.
- Cayuela, L., G. M. & Bolancé, C. (2016), 'introducción GLMs- función vínculo'. Universidad de Valencia, Notas de Clase.
- Chang, C. L. (1968), *Introduction to stochastic processes in biostatistics*, John Wiley And Sons, Inc; New York.
- CIA (2017), *Central Intelligence Agency*. Accedió: 07/11/2018.
URL: <https://www.cia.gov/library/publications/the-world-factbook/fields/2066.html>
- Coates, D. & Diggle, P. (1986), 'Tests for comparing two estimated spectral densities', *Journal of Time Series Analysis* **7**(1), 7–20.
- Colditz, G. A. (2015), *The SAGE Encyclopedia of Cancer and Society*, SAGE Publications.
- Corazziari, I., Quinn, M. & Capocaccia, R. (2004), 'Standard cancer patient population for age standardising survival ratios', *European Journal of Cancer* **40**(15), 2307–2316.
- Cressie, N. (1993), *Statistics for spatial data*, John Wiley & Sons. London, UK.

Cutler, S. J. & Ederer, F. (1958), 'Maximum utilization of the life table method in analyzing survival', *Journal of chronic diseases* **8**(6), 699–712.

de Cali, A. (2008), 'Plan de desarrollo 2008-2011'.

URL: <http://www.cali.gov.co/publico2/documentos/planeacion/plandesarrollo2008-2011.pdf>

De la Cruz, M. (2006), 'Introducción al análisis de datos mapeados o algunas de las (muchas) cosas que puedo hacer si tengo coordenadas', *Revista Ecosistemas* **15**(3).

Del Canto, M., García-Martínez, L., Fernández-Villa, T., Molina, A., Campanario, F., García-Sanz, M., López-Abente, G., Honrado, E. & Martín-Sánchez, V. (2015), 'Tendencia y distribución municipal de la incidencia de cáncer de vejiga en el área de salud de león (1996-2010)', *Actas Urológicas Españolas* **39**(6), 349–353.

Departamento Administrativo Nacional de Estadística (DANE). Accedió: 30/04/2018.

URL: <https://www.dane.gov.co/>

Diggle, P. (1985), 'A kernel method for smoothing point process data', *Applied statistics* pp. 138–147.

Dobson, A. J. & Barnett, A. (2008), 'An Introduction to Generalized Linear Models'. USA.

Ederer, F. (1961), 'The relative survival rate: a statistical methodology', *NCI Monograph* **6**, 101–121.

Ederer, F. & Heise, H. (1959a), 'The effect of eliminating deaths from cancer on general population survival rates', *Methodological Note No* **11**(23), 169.

Ederer, F. & Heise, H. (1959b), Instructions to ibm 650 programmers in processing survival computations, Technical report, Methodological note.

Espinal, E. K. & Aruner, M. E. P. (2014), 'Modelización cartográfica mediante funciones kernel para la ubicación óptima de centros de salud mental, que requieren limeñas agredidas psicológicamente por su pareja', *Revista ECIPerú Volumen* **11**(1).

Esteve, J., Benhamou, E., Raymond, L. et al. (1994), 'Statistical methods in cancer research. volume iv. descriptive epidemiology', *IARC Sci publ* **128**(1), 302.

Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012, GLOBOCAN (2012). Accedió: 09/10/2017.

