Análisis espacial y modelación de la ocurrencia de tres tipos de cáncer infantil en la ciudad de Santiago de Cali en el periodo del 2009-2016
Programa Académico de Estadística

Maria Elena Colina Hincapie
Mercedes Andrade Bejarano PhD
Doc. Oscar Ramírez

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2017

Análisis espacial y modelación de la ocurrencia de tres tipos de cáncer infantil en la ciudad de Santiago de Cali en el periodo del 2009-2016 Programa Académico de Estadística

Maria Elena Colina Hincapie
Mercedes Andrade Bejarano PhD
Doc. Oscar Ramírez

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de:
Estadístico(a)

Director(a):
Mercedes Andrade Bejarano PhD
Codirector(a):
Doc. Oscar Ramírez

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2017

"Seguir cuando crees que no puedes más es lo que te hace diferente a los demás. La vida no se trata de la fuerza de tus golpes, sino de la fortaleza de tu carácter. No te des por vencido".

Sylvester Stallone

"Todo parece imposible hasta que se hace"

Nelson Mandela

"Pobre no es el hombre cuyo sueños no se han realizado, sino aquel que no sueña"

Marie Von Ebner

Agradecimientos

Quiero agradecer inicialmente a Dios, por siempre darme la fortaleza de enfrentar cada reto que se presenta en mi vida y por acompañarme a realizar todos los pasos con firmeza, por nunca abandonarme y siempre darme las herramientas espirituales para nunca decaer.

Agradecerle a mis padres por siempre estar a mi lado en los momentos difíciles y darme el apoyo necesario para este proceso, porque con sus amor y con sus palabras sabias, hacen que todo lo haga con amor y dedicación.

Agradecerle a mi prima Monica Castaño Hincapie y a mi tia Maria Nora Hicapie, por acompañarme en todos los procesos logrados en mi vida, por enseñarme ese camino de fortaleza y conocimiento, que espero seguir compartir con Veronica González Castaño.

A mis amigos les agradezco todas las experiencias vividas, las explicaciones, las tristezas y alegrías vividas en el transcurso de la carrera.

A la profesora Mercedes Andrade y al Medico Oscar Ramirez, les agradezco por su asesoría, su apoyo y conocimiento transmitidos para el desarrollo de este proyecto.

A todos los profesores les agradezco, por compartir todo el conocimiento que tengo hoy, por darme las herramientas necesarias para ser una gran profesional.

Y cada una de las personas que aportaron su granito de arena para que este sueno se cumpliera, a cada uno mil gracias por hacer parte de mi vida y de esta experiencia.

Maria Elena Colina Hincapie

Resumen

La red de gestores sociales informa que en Colombia, los casos de cáncer infantil se encuentra en aumento, siendo esta la segunda causa de muerte en niños entre los 5 y 15 años, teniendo en cuenta que el cáncer pediátrico es poco frecuente y representa menos del 3 % de las neoplasias malignas en el mundo. Las técnicas estadísticas implementadas en este estudio contribuyen a la planeación, ejecución y evaluación de políticas públicas sectoriales. Durante esta investigación se implementó métodos de análisis espacial, con los cuales se realizó un análisis de intensidad y de aleatoriedad de los casos de Leucemia Linfoblastica Aguda, Linfoma de Burkitt y Meduloblastomas en la ciudad de Santiago de Cali, aplicando finalmente un modelo lineal generalizado con Variables socio-demográficas. Concluyendo se encontró que estas enfermedades son de poca frecuencia y que los casos de Leucemia Linfoblastica Aguda siguen un patrón agrupado, por otra parte los casos de Linfoma de Burkitt y Meduloblastomas siguen un patrón aleatororio, y el mejor modelo implementado es un modelo lineal generalizado con variable de respuesta Gamma y función de enlace Logaritmica, utilizando como variable de respuesta la tasa de casos de cada tipo de cáncer infantil, con las variables explicativas de Etnia del paciente y Densidad de Niños menores de 15 años por comuna.

Abstract

The social managers network report that in Colombia the childhood cancer cases is growing, this being the second cause of death in children between 5 and 15 years, taking into account that pediatric cancer is little frequent and represents less than 3 % of the malignant neoplasms in the world. The statistical techniques implemented in this study are to the planning, execution and evaluation of sectoral public policies. During this investigation, methods of spatial analysis were implemented, which was carried out an intensity and randomness analysis of the cases of Leukemia of precursor cells, Burkitt's lymphoma and Medulloblastomas in the Santiago de Cali space, finally applying a generalized linear model. with socio-economic variables. Concluding that these diseases are infrequent and that the cases of Precursor Cell Leukemia continue to be a grouped pattern, in the change the cases of Burkitt's Lymphoma and Medulloblastomas follow a random pattern, and the best model implemented is a generalized linear model with Gamma response variable and Logarithmic link function, for the variable response rate of cases of this type of childhood cancer, with the explanatory variables of the patient's ethnicity and children's density by commune.

Cuadro **0-1**: Tabla de abreviaturas

RPCC	Registro Poblacional de Cáncer de Cali
OMS	Organización Mundial de la Salud
OPS	Organización Panamericana de la Salud
ICCC	Clasificación Internacional de Cáncer Infantil
GLM	Modelo Lineal Generalizado
ANOVA	Análisis de la Varianza
AIC	Criterio de Información de Akaike
VIF	Factor de Incremento de Varianza
UTM	Sistema de Coordenadas Universal Transversal de Mercator
CV	Criterio de Validación Cruzada
LCV	Criterio de Verosimilitud de Validación Cruzada
R^2	Coeficiente de Determinación
SE	Error Estandar
LLA	Leucemia Linfoblastica Aguda

Contenido

Resumen vii

Lista de Figuras xv

Lista de Tablas xviii

1 Introducción 1
- 1.1 Descripción del problema . . . 1
- 1.2 Justificación . . . 2
- 1.3 Antecedentes . . . 3
- 1.4 Objetivos . . . 7
  - 1.4.1 Objetivo General . . . 7
  - 1.4.2 Objetivos Específicos . . . 7

2 Marco Teórico 8
- 2.1 Marco Conceptual . . . 8
- 2.2 Marco teórico del análisis espacial . . . 14
  - 2.2.1 Estadística Espacial . . . 14
  - 2.2.2 Patrones puntuales . . . 15
  - 2.2.3 Test de completa aleatoriedad espacial . . . 17
  - 2.2.4 Función de intensidad . . . 21
- 2.3 Modelación . . . 23
  - 2.3.1 Prueba Anderson Darling para Bondad de Ajuste . . . 23
  - 2.3.2 Modelos Lineales Generalizados . . . 24
  - 2.3.3 Modelo de Regresión Lineal Generalizado con Variable de Respuesta Gamma . . . 26
  - 2.3.4 Estimación de los parámetros . . . 27
  - 2.3.5 Pruebas de Bondad de Ajuste . . . 29
  - 2.3.6 Multicolinealidad . . . 33

3 Metodología 35
- 3.1 Población de estudio . . . 35
- 3.2 Definición de Caso . . . 35

3.3 Criterios de Inclusión y Exclusión 37
3.3.1 Criterios de Inclusión 37
3.3.2 Criterios de Exclusión 37
3.4 Base de datos 37
3.5 Georeferenciación 40
3.6 Análisis Estadístico 40
3.6.1 Análisis Exploratorio 40
3.6.2 Test de completa aleatoriedad espacial 40
3.6.3 Estimación de intensidad 44
3.7 Modelación 46

4 Resultados **54**
4.1 Análisis Exploratorio 54
4.1.1 Características de las variables, por comunas 57
4.2 Análisis Espacial 60
4.2.1 Pruebas de Aleatoriedad Espacial 61
4.2.2 Estimación de la intensidad 62
4.3 Modelación 65

5 Conclusiones y Recomendaciones **71**
5.1 Conclusiones 71
5.2 Limitaciones del Estudio 72
5.3 Recomendaciones 73

6 Anexos **74**

Bibliografía **82**

Lista de Figuras

2-1. Tipos de patrones espaciales 16

3-1. Diagrama de depuración de base de datos 39
3-2. Función distancia del vecino más cercano *G(h)* para cada tipo de patrón . . 42
3-3. Función distancia del vecino más cercano *F(h)* para cada tipo de patrón . . . 43
3-4. Función de la medida reducida *K(h)* para cada tipo de patrón 44
3-5. Función de la medida reducida modificada *L(h)* para cada tipo de patrón . . 44
3-6. Prueba Anderson Darling para la distribución de la tasa de casos de Leucemia 48
3-7. Validación de los supuestos para el modelo completo 52

4-1. Frecuencias del conteo por comuna, de acuerdo al genero de los registros de Leucemia 57
4-2. Frecuencias del conteo por comuna, de acuerdo a la edad categorizada de los registros de leucemia 58
4-3. Frecuencias del conteo por comuna, de acuerdo a la característica de afrocolombiano de los registros de leucemia 59
4-4. Frecuencias del conteo por comuna, de acuerdo a la característica de células B o T de los registros de leucemia 60
4-5. Área urbana de la ciudad de Santiago de Cali y los casos de cáncer infantil en el período 2009-2016 61
4-6. Estimación Kernel de la intensidad para los casos de cáncer infantil en Cali . 63
4-7. Estimación de intensidad por medio del criterio de validación cruzada para los casos de cáncer infantil en Cali 64
4-8. Validación de los supuestos para el modelo seleccionado 70