URL: <http://globocan.iarc.fr/old/FactSheets/cancers/cervix-new.asp>

- García-Soto, G., de Jesús Pachón-Rincón, J., Meneses-Cubides, R. S. & Zuleta-Tobón, J. J. (2016), 'Cáncer de cuello uterino: experiencia durante un período de seis años en un hospital universitario colombiano', *Revista Colombiana de obstetricia y ginecología* **58**(1), 21–27.
- Giraldo, R. (2002), 'Introducción a la geoestadística: Teoría y aplicación', *Bogotá: Universidad Nacional de Colombia*.
- Giraldo, R. (2011), 'Notas de clase: Estadística espacial', *Universidad Nacional de Colombia*. Universidad Nacional de Colombia - Bogotá.
- Hakulinen, T. (1982), 'Cancer survival corrected for heterogeneity in patient withdrawal', *Biometrics* pp. 933–942.
- Holleczeck, B. & Brenner, H. (2013), 'Model based period analysis of absolute and relative survival with R: Data preparation, model fitting and derivation of survival estimates', *Computer methods and programs in biomedicine* **110**(2), 192–202.
- Imaz, M. S., Eimann, M., Poyard, E. & Savy, V. (2006), 'Exceso de mortalidad asociada a influenza en argentina: 1992-2002', *Revista chilena de infectología* **23**(4), 297–306.
- Instituto Nacional del Cáncer* (NCI). Accedió: 07/09/2017.
URL: <https://www.cancer.gov/español>
- Isaaks, E. & Srivastava, M. (1989), 'Applied geostatistics'. Oxford University Press.
- Liga Contra el Cáncer* (Seccional Bogotá). Accedió: 09/10/2017.
URL: <http://www.ligacontraelcancer.com.co/>
- Loader, C. (2006), *Local regression and likelihood*, Springer Science & Business Media.
- Madsen, H. & Thyregod, P. (2010), *Introduction to general and generalized linear models*, CRC Press.
- Manotas, E. N., Castellar, A. V., García, D. R. & Ibarra, K. H. (2014), 'Regresión logística ordinal aplicada a la identificación de factores de riesgo para cáncer de cuello uterino', *Ingeniare* (17), 87–105.
- McCullagh, P. & Nelder, J. A. (1989), 'Generalized linear models, vol. 37 of monographs on statistics and applied probability'. Chapman and Hall, London.
- Monroy Pacheco, J. A. & Vega Valero, A. E. (2014), Frontend del sistema de información para la identificación de vulnerabilidad territorial, B.S. thesis. Repositorio Institucional Universidad Católica de Colombia - RIUCaC.

- Montgomery, D., Peck, E. & Vining, G. G. (2006), 'Introducción al análisis de regresión lineal', *Compañía Editorial Continental. Tercera edición. México*.
- Mostacilla, B. & Elena, K. (2018), Redes sociales y economía local en el sector de Floralia del Jarrillón del Río Cauca en la ciudad de Cali, PhD thesis.
- Muñoz, N., Knaul, F. & Lazcano, E. (2014), '50 años del registro poblacional de cáncer de Cali, colombia'. SciELO Public Health.
- Nelder, J. A. & Baker, R. J. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society* **3**, 370–384.
- Organización Mundial de la Salud (2016), 'Control integral del cáncer cervicouterino: guía de prácticas esenciales, segunda edición'. World Health Organization.
- Ortega, G. L.-A. & Martí, C. I. (2001), *Aplicación de técnicas de análisis espacial a la mortalidad por cáncer en Madrid*, Comunidad de Madrid: Instituto de Salud Carlos III.
- Pardo, C., De Vries, E., Buitrago, L. & Gamboa, O. (2017), 'Atlas de mortalidad por cáncer en Colombia', *Bogotá DC: Instituto Nacional de Cancerología*.
- Ramos, C. P. & Duarte, R. C. (2015), 'Incidencia, mortalidad y prevalencia de cáncer en colombia, 2007-2011', pp. 978–958.
- Red de Salud de Ladera, I. F. P. f. C. C. & Prevention (2016), 'I simposio regional prevención primaria en cáncer de cuello uterino'. Hotel Four Points By Sheraton.
- Registro Poblacional de Cáncer de Cali (RPCC)*. Accedió: 09/10/2017.
URL: <http://rpcc.univalle.edu.co/>
- Registry, S. C. (2011), *periodR – an R package to calculate long term cancer survival estimates using period analysis*. Accedió: 1/9/2018.
URL: http://www.krebsregister.saarland.de/improve/periodR_en.html
- Ripley, B. D. (1976), 'The second-order analysis of stationary point processes', *Journal of applied probability* **13**(2), 255–266.
- Royo, M. & Moreno, J. (2009), 'Manual docente de la escuela nacional de sanidad'.
- Schabenberger, O. & Gotway, C. A. (2005), *Statistical methods for spatial data analysis*. Inglaterra, Chapman Hall.
- Silverman, B. W. (1986), *Density estimation for statistics and data analysis*, Vol. 26, CRC press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. (2002), ‘Bayesian measures of model complexity and fit’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 583–639.

UN (2013), *United Nations Population Division*. Accedió: 04/10/2018.

URL: <http://www.un.org/en/development/desa/population/publications/mortality/mortpak.shtml>

Vásquez, E. (2001), ‘Historia de Cali en el siglo 20. sociedad, economía, cultura y espacio’.

Wald, A. (1944), ‘On a statistical problem arising in the classification of an individual into one of two groups’, *The Annals of Mathematical Statistics* **15**(2), 145–162.

World Health Organization (OMS). Accedió: 07/09/2018.

URL: <http://www.who.int/>

Xie, Z. & Yan, J. (2008), ‘Kernel density estimation of traffic accidents in a network space’, *Computers, environment and urban systems* **32**(5), 396–406.