6-1. Función G Leucemia 74
6-2. Función G Burkitt 75
6-3. Función G Meduloblastoma 75
6-4. Función F Leucemia 76
6-5. Función F Burkitt 76
6-6. Función F Meduloblastoma 77
6-7. Función K Leucemia 77
6-8. Función K Burkitt 78
6-9. Función K Meduloblastoma 78

6-10. Ancho de Banda para los casos de Leucemia 79
6-11. Ancho de Banda para los casos de Burkitt 79
6-12. Ancho de Banda para los casos de Meduloblastoma 80
6-13. Estamación de intensidad por ambos métodos para los casos de Leucemia . . 80
6-14. Estimación de Intensidad por ambos métodos para los casos de Linfoma de Burkitt 81
6-15. Estimación de Intensidad por ambos métodos para los casos de Meduloblastomas 81

Lista de Tablas

0-1. Tabla de abreviaturas IX

2-1. Riesgo en las Anomalías cromosómicas (Carroll and Bhatla, 2016) 10
2-2. Diagnóstico diferencial entre el Linfoma de Burkitt (Imbach et al., 2011a) . . 11
2-3. Categorias de riesgo de meduloblastoma (Hanson and Atlas, 2016) 12
2-4. Funciones de vínculo más comunes utilizadas por los GLM (Cayuela, 2010) . 26
2-5. Evaluación inicial de la bondad del ajuste de un modelo H_1, H_0 y $\hat{\mu_0}$ se refiere a modelo mínimo, es decir, un modelo con todas las observaciones que tienen el mismo valor medio. 31

3-1. Cantidad de Habitantes por comuna en Santiago de Cali (Alcaldia de Santiago de Cali, 2017) 36
3-2. Tabla de Contingencia para cuantificar las variables a emplear en el modelo . 47
3-3. Tabla de asociación entre las variables categóricas 47
3-4. Expresiones de los modelos simples 49

4-1. Número de casos para el tipo de cáncer por genero 55
4-2. Estadísticas descriptivas para la edad por genero 55
4-3. Numero de casos para los periodos por Tipo de cáncer 55
4-4. Estadísticas descriptivas para las Comunas por genero 56
4-5. Estadísticas descriptivas para etnia por enfermedad 56
4-6. Estadísticas descriptivas para tipo de células de Leucemia 56
4-7. Prueba de Aleatoriedad para los casos de Leucemia Linfoblastica Aguda . . . 62
4-8. Prueba de Aleatoriedad para los casos de Linfoma de Burkitt 62
4-9. Prueba de Aleatoriedad para los casos de Meduloblastoma 62
4-10.Ancho de Banda para la estimación de intensidad por medio de los criterios . 63
4-11.Aporte individual de cada variable a la tasa de casos de LLA 66
4-12.Análisis de Desvianzas para el modelo completo 67
4-13.Análisis de Desvianzas para el modelo completo 67
4-14.Estimaciones del modelo completo 67
4-15.Raíz cuadrada del factor de inflación de la varianza (VIF) 68
4-16.Proceso de selección de variables *Srepwise* 68
4-17.Análisis de Desvianzas para el modelo seleccionado 68
4-18.Análisis de Desvianzas para el modelo completo 69

4-19.Resumen del modelo completo 69
4-20.Raíz cuadrada del factor de inflación de la varianza (VIF) 69

1 Introducción

1.1. Descripción del problema

La red de gestores sociales indica que en Colombia los casos de cáncer en niños (menores de 15 años) se encuentran en aumento (Garzón F, 2014), siendo el cáncer infantil la segunda causa de muerte en niños entre los 5 y 15 años. El cáncer pediátrico es una de las enfermedades con poca frecuencia y solo representa entre el 0.5 % y 3 % de todos los casos de cáncer en la población en el mundo (Wurttemberger, 2016). Según los datos del Registro Poblacional de Cáncer de Cali (RPCC), la probabilidad de sobrevida a los 5 años en Cali, es de 48-53 %; en cambio en otros países esta expectativa oscila entre el 70-80 % (Garzón F, 2014); estas cifras crean una alarma, ya que la mortalidad de niños con cáncer puede ir aumentando si no se mejoran los servicios de salud (Wurttemberger, 2016).

De acuerdo Fajardo-Gutiérrez et al. (1999), las neoplasias malignas más frecuentes en niños menores de 15 años, en algunos países del mundo son las leucemias con un 30-34 %, seguido de los linfomas con un 13-18 % y para los tumores del sistema nervioso central con un 10-14 %. Según los reportes de Registro Poblacional de Cáncer en Cali (RPCC), se estima una tasa de mortalidad en Colombia de 5.74 por millón-niños año a causa del cáncer, en cambio en Cali durante el periodo de 1994 al 2004 la tasa de mortalidad se ha mantenido estable con un índice alrededor de 5.5 por millón-niños año (Garzón F, 2014).

Según Bravo et al. (2013), se identificaron 1548 casos nuevos de cáncer infantil en Santiago de Cali entre los años 1992-2011, lo que significa que en promedio se tienen 77,4 casos nuevos de cáncer pediátrico por año (Bravo et al., 2013). De acuerdo a la Organización Panamericana de la Salud (OPS), la probabilidad de sobrevida son más bajas, para niños que viven en entornos de bajos recursos, donde aproximadamente uno de cada dos niños fallecen por esta enfermedad en América.

Son pocas las investigaciones realizadas a enfermedades de poca frecuencia que no son transmisibles y en el que utilicen métodos estadísticos y espaciales. En la ciudad de Santiago de Cali, Colombia, no hay estudios en el que muestren la ubicación exacta de los casos de cáncer infantil, y mucho menos que realicen un estudio de patrones espaciales de esta enfermedad. Por lo que a partir de esto, surge la pregunta: ¿Existe un patrón espacial en los casos de cáncer pediátrico en Santiago de Cali, mostrando mayor incidencia en ciertas

zonas de la ciudad?

1.2. Justificación

El cáncer infantil es el segundo causante de muertes entre niños y adolescentes, en Santiago de Cali se tiene una probabilidad de sobrevida muy baja, con una tasa de mortalidad que se ha mantenido estable alrededor de 20-23 millón-niño (Wurttemberger, 2016), pero si estas tasas llegan a aumentar, se pueden llegar a existir cambios demográficos, causando un gran impacto en la población.

En epidemiologia se suelen utilizar indicadores como: la incidencia, la mortalidad, la supervivencia, pero nunca se ha identificado por medio de patrones espaciales la ocurrencia de enfermedades no transmisibles (Bravo et al., 2013). Los indicadores paremiológicos antes mencionados contribuyen a la planeación, ejecución y evaluación del cáncer en niños. Además los profesionales de la salud pública pueden utilizar este tipo de análisis para centrar su atención en algunas áreas de Santiago de Cali y disminuir el riesgo de incidencia de cáncer infantil.

La estadística geoespacial, permite obtener de manera visual el riesgo de presentar algún evento de interés teniendo en cuenta el aspecto espacial (Cressie, 2015), en este caso el estudio se enfoca en tres tipos de cáncer: Leucemia Linfoblastica Aguda(LLA), Linfomas de Burhitt y Meduloblastomas en niños menores de 15 años en la ciudad de Santiago de Cali entre el periodo 2009 al 2016. La metodología de geoestadística espacial permite visualizar el comportamiento epidemiológico para luego asociarlo con condiciones sociales o demográficos.

De igual manera, es necesario examinar los efectos de primer orden espacial, que analizan la estructura de interacción (dependencia) espacial de los casos puntuales de patrones espaciales, lo que permite estudiar si en la ciudad de Santiago de Cali los eventos de cáncer infantil suceden de manera aleatoria, agrupado o regular en el espacio.

De igual manera, es de interés en este trabajo, la modelación de las tasas de casos de cáncer infantil enfocado a las tres enfermedades de interés en el perímetro urbano de Santiago de Cali a través de variables que lo explique, como las características de los pacientes y de las comunas de la ciudad. Al modelar una variable continua positiva, se descarta inmediatamente la posibilidad de utiliza un modelo lineal con distribución normal, por lo que es mejor trabajar con un modelo lineal generalizado, ya que la teoría de este tipo de modelos se ajusta para variables de respuesta cuya distribución pertenece a la familia exponencial (Dobson and Barnett, 2008).

1.3. Antecedentes

A continuación se citan algunos trabajos nacionales e internacionales sobre estudios de modelaciones y estadística espacial aplicado en cáncer infantil y se comentara la metodología empleada en las investigaciones.

Tovar et al. (2016)
Desarrollaron un estudio en Santiago de Cali, Colombia, con pacientes entre los 0 y 15 años de edad, diagnosticados durante el periodo del 2009 al 2013 con algún cáncer de acuerdo al ICCC-3, con el objetivo de describir el comportamiento del número de casos en las comunas de la ciudad.

Por medio de los datos del sistema de vigilancia epidemiológica de cáncer infantil (Vigicancer) y las variables sociodemográficos suministradas por los mismos, realizaron el calculo de las tasas de incidencia estandarizadas por edad y comuna, tomando como base la población mundial de menores con el rango de edades de interés. El cálculo de la tasa se puede implementar en este proyecto, utilizando una población en riesgo que se encuentre dentro de la edad establecida.

Además desarrollaron inferencia bayesiana para calcular las probabilidades de riesgo por comuna dentro de la ciudad de Santiago de Cali, procediendo a realizar una elicitación de la información a priori (comportamiento de la naturaleza) conocida. Al ser estudiado el número de casos, trabajaron con una distribución Poisson, tomando como a priori una distribución Gamma y finalmente teniendo como a posteriori y calculando con esta las probabilidades predictivas una distribución Binomial Negativa.

En este estudio Concluyeron que las tasas de incidencia de cáncer infantil observadas para la ciudad fueron menores en comparación con las reportadas en la literatura; encontraron una incidencia de 121 casos por millón de habitantes.

Hurtado González and Ramírez Rodriguez (2015)
Desarrollaron un estudio en Santiago de Cali, Colombia, con pacientes menores de 15 años de edad, diagnosticados durante el periodo del 2009 al 2013 con tres tipos de cáncer infantil, con el objetivo de realizar un geoportal que ofrece información geográfica actualizada de las enfermedades y con mayor relevancia clínica en las comunas de la ciudad.

Con el fin de permitir visualizar mapas temáticos relacionados con la ocurrencia, incidencia y probabilidad de predicción de los tres tipos de cáncer infantil, el geoportal realiza una integración de la modelación estadística que involucra modelos bayesianos y análisis espaciales.

En este estudio se trabajó con los registros suministrados por VIGICANCER; el modelo estadístico implementado fue a partir de una variable de conteo por comuna, asumiendo que esta se distribuye Poisson. Por medio de las métodos bayesianos calcularon una distribución a priori Gamma, obteniendo como resultados probabilidades predictivas con el modelo Poisson-Gamma y a partir de estas estimaciones realizaron la representaciones geográficas por comunas y las plasmaron en el geoportal.

Ortega-García et al. (2011)
Desarrollaron un estudio en Murcia, España, con pacientes menores de 15 años de edad, diagnosticados entre el 1 de enero de 1998 y 31 de diciembre de 2009 con algún tipo de cáncer de acuerdo al ICCC-3, con el objetivo de crear mapas de incidencia y analizar la distribución geográfica del cáncer pediatrico.

Para el desarrollo de este estudio, obtuvieron 3 direcciones postales (durante el embarazo, periodo posnatal y en el momento del diagnóstico) para realizar un contraste durante todo el proceso de crecimiento del menor. Junto a esta realización geográfica calcularon la tasa cruda y la tasa de incidencia estandarizada considerando en estos cálculos a todos los niños inmigrantes del proceso de estudio. Para este proyecto en los cálculos de las tasas se tendrá en cuenta a todos los niños estudiados durante el tiempo de interés.

Para el análisis utilizaron las variables: Zona básica de salud, sexo, fecha de nacimiento, fecha de diagnóstico y diagnostico patológico. El cual fueron utilizados para realizar una tabulación por barrio de Murcia y por medio de este, encontrar agregaciones, mostrando de esta forma cada zona con el cálculo de las tasas antes mencionadas.

En el documento indicaron que la identificación de patrones geográficos puede sugerir estudios posteriores más específicos, que permitan identificar aquellos factores que inciden en el riesgo de presentar algunos de los tipos de cáncer.

Durá et al. (2016)
Desarrollaron un estudio en la provincia de Villa Clara, con pacientes entre los 0 y 18 años de edad, diagnosticados durante el periodo de 2009 al 2012 con Leucemia Aguda, con el objetivo de establecer la distribución del riesgo de enfermar Leucemia Aguda en niños, en el tiempo, el espacio y el espacio/tiempo.

Para el cumplimiento del objetivo implementaron tres estudios de tasas puramente temporales (por año), tasas puramente espaciales (por barrios) y tasas espacio/temporales (cada barrio tiene un tiempo estudiado) que lograron representar en mapas. Para cada uno de los estudios calcularon los casos observados, casos esperados, tasa de incidencia y riesgo

relativo con su respectivo valor p, el cual les ayudo a identificar si el riesgo relativo era significativo, indicando que hay más probabilidad de enfermar dentro del barrio que fuera de él.

Finalmente concluyeron que el conglomerado de incidencia es detectado en el noroeste de Villa Clara, en especial un patrón de distribución de altas tasas en la franja costera, haciendo un llamado a mantener vigilancia en estas zonas para no aumentar el riesgo de contraer esta enfermedad.

Wheeler (2007)
Desarrollaron un estudio en Ohio, E.E.U.U., con pacientes entre los 0 y 14 años de edad, diagnosticados durante el periodo de 1996 al 2003 con Leucemia, con el objetivo de analizar la distribución espacial de la incidencia de Leucemia.

Utilizaron los datos del Sistema de Vigilancia de la Incidencia del Cáncer de Ohio (OCISS), y exploraron que los conglomerados globales y los grupos locales fueran estadísticamente significativos de los casos de leucemia individual en Ohio, es decir, que existieran conglomerados específicos y un conglomerado dentro del área estudiada. Por esta razón, se centraron en un estudio espacial de casos y controles, adquiriendo una muestra de controles aquellos niños que no desarrollaron leucemia durante el mismo período de tiempo de los nacimientos, de los casos estudiados.

Para cumplir con el objetivo de este estudio utilizaron patrones puntuales, para definir los casos y los controles dentro del mismo espacio. Utilizaron la función de intensidad para identificar la existencia de varios conglomerados dentro del área y el método del vecino más cercano o función K, como prueba para encontrar una agrupación global. Para este Para este proyecto se utilizara el método de estimación de aleatoriedad espacial por medio de distancias, para ver qué tipo de patrones siguen los datos estudiados en esta investigación.

Para la función de intensidad utilizaron la estimación del centro del ancho de banda, y finalmente utilizaron la función de intensidad del kernel y la estadística de exploración espacial de Kulldorff, con el fin de probar que los clusters locales eran significativos. Para la investigación de la agrupación potencial y los conglomerados locales, asumieron una realización de un proceso de punto Poisson heterogéneo para los casos y controles, con una hipótesis nula de riesgo constante donde se esperaban más casos con una población en mayor riesgo.

Este estudio comparativo para clusters y agrupamientos de leucemia infantil en Ohio es el primero con datos de casos y controles a nivel individual. Por lo que lograron concluir, que el método de función de intensidad de kernel sugiere agrupamientos estadísticamente

significativos en áreas del centro, sur, y el este de Ohio, además de que los hallazgos son consistentes para las diferentes pruebas de agrupamiento global, donde no hay una agrupación significativa cuando se consideran todos los casos de edad. Durante esta investigación el método de intensidad ayudara a identificar agrupaciones en Santiago de Cali para los tipos de cáncer estudiados, el cual se realizaran con estimación Kernel y un ancho de banda que se logre ajustar a los datos.

Thompson et al. (2008)
Desarrollaron un estudio en Texas, E.E.U.U., con pacientes menores a 18 años de edad, diagnosticados durante el periodo de 1990 al 2003 con algún tipo de los 19 grupos de cáncer según el ICCC-3, con el objetivo de evaluar los posibles patrones de riesgo geográfico en Texas.

En este estudio se utilizó un modelo de riesgo espacio-temporal, donde los 19 grupos de cáncer infantil se modelaron como potencialmente correlacionados, a partir de un modelo jerárquico con inferencia bayesianas y estimaciones MCMC de 100.000 interacciones. Las variables que utilizaron son: Zonas de cultivo agricola, la liberación intensiva de contaminantes peligrosos del aire, la densidad de población y el rápido crecimiento de la población, siendo siempre la unidad de medida el condado-año.

Concluyeron que la implementación bayesiana de un modelo autorregresivo condicional multivariante proporcionó un enfoque flexible para el modelo espacial de múltiples tipos de cáncer infantil. Además este estudio logró identificar factores geográficos que respaldan estudios más enfocados de tumores de células germinales y otros tipos de cáncer en áreas de grandes cultivos.

Rainey et al. (2007)
Desarrollaron un estudio en Kenia, África, con pacientes menores a 15 años de edad, diagnosticados durante el periodo de 1988 al 1997 con Linfoma de Burkitt pediátrico, con el objetivo de identificar si existe una relación espacial entre el Linfoma de Burkitt y la Malaria.

Para obtener los datos realizaron un registro clínico de los pacientes diagnosticados con Linfoma de Burkitt en siete hospitales provinciales y un hospital nacional. Los datos se recolectaron retrospectivamente para el período 1988-1992 y prospectivamente para el periodo 1993-1997.

Una de las variables utilizadas en este estudio fue la tribu, ya que es un factor de riesgo independiente para el Linfoma de Burkitt y se encuentra asociado con el riesgo de malaria; para ver la asociación de estas variables utilizaron una prueba chi-cuadrado, seguidamente aplicaron un modelo de regresión logarítmica-lineal, para evaluar el papel independiente de

estos factores en las tasas de incidencia a nivel de distrito en Kenia. La metodología de la prueba de asociación será empleada en el proyecto, para mirar la relación entre las variables sociodemográficas.

Entre los resultados del modelo logarítmico lineal, encontraron que el Linfoma de Burkitt fue 3.5 veces mayor en la incidencia de transmisión independientemente de la tribu. A partir de esto concluyeron que el modelo de regresión log-lineal confirma que los niveles de parasitemia y la morbilidad varían dentro de cada categoría, observando que los niveles de transmisión deben analizarse más a fondo. Además este modelo mostró que solo la transmisión de la malaria y otra enfermedad denominada la epidemia del lago, presentaba un riesgo estadísticamente elevado de tener el Linfoma de Burkit en comparación con los distritos de malaria de bajo riesgo.

1.4. Objetivos

1.4.1. Objetivo General

Realizar un análisis espacial y modelar la tasa de casos de Leucemia Linfoblastica Aguda, Linfomas de Burhitt, Meduloblastomas en niños menores de 15 años en la ciudad de Santiago de Cali entre el 2009 al 2016.

1.4.2. Objetivos Específicos

- Realizar un análisis descriptivo del comportamiento espacial de los casos de Leucemia Linfoblastica Aguda, Linfomas de Burhitt, Meduloblastomas en niños menores de 15 años en la ciudad de Santiago de Cali entre el 2009 al 2016.
- Definir si existe un patrón espacial y su intensidad en caso de exista en los casos de Leucemia Linfoblastica Aguda, Linfomas de Burhitt, Meduloblastomas en niños menores de 15 años en la ciudad de Santiago de Cali entre el 2009 a 2016.
- Modelar la tasa de casos de Leucemia Linfoblastica Aguda, Linfomas de Burhitt, Meduloblastomas en niños menores de 15 años en la ciudad de Santiago de Cali entre el 2009 a 2016 a través de variables demográficas que lo expliquen.

2 Marco Teórico

En esta sección se presenta el marco teórico de la investigación. En la primera parte se muestra el marco conceptual, en el cual se incluye el tipo de estudio epidemiológico, definiciones importantes sobre el cáncer infantil y los tres tipos de cáncer que se tratan en este estudio. En la segunda parte se presenta lo concerniente al análisis espacial, conceptos sobre georreferenciación, algunas definiciones importantes de la estadística para datos espaciales y por último el marco teórico del tipo de modelo que se utilizó en este estudio.

2.1. Marco Conceptual

El marco conceptual esta compuesto por algunas definiciones epidemiologicas necesarias en la investigación.

Cáncer Infantil

El cáncer infantil es el cáncer que afecta a niños y jóvenes que se encuentran entre los 0 y 15 años. Los tipos de cáncer infantil más frecuentes son las leucemias, tumores de sistema nervioso y del sistema linfático, cada uno de estos se comporta de forma diferente, pero todos se caracterizan por la proliferación descontrolada de células anormales (Instituto Nacional de Cancerologia, 2016). Un niño o joven diagnosticado con cáncer, debe tener un diagnóstico preciso y debe ser tratado por equipos de especialistas en oncología pediátrica. De acuerdo a la Organización Panamericana de la Salud(OPS) y la Organización Mundial de la salud(OMS) cada año se diagnostican mas de 27.000 casos de cáncer en niños menores de 14 años en la región de las Americas (Organizacón Mundial de la Salud, 2016; Organizacón Panamerica de la Salud, 2016). Además, encontraron que gracias a los avances en los tratamientos, se han logrado altas probabilidades de supervivencia, la cual se aproximan al 80 %; sin embargo estas tasas son significativamente mas bajas para aquellos niños que viven en entornos de bajos recursos.

Durante el estudio se van observar tres tipos de cáncer infantil, la cual cada una tiene origen diferente de células cancerosas. Estas clases de Cáncer pediátrico son:

Tipos de Cáncer infantil mas frecuentes:

- **Leucemia Linfoblastica Aguda (LLA):**

  Las leucemias agudas representan una expansión y detención clonal en una etapa específica de la hematopoyesis mieloide o linfoide normal. Constituyen el 97 % de todas las leucemias infantiles y consisten a la Leucemia Linfoblástica Aguda (ALL), Leucemia Mieoblástica Aguda (LMA), Leucemia Aguda Indiferenciada, Leucemia Aguda de Linaje Mixto, Leucemia Mieloide positiva al cromosoma Philadelphia y la Leucemia Mielomonocítica Juvenil (Carroll and Bhatla, 2016).

  La leucemia linfoblástica aguda (LLA), también conocida como leucemia linfocítica aguda, es una neoplasia maligna de las células precursoras de los linfocitos, o linfoblastos. Los linfoblastos leucémicos tienen un crecimiento exagerado y descontrolado, no logran una respuesta inmune normal y causan una disminución en la producción de células normales de la médula ósea que conduce a una deficiencia de glóbulos rojos circulantes (anemia), plaquetas (trombocitopenia) y glóbulos blancos distintos de linfocitos (especialmente neutrófilos o neutropenia) (Wartenberg et al., 2014).

  La leucemia Linfoblástica Aguda representa el 75 % de las leucemias infantiles, tambien pueden ser llamados como ?Neoplasia de células precursoras linfoides? o ?Leucemia B o T linfoblástica? (Carroll and Bhatla, 2016). En Estados Unidos se diagnostican entre 2500 y 3000 niños al año, con una incidencia máxima en niños entre los 2 y 5 años de edad (Carroll and Bhatla, 2016).

  La etiología de la leucemia aguda es desconocida. Los siguientes factores son importantes en la patogenia de la leucemia: Radiación ionizante, productos químicos (por ejemplo, benceno en AML), los medicamentos (p. Ej., El uso de agentes alquilantes, ya sea solo o en combinación con radioterapia, aumenta el riesgo de AML). En las consideraciones genéticas se tiene que si un gemelo identico desarrolla leucemia durante los primeros 5 años de vida, el riesgo de desarrollar la segunda leucemia gemelar es del 20 %. En las anomalías cromosómicas se logra medir el riego, en la Tabla (**2-1**) se presentan las características para medir tal riesgo.

  La mayoría de los casos de leucemia no provienen de una predisposición genética heredada, sino de alteraciones genéticas somáticas. Sin embargo, estudios recientes indican un posible vínculo genético con polimorfismos heredados en los genes ARID5B e IKZF1 para la LLA infantil (Carroll and Bhatla, 2016).

Tabla 2-1: Riesgo en las Anomalías cromosómicas (Carroll and Bhatla, 2016)

Grupo	Riesgo	Intervalo de tiempo
Trisomía 21 (síndrome de Down)	1 en 95	10 años de edad
Síndrome de Bloom	1 en 8	30 años de edad
Anemia de Fanconi	1 en 12	16 años de edad

Para diagnosticar la LLA, se realizan estudios de laboratorio (Conteo sanguíneo, Recuento de Leucocitos, Frotis de sangre y Trombocitopenia) (Carroll and Bhatla, 2016), además se realiza examen de médula ósea el cual sirve para caracterizar las células blásticas y determinar el grado de reducción de la eritro, mielo y trombopoyesis normales, así como de la hiper- o hipocelularidad (Imbach et al., 2011b). El distintivo del diagnóstico de leucemia aguda es la célula blástica, una célula relativamente indiferenciada con cromatina nuclear distribuida de forma difusa, uno o más nucleolos (más prominente en AML) y citoplasma escaso (más abundante en AML). Los estudios especiales de la médula ósea, que ayudan en la clasificación celular detallada como los blastos de linaje B o T (Carroll and Bhatla, 2016).

Para lograr diferenciar el linaje de los blastos se usa un panel de anticuerpos para establecer el diagnóstico de leucemia y para distinguir entre los subclones inmunológicos. El panel debe incluir al menos un marcador que sea altamente específico del linaje, por ejemplo, CD19 para linaje B, CD3 citoplásmico para linaje T y mieloperoxidasa o marcadores de diferenciación monocítica como esterasa no específica, CD11c, CD14, CD64, lisozima para neoplasias de linaje mieloide. Además, el uso de CD79a citoplasmático, CD22 citoplásmico, CD10 para linaje B, CD3 superficial, CD7 y CD5 para el linaje T y CD13 y CD33 para células mieloides puede ser útil para diferenciar inmunofenotipos poco claros (Carroll and Bhatla, 2016).

La célula T representa el 15-20 % de todos los casos, este subtipo está asociado con un alto recuento inicial de glóbulos blancos, presencia de enfermedad extramedular (masa mediastínica). La célula B-precursora representa el 80 % de todos los casos. La célula B madura representa el 1-2 % de TODOS los casos. Estos son inmunoglobulina de superficie positiva y se tratan como linfoma de Burkitt. El pronóstico es similar a otros subtipos de LLA de alto riesgo (Carroll and Bhatla, 2016).

La supervivencia de la leucemia infantil es mucho mejor que la de los adultos, con más de las tres cuartas partes de todos los niños con leucemia y más de las cuatro quintas partes de todos los casos que sobreviven al menos 5 años después del diagnóstico

(Wartenberg et al., 2014).

- **Linfoma de Burkitt:**

  El linfoma de Burkitt pertenece a un 50 % de los subtipos principales del linfoma no Hodgkit pediátrico (**?**Imbach et al., 2011a). La morfología de esta enfermedad es de células grandes vacuoladas con cromatina nuclear fina, dos a cinco nucleolos, citoplasma basófilo, morfología L3 (Imbach et al., 2011a).

  El linfoma de Burkitt no endémico es un tumor de frecuente localización abdominal, especialmente en niños y jóvenes, Entre los tumores abdominales más frecuentes se encuentra el linfoma de Burkitt (Pinilla et al., 2009). Los niños que desarrollan el Linfoma de Burkitt en áreas endémicas del mundo a menudo tienen una masa en la región de la cabeza o el cuello (especialmente la mandíbula) en contraste con la presentación abdominal típica del Linfoma de Burkitt no endémica. Tanto los casos endémicos como esporádicos del Linfoma de Burkitt tienen las mismas translocaciones cromosómicas.

  El linfoma de Burkitt se divide en esporádica y endémica, en la Tabla (**2-2**) se encuentra el diagnóstico diferencial entre estos dos tipos.

Tabla 2-2: Diagnóstico diferencial entre el Linfoma de Burkitt (Imbach et al., 2011a)

Linfoma de Burkitt esporádico	**Linfoma de Burkitt endémico**
* Abdomen (25 %) con ascitis y derrame pleural	* Área mandibular en 70 % de los niños menores de 5 años y en el 25 % de los niños mayores de 14 años
* Área faríngea y retrofaríngea, incluidos los senos paranasales	* La mayor frecuencia del tumos se encuentra en el abdomen
* Implicación de médula ósea en un 20-40 %	* Frecuencia de compromiso de la médula ósea en un 8 %
* La participación del Sistema Nervioso Central es rara	* Participación del Sistema Nerviosos Central, nervios craneales y periféricos

- **Meduloblastomas:**

  El meduloblastoma es el tumor del SNC más común en los niños, representa aproximadamente el 20 % de todos los tumores cerebrales infantiles, y el 80 % de los casos se presenta antes de los 15 años. El tumor se presenta en la fosa posterior y puede producirse una diseminación generalizada del espacio subaracnoideo. La frecuencia de propagación del SNC fuera del tumor primario puede ser tan alta como 40 % en el momento del diagnóstico (Hanson and Atlas, 2016). La edad media de detectar medulloblastoma es de 4 a 8 años, donde se presenta con mayor frecuencia en

niños que en niñas (Imbach et al., 2011c).

Los estudios de estadificación deben incluir resonancia magnética de la columna vertebral (preferiblemente preoperatoriamente), citología de la recolección de líquido cefalorraquídeo (LCR) lumbar, pruebas de función hepática y, si es clínicamente sintomático, exploración ósea y/o examen de médula ósea. La histología y la citogenética del tumor original son esenciales para evaluar el subtipo anaplásico de células grandes (Hanson and Atlas, 2016).

Los pacientes se dividen en categorías de riesgo medio y alto en función de la extensión de la enfermedad, el volumen de tumor residual, la histología y la edad al momento del diagnóstico (Tabla **2-3**).

Tabla 2-3: Categorias de riesgo de meduloblastoma (Hanson and Atlas, 2016)

	Riesgo promedio	**Riesgo Alto**
Alcance de la enfermedad	Citología negativa del LCR Resonancia magnética normal de la columna vertebral	Citología Positiva del LCR Resonancia magnética positiva con gadolinio en la columna vertebral
Alcance del tumos residual (en una medida bidimensional)	$< 1{,}5cm^2$ residual	$> 1{,}5cm^2$ residual
Histología	Indiferenciado	Anaplástico de células grandes
Edad en el momento del diagnóstico	>3 años	<3 años

La supervivencia de los pacientes con meduloblastoma de riesgo promedio tratados con radiación adyuvante y quimioterapia es de al menos 82 % a los 5 años. En pacientes de alto riesgo que usan radiación craneoespinal con quimioterapia, 45 a 50 % de los pacientes no tienen enfermedad a los 5 años (Hanson and Atlas, 2016).

Estudios Ecológicos en Epidemiológia

Los estudios epidemiológicos clásicamente se dividen en Experimentales y No experimentales. En los estudios experimentales se produce una manipulación de una exposición determinada en un grupo de individuos que se compara con otro grupo en el que no se intervino, o al que se expone a otra intervención. Cuando el experimento no es posible se diseñan estudios no experimentales que simulan de alguna forma el experimento que no se ha podido realizar, entre estos se encuentran los estudios ecológicos (Fernández, 1995). Los estudios ecológicos se distinguen de otros diseños en su unidad de observación, pues se caracterizan por estudiar grupos, más que individuos por separado. Frecuentemente se les denomina estudios generadores de hipótesis o diseños incompletos debido a que, por emplear promedios grupales, se les desconoce la distribución conjunta de las características en estudio

a nivel de cada individuo. Comúnmente las unidades de observación son diferentes áreas geográficas o diferentes periodos de tiempo en una misma área, a partir de las cuales se comparan las tasas de enfermedad y algunas otras características del grupo. La principal motivación para desarrollar estos estudios es la fácil disponibilidad de los datos, ya que generalmente se utilizan datos registrados rutinariamente con propósitos administrativos o legales (Borja-Aburto, 2000), y a su vez se clasifican como exploratorios, de grupos múltiples, de series de tiempo y mixtos.

Tipos de estudios ecológicos

Según Borja-Aburto (2000) se clasifican en:

- **Estudios exploratorios:** Tiene como objetivo comparar tasas o frecuencia de las enfermedad o eventos de interés entre muchas zonas continuas en el mismo período. En estos estudios no se hace comparación formal con otras variables, el propósito es buscar patrones espaciales o temporales que puedan dar indicios para las hipótesis sobre las causas.

- **Estudios de grupos múltiples:** Son estudios analíticos que evalúan la asociación entre los niveles de exposición promedio y la frecuencia de la enfermedad o eventos de interés entre varios grupos. Generalmente los datos provienen de estadísticas de morbilidad y mortalidad.

- **Estudios de series de tiempo:** Comparan variaciones temporales de los niveles de exposición con otra serie de tiempo que refleja los cambios en el frecuencia del evento en la población de una zona. La inferencia causal de este tipo de análisis puede limitarse debido a dificultades de los periodos de observación entre la exposición y los efectos, de la medición de la exposición.

- **Estudios mixtos:** Estos estudios son una combinación de los estudios de series de tiempo y los de grupos múltiples.

La principal limitación de estos estudios es que al no poder determinar si existe una asociación entre una exposición y una enfermedad a nivel individual, se genera una *"falacia ecológica"*, la cual consiste en obtener conclusiones inadecuadas a nivel individual, ya que se encuentran basadas en datos poblacionales (Borja-Aburto, 2000). Igual que otros estudios observacionales en epidemiología puede existir una tercera variable que explique la relación entre la enfermedad y la exposición objeto de estudio.

Tasa

Una tasa puede ser definida como una medida del cambio que expresa una cantidad y por cada unidad de otra cantidad x, de la cual y es dependiente. Por lo tanto, si $y = y(x)$ y

$\triangle y = y(x + \triangle x) - y(x)$, entonces la tasa promedio de cambio es $\triangle y / \triangle x$ (es decir, el cambio promedio de y por unidad de x en el intervalo $(x, x + \triangle x)$(Elandt-Johnson, 1997).

Puesto que x suele ser la medida del tiempo y la cantidad y describe un proceso continuo a lo largo del tiempo, entonces $\frac{\triangle y}{\triangle x}$ es la velocidad promedio que corresponde a ese proceso. El resultado puede ser positivo o negativo, dependiendo de que y se incremente o disminuya a lo largo del tiempo.

En el área de epidemiología, Las estimaciones de las tasas de incidencia de las enfermedades son de utilidad para las autoridades en el momento de realizar asignación de recursos (Granados, 1994). Las tasas de incidencia se miden con los nuevos casos en un rango de tiempo y la población que se encontró en riesgo de sufrir la enfermedad (ecuación 2-1), mostrando de esta forma la dinámica de ocurrencia de un determinado evento en una población dada (Granados, 1994).

$$Tasa = \frac{\#\ de\ casos\ nuevos\ en\ x\ tiempo}{\#\ de\ personas\ que\ tuvieron\ el\ riesgo\ de\ sufrir\ la\ enfermedad} * 1?000{,}000 \quad (2\text{-}1)$$

Como se muestra en la ecuación 2-1, la tasa se encuentra estandarizada por 1?000,000 de niños, ya que el cáncer en niños menores de 15 años es muy raro de encontrar y por lo tanto de medir.

Densidad de Población

La densidad de Población se refiere al número promedio de habitantes de un área urbana o rural en relación a una unidad de superficie dada, es decir, mide el número de habitantes que viven por kilómetro cuadrado a través de la siguiente formula (Densidad de población, 2017):

$$indice = \frac{numero\ de\ habitantes}{superficie} \quad (2\text{-}2)$$

Indicando que existe una cantidad X de habitantes por kilómetro cuadrado, este tipo de indicadores suele utilizarse en epidemiologia y economía, ya que a partir de este indicador se puede medir los recursos y la infraestructura de una área urbana o rural.

2.2. Marco teórico del análisis espacial

2.2.1. 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 sitos (puntos del espacio) de

una región; se puede decir que la estadística espacial es el análisis de realizaciones de un proceso estocástico $[Z_{(s)} : s \epsilon D]$; donde s pertenece a R^d y representa una ubicación en el espacio euclidiano dimensional, $Z_{(s)}$ es una variable aleatoria en la ubicación s que varia sobre un conjunto de indices $D \subset R^d$ (Giraldo, 2002).

Giraldo (2002) muestra que la estadística espacial se divide en tres grandes áreas de estudio:

- **Geoestadística:** Estudia los datos donde las ubicaciones s provienen de un conjunto D pertenece a R^d continuo, pero es el investigador quien selecciona en que sitios del área de estudio realiza la medición de las variables bajo algún esquema de muestreo probabilístico (D fijo). La geoestadística tiene como propósito la interpolación; si no existe continuidad espacial puede hacerse predicciones carentes de sentido, teniendo en cuenta que las mediciones se encuentran georeferenciadas.
- **Enmallados:** Estudia los datos donde las ubicaciones s pertenecen a un conjunto D discreto y al igual que la geostadística son seleccionadas por el investigador, estos datos pueden encontrarse de manera regular o irregularmente espaciadas. Esta técnica corresponde a agregaciones espaciales mas que aun conjunto de puntos del espacio.
- **Patrones Puntuales:** A diferencia de las dos técnicas anteriores, este maneja el conjunto D que puede ser discreto o continuo (aleatorio), por lo tanto el investigador no puede tomar la decisión de seleccionar en que sitios del área de estudio realizar las mediciones.

2.2.2. Patrones puntuales

Esta área se diferencia de la Geoestadistica y de los Lattices, ya que pertenece a un conjunto del dominio D que puede ser discreto o continuo y la selección del lugar para realizar las medidas no dependen del investigador, es decir que dicho conjunto D puede ser aleatorio, pero los lugares donde ocurre el fenómeno de interés está dado por la naturaleza; luego de la selección del sitio es posible hacer medidas de variables aleatorias en cada uno de ellos (Giraldo, 2002).

Algunos ejemplos de patrones espaciales son la ubicación de nidos de pájaro en una región dada, ubicación de los sitios de terremoto en Colombia, la ubicación de árboles de pino dentro de un bosque o los cuadrantes de una región con presencia de una especie en particular. En general el propósito de análisis en estos casos es el de determinar si la distribución de los individuos dentro de la región es aleatoria, agrupada o uniformes como se logra ver en la figura **2-1** (Giraldo, 2002).

Formalmente un proceso puntual es un arreglo (patrón) de puntos de un conjunto aleatorio D. Cuando se tienen solo parte de los eventos, el patrón se llama patrón muestreado (sampled

point mapped). Cuando todos los eventos de la realización se registran, entonces se dice que este es puntual (mapped point patterns) (Schabenberger and Gotway, 2005). .

Tipos de Patrones Espaciales

Muchos eventos puede ser independientes en sub-regiones que no se traslapan, pero no necesariamente la intensidad $\lambda(s)$ es homogénea en el conjunto D. Otros eventos pueden estar localizados en regiones donde $\lambda(s)$ es grande, y unos pocos en regiones donde la intensidad en mínima. Los eventos puede tener una intensidad constante o promedio, $\lambda(s) = \lambda$, pero tener algún tipo de interacción. La presencia de un evento puede atraer o alejar otros eventos cercanos. Debido a esto, las derivaciones de un patrón espacial son completamente aleatorio, agregados (agrupados) o regulares, como se muestra en la Figura (**2-1**).

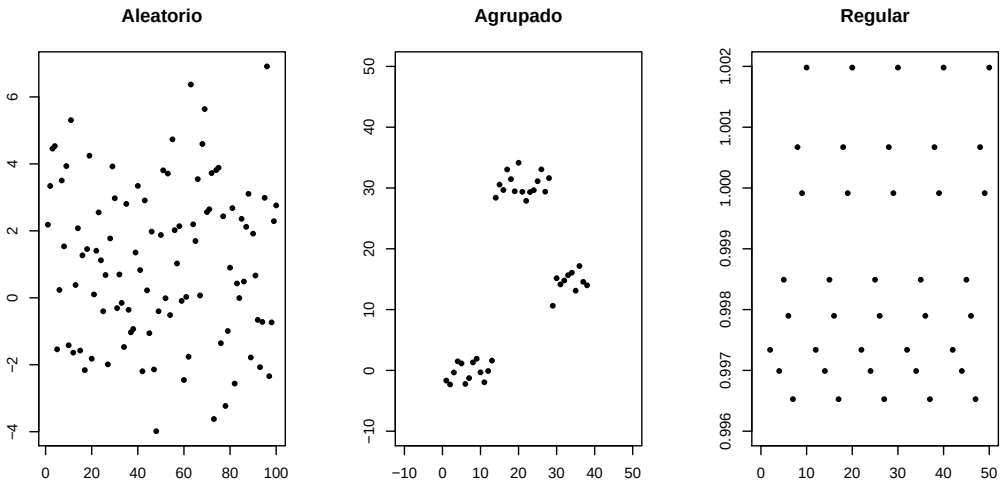

Figura 2-1: Tipos de patrones espaciales

La razón de esta diferenciación entre los tipos de patrones espaciales puede ser la variación determinística espacial de la función de intensidad $\lambda(s)$ o los resultados de un elemento estocástico. Las caracteristics que tienen los diferentes patrones puntuales son (Schabenberger and Gotway, 2005):

- **Patrón aleatorio:** El promedio de los eventos por unidad de área es homogénea en el conjunto D, el número de eventos de dos sub-regiones que no se solapan A_1 y A_2 son independientes, y el número de eventos en cualquier sub-región sigue una distribución de Poisson.

- **Patrón Regular:** La distancia promedio entre los eventos y su vecino mas cercano es más pequeña que la misma distancia media en un patrón espacial aleatorio.
- **Patrón Agrupado:** La distancia media entre un evento y su vecino más cercano es más grande que la esperada bajo aleatoriedad.

2.2.3. Test de completa aleatoriedad espacial

El test de completa aleatoriedad espacial, aborda si el patrón de puntos observados podría ser realizado de un proceso Poisson homogéneo o para un proceso binomial con n fijo; al igual que las propiedades estocásticas estos pueden ser descritos a través de puntos de conteo o medidas, ademas pueden estar basados en los métodos de recuento de cuadrantes y en los métodos basados en la distancia (Schabenberger and Gotway, 2005).

Dentro un proceso Poisson homogéneo, se puede considerar como una variable el número de eventos en una región A, teniendo en cuenta que las regiones no se encuentran solapadas, se garantiza de esta forma que las variables sean independientes. La distribución del estadístico de prueba basados en conteos de cuadrantes, se conoce que es una distribución asintotica, garantizando ser un test preciso. Pero para dominios espaciales irregulares, debido a los efectos de borde y cuadrantes pequeños, las estimaciones por medio de este método pueden no funcionar bien. En caso de que se aplique estadísticas basada en la distancia entre los eventos y sean muestras intratables, se puede usar los métodos de simulación de análisis de patrones por medio de las herramientas básicas como la prueba de Monte Carlo y el examen de simulaciones de patrones (Schabenberger and Gotway, 2005).

Métodos basados en Cuadrantes

Uno de los métodos para probar la aletoriedad espacial de un patrón son basadas en la división del dominio D en sub-regiones no traslapadas (cuadrantes), siendo $A_1, ..., A_k$ de igual tamaño. Matemáticamente sea $A_1, ..., A_K$ son tal que $U_{i=1}^{k} A_i = D$ donde A_i y A_i' no se troslapan.

Test de bondad de ajuste x^2

La prueba de ajuste estadístico x^2, tiene como hipótesis nula: que los n puntos son distribuidos uniformemente independientes a lo largo de D, es decir, los conteos de sitios por cuadrante son variables Poisson independientes con media común.

n : numero total de eventos.

n_{ij} : numero total de eventos en la cuadricula ij.

$\bar{n}$: promedio de eventos totales.

El estadístico de prueba en este caso es:

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-\bar{n})^2}{\bar{n}} \sim \chi^2_{rc} \tag{2-3}$$

donde r es el numero de filas y c el numero de columnas en las que se encuentra divida la cuadricula y sea χ^2 una distribución chi-cuadrado con $r * c$ grados de libertad.

Test de Monte Carlo

La prueba de Monte Carlo funciona como un caso especial de la prueba chi-cuadrado. En esta prueba se simulan muestras de patrones de puntos aleatorios a partir de la hipótesis nula, en la que se contrasta la posibiidad de que se presenten patrones aleatorios.

Ademas la prueba de Monte Carlo tiene numerosas ventajas, ya que los valores p de las pruebas son exactas, en el sentido de que no se requiere ninguna aproximación de la distribución teórica del estadístico; los valores p son inexactos cuando el número de posibles realizaciones de los puntos es infinita, es decir que el tamaño de los puntos sea desconocido y se simule una gran cantidad de puntos. Por lo que se recomienda un numero grande de simulaciones para obtener mayor precisión en las las pruebas (Schabenberger and Gotway, 2005).

Método basados en distancias:

La elección de el número de cuadrantes y la forma para probar la hipótesis de aleatoriedad basados en conteos por áreas puede influenciar los resultados. Las pruebas estadísticas basadas en distancias entre eventos o entre puntos muestreados y eventos, elimina la influencia del área de los cuadrantes sobre la decisión respecto a la hipótesis (Schabenberger and Gotway, 2005). Entre las pruebas basadas en distancias se encuentran:

Función G

La función *G(h)* tiene en cuenta la mínima distancia entre eventos (distancia al vecino más cercano), en la cual se plantea la función de distribución empírica:

$$\hat{G(D)} = \frac{N(d_i \leq d)}{n} \tag{2-4}$$

Donde n es el numero de eventos presentados, d_i es la mínima distancia entre un evento y sus vecinos, D es la variable aleatoria de distancia al lugar del evento más cercano y sea $G(D)$ la función de distribución de D.

La distribución teórica de $G(D)$ surge de $G(d) = P(D \leq d) = 1 - P(D > d) = 1 - P(N(A) = 0)$, donde $N(A)$ se encuentra bajo aleatoriedad.

Si $N(A) \sim Poisson(\lambda\pi d^2)$ entonces $P(N(A) = 0) = e^{-\lambda\pi d^2}$, por lo tanto:

$$\hat{G(d)} = 1 - e^{-\lambda\pi d^2} \tag{2-5}$$

Donde el patrón es agregado cuando d es pequeño y la distancia comienza a crecer entonces la función $\hat{G(d)}$ crece rápidamente, hasta que llega a una distancia donde la diferencia en $\hat{G(d)}$ deja de ser significativa. En cambio cuando el patrón es regularmente espaciado la distancia d entre los eventos no tiene diferencia significativa, por lo tanto la función $\hat{G(d)}$ crece lentamente.

Función F

Otra función planteada es la $F(h)$, la cual cuenta con la mínima distancia punto-evento, donde un punto es un suceso escogido aleatoriamente dentro de la región de estudio y un evento es el caso estudiado dentro de la investigación. Para la construcción se siguen los siguientes pasos:

1. Se selecciona aleatoriamente m puntos $(p_1, ..., p_m)$
2. Se calcula $d_i = d(p_i, s_i)$, la distancia de cada punto escogido al sitio del evento más cercano.

Sea D la variable aleatoria de mínima distancia de un punto a un evento, entonces se tiene que la función de distribución empírica es la que se muestra en la ecuación (2-6) y la distribución teórica es la que se muestra en la ecuación (2-7).

$$\hat{F(D)} = \frac{N(d_i \leq d)}{n} \tag{2-6}$$

$$\hat{F(d)} = 1 - e^{-\lambda \pi d^2} \tag{2-7}$$

Donde n es el numero de eventos presentados, d_i es la distancia mínima entre un punto y un evento, D es la variable aleatoria de distancia al punto más cercano.

Cuando los procesos de puntos se estudian a través de medidas de recuento, como el número de eventos en la región A ($N(A)$) siendo $N(A)$ una variable aleatoria. La función $F(d)$ puede ser encontrada a partir de la función de masa de $N(A)$, donde hay una función $\lambda(s)$ cuya integración sobre A da su valor esperado, por lo que anteriormente esta función se denomina intensidad de primer orden y se define como limite de la forma de la ecuación (2-7) (Schabenberger and Gotway, 2005).

Si el patrón es agregado, indica que la distancia entre los punto y los eventos es pequeño y luego comienza a crecer, la función $\hat{F(d)}$ crece lentamente. En cambio si la función $\hat{F(d)}$ crece rápidamente, quiere decir que la distancia entre los punto y los eventos es muy grande (Schabenberger and Gotway, 2005).

Función *K* de Ripley

La función K de Ripley es una función de segundo orden para procesos estacionarios e isotropicos. También se conoce como la medida reducida del segundo momento o segunda función de momento reducido (Cressie, 2015) de un proceso de punto estacionario d. Además se define de manera que $\lambda K(d) = E(M)$, es igual al número esperado de puntos aleatorios adicionales dentro de una distancia d de un punto aleatorio típico de d. Aquí λ es la intensidad del proceso, es decir, el número esperado de puntos d por unidad de área. En los análisis exploratorios, la estimación de K es una estadística útil para resumir los aspectos de la ínter-punto, es decir, la distancia de cada punto a todos los otros registros, la cual se encuentra definida por:

$$K(d) = \frac{E(M)}{\lambda} \tag{2-8}$$

Donde $E(M)$ es la esperanza del número de eventos adicionales a una distancia d de un evento elegido aleatoriamente y λ es la intensidad.

$$\hat{K(d)} = \frac{\sum \sum_{i \neq j} I_d(d_{ij})}{\hat{\lambda_n}} \tag{2-9}$$

Donde $\hat{\lambda_n} = \frac{n}{A}$, A es el área de la región R y d_{ij} es la distancia entre el *i-ésimo* y el *j-ésimo* evento.

Bajo aleatoriedad se tiene que:

$$K(d) = \frac{\lambda \pi d^2}{\lambda} = \pi d^2 \tag{2-10}$$

Donde si $K(d) < \pi d^2$ se dice que hay regularidad y si $K(d) > \pi d^2$ hay agregación.

Función *K* de *L*

Otra función de segundo orden implementada en el estudio de dependencia en patrones puntuales basada en la función $k(d)$, es llamada la función L, la cual esta dada por:

$$L(d) = \sqrt{\frac{K(d)}{\pi}} \text{ Bajo aleatoriedad } L(d) = \sqrt{\frac{\pi d^2}{\pi}} = d \tag{2-11}$$

si el patrón es aleatorio no debe haber desviación de la recta construida a partir de la transformación de la función $K(L)$ (Schabenberger and Gotway, 2005).

2.2.4. Función de intensidad

La función de intensidad es un factor decisivo en el desarrollo del análisis espacial de un área determinada, ya que es vital el número promedio de eventos ocurridos por unidad de área(λ), y éste se puede definir de manera matemática para áreas pequeñas (ecuación 2-12) y para patrones aleatorios en la ecuación (2-13).

$$\lambda(s) = \lim_{|s| \to \infty} \frac{E[N(s)]}{|s|} \tag{2-12}$$

$$\lambda(s) = \frac{n}{|A|} = \lambda \tag{2-13}$$

A través de $E[N(s)]$ se elimina la dependencia del tamaño y la forma de la zona s. Este método es útil para estudiar la intensidad de los eventos más localmente, por ejemplo, para determinar si se debe proceder con un análisis del comportamiento de segundo orden. En la practica, las estimaciones de variables espaciales $\hat{\lambda(s)}$ de la intensidad en la ubicación s se puede obtener por medio de suavización no paramétrica de los recuentos de cuadrantes o por métodos de estimación de la densidad (Schabenberger and Gotway, 2005).

Estimación Kernel de la función de intensidad

La función Kernel univariada necesita ser reemplazada por una función que se acomode a las coordenadas, la cual se puede realizar por medio de la multiplicación de las funciones Kernel, este proceso se puede usar por conveniencia, ya que puede existir ausencia de la interacción entre coordenadas. Los anchos de banda utilizados en esta estimación pueden ser elegidos por el investigador y pueden ser diferentes a dos dimensiones, pero las estimaciones más comunes son los Kernel esféricos y Kernel gausianos (Schabenberger and Gotway, 2005). Un patrón espacial completamente aleatorio tiene:

$$\hat{\lambda} = \frac{n}{|A|} = \frac{\text{número de eventos}}{area\ de\ s} \tag{2-14}$$

Si x_i y y_i son las coordenadas de localización del sitio $s_i(x_i, y_i)$, entonces el producto kernel conlleva a un estimador de la intensidad dado por:

$$\hat{\lambda}(s) = \frac{1}{|A|h_x h_y}\sum_{i=1}^{n} K(\frac{x-x_i}{h_x})K(\frac{y-y_i}{h_y}) = \frac{1}{|A|}\sum_{i=1}^{n} K(\frac{s-s_i}{n}) \tag{2-15}$$

Donde se tiene que h_x y h_y son los anchos de banda en las direcciones respectivas del sistema de coordenadas. Por otra parte, si se presentan problemas con el efecto de borde $\delta(S)$, se obtiene un estimador que incluya el efecto de borde, como se muestra en la ecuación (2-16).

$$\hat{\lambda(s)} = \frac{1}{\delta(s)}\sum_{i=1}^{n} K\left(\frac{s-s_i}{n}\right) \quad \text{Donde} \quad \delta(s) = \int_A \frac{1}{h^2} K\left(\tfrac{s-u}{n}\right) du \tag{2-16}$$

La estimación de la intensidad mediante funciones Kernel es muy útil, pero los casos de cáncer infantil en la ciudad de Santiago de Cali se da en diferentes cuadras de la ciudad, por lo tanto el espacio que éste maneja no es continuo, existiendo lugares de la ciudad en los cuales no se encuentran casos con este tipo de enfermedades. Debido a esto, como alternativa de análisis se propone encontrar el ancho de banda óptimo para la estimación de la función de la intensidad, esto a pesar de no solucionar el problema, reduce el sesgo en la estimación de la función de intensidad (Espinal and Aruneri, 2014). Para determinar el ancho de banda adecuado debido a las condiciones especiales de este estudio se propone utilizar el criterio de validación cruzada, que a su vez se puede realizar mediante los criterios:

- **selección del ancho de banda de la densidad Kernel con validación cruzada:** El ancho de banda σ se escoge para minimizar el criterio de error cuadrático medio definido por Diggle (1985). El algoritmo calcula el error cuadrático medio por el método de Berman y Diggle (1989).

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

donde $MSE(\sigma)$ es el error cuadratico medio en el ancho de banda σ, $g(0)$ es la función de correlación de pares y λ es la intensidad media.

- **selección del ancho de banda para la densidad Kerner con a verosimilitud de la valización cruzada:** El ancho de banda σ se elige para maximizar la probabilidad del proceso puntual con el criterio de validación cruzada.

$$LCV(\sigma) = \sum_i log(\lambda_i(x_i)) - \int_W \lambda(u)du$$

donde se toma la suma sobre todos los puntos de datos x_i, y $\lambda_i(x_i)$ es la estimación Kernel suavizado, dejando uno fuera de la intensidad x_i con ancho de banda σ, $\lambda(u)$ es la estimación Kernel suavizada de la intensidad en una localización espacial u con suavizado σ (Loader, 1999)

2.3. Modelación

En la fase de la modelación de las tasas de casos de cáncer infantil, se busca indagar la asociación de las variables disponibles como las sociodemograficas de cada paciente y que pertenecen al área de Santiago de Cali. Para esta parte se va ajustar un Modelo Lineal Generalizado. Para identificar el tipo de distribución de la variable de respuesta se utilizó pruebas de bondad de ajuste.

2.3.1. Prueba Anderson Darling para Bondad de Ajuste

Esta prueba fue desarrollada en 1952 por Theodoro Anderson y Donald Darling (Anderson and Darling, 1954) para identificar si un conjunto de datos específico proviene de cierta distribución de probabilidad. Esta prueba es un caso especial de las pruebas de Kolmogorov-Smirnov y Cramer Von Mises (Marsaglia et al., 2004). La hipótesis es:

H_0 : Los datos siguen una distribución específica
H_a : Los datos no siguen una distribución específica

$$H_0 : F_n(x) = F(x) \qquad H_1 : F_n(x) \neq F(x) \tag{2-17}$$

Siendo $F_n(x)$ la función de distribución empírica (los datos siguen la distribución especificada) y $F(x)$ es la función de densidad teórica . El estadístico de prueba corresponde

a la ecuación (2-18), donde n es el número de datos y k es un parámetro que ayuda a la secuencia de la sumatoria.

$$A_n = -n - \sum_{k=1}^{n} \frac{2k-1}{n}[ln(F(x_k)) + ln(1 - F(x_{n-1+k})] \tag{2-18}$$

De aquí se tiene que la prueba Anderson-Darling se basa en la medición de las diferencias de la distribución entre los valores críticos de $F_n(x)$ y $F(x)$, como se muestra a continuación:

$$n \int_{-\infty}^{\infty} (F_n(x) - F)^2 w(x) dF(x) \qquad w(x) = [F(x)(1 - F(x))]^{-1} \tag{2-19}$$

Los valores críticos de la distribución normal (z) se encuentran tabulados o en los softwares implementados. El *valor-p* se calcula como:

$$valor - p = Prob(A < z) \tag{2-20}$$

Si el *valor-p* es menor que 0.05 se rechaza la hipótesis nula, es decir, los datos no se distribuyen $F(X)$ con un $\alpha = 0{,}05$ de significancia. Una vez identificada la distribución de referencia se plantea el modelo lineal generalizado para el tipo de cáncer infantil.

2.3.2. Modelos Lineales Generalizados

Nelder and Baker (1972), crearon la idea de Modelo Lineal Generalizado (GLM de las siglas en inglés de Generalized Linear Models). Tanto el modelo lineal generalizado como el modelo lineal ajustado por mínimos cuadrados ordinarios son herramientas metodológicas que permiten codificar todas las situaciones de análisis dentro de un mismo esquema general. El GLM se define en términos de un conjunto de variables aleatorias independientes $Y_1, Y_2, ..., Y_n$ y de un componente sistemático, en el cual, el componente aleatorio proviene de las distribuciones pertenecientes a la familia exponencial y cumplen con dos propiedades (Dobson and Barnett, 2008):

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

   $$f(y_i, \theta_i) = exp[a(y_i)b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)] \tag{2-21}$$

   Donde $a(y_i)$, $b_i(\theta_i)$, $c_i(\theta_i)$, y $d_i(y_i)$ son funciones conocidas, y $b_i(\theta_i)$ el parámetro natural de la distribución o función canónica.

2. La distribución de todo Y_i es de la misma distribución, de modo que no son necesarios los subíndices en *b, c y d*. Por lo tanto, la función de densidad de probabilidad conjunta de $Y_1, Y_2, ..., Y_n$ es:

$$f(y_1, ..., y_n) = \prod_{i=1}^{n} exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)] = exp[\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)] \tag{2-22}$$

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

Identificando que $Y_i \sim N(\mu, \sigma)$ siendo $\mu_i = E(Y_i)$. Donde el efecto marginal de X_k viene dado por: $\frac{\partial E(Y_i)}{\partial X_{ki}} = \beta_k \; \forall i = 1, ...n$, el significado de las marginales está dado a las unidades de medida de cada variable explicativa (Cayuela et al., 2016). Entonces para un modelo lineal generalizado se tiene una transformación para μ_i de la forma (Dobson and Barnett, 2008):

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

De la ecuación anterior se tiene que $g(.)$ es monótona y diferenciable también llamada **función de enlace**, ya que ésta ayuda a que las predicciones del modelo queden acotadas, ademas el vector de variables explicativas $\vec{x_i}$ puede ser de covariables o niveles de factores con dimensión $p * 1$ ($\vec{x_i^T} = [x_{i1}, ..., x_{ip}]$), también el vector de parámetros $\vec{\beta}$ es de dimensión $p * 1$ ($\vec{\beta^T} = [\beta_1, ..., \beta_p]$).

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

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

Lo que hace la función de vínculo es básicamente transformar la variable respuesta de modo similar a cómo se haría en una regresión cuando tenemos problemas de linealidad, pero teniendo en cuenta los valores estimados por el modelo mediante la transformación inversa de la función de vínculo. Otra de las utilidades, es la de conseguir que las predicciones de nuestro modelo queden acotadas. Por ejemplo, si tenemos datos de conteo, no tiene sentido que nuestras predicciones arrojen resultados negativos, otro ejemplo, si la variable respuesta es una proporción, entonces los valores estimados tienen que estar entre 0 y 1 o 0 y 100 (valores por debajo de 0 o por encima de 1 o 100 no tienen ningún sentido) (Cayuela, 2010).

Las funciones de enlace son diferentes dependiendo del caso. Algunas de estas se muestran en la Tabla (**2-4**).

Tabla 2-4: Funciones de vínculo más comunes utilizadas por los GLM (Cayuela, 2010)

Función de vínculo	**Fórmula**	**Uso**
Identidad	μ	Datos continuos con errores normales (regresión y ANOVA)
Logarítmica	$Log(\mu)$	Conteos con errores de tipo Poisson
Logit	$Log\left(\frac{\mu}{n-\mu}\right)$	Proporciones (datos entre 0 y 1) con errores binomiales
Recíproca	$\frac{1}{\mu}$	Datos continuos con errores gamma
Raíz cuadrada	$\sqrt{\mu}$	Conteos
Exponencial	μ^n	Funciones de potencia

Parte del trabajo de construcción y evaluación del modelo, es determinar cuál de todos estos modelos son adecuados, y entre todos los modelos adecuados, cuál es el que explica la mayor proporción de la varianza, sujeto a la restricción de que todos los parámetros del modelo deberían ser estadísticamente significativos (Cayuela, 2010).

2.3.3. Modelo de Regresión Lineal Generalizado con Variable de Respuesta Gamma

Los modelos lineales generalizados (GLM) son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores (binomiales, Poisson, gamma, etc) y varianzas no constantes. Una distribución Gamma, es muy útil con datos que muestran un coeficiente de variación constante, esto es, en donde la varianza aumenta según aumenta la media de la muestra de manera constante (p.e. número de presas comidas por

un predador en función del número de presas disponibles) (Cayuela, 2010).

La distribución Gamma es útil para modelar variables que son estrictamente no negativas, ya que es muy flexible para modelar distintas formas de la variable de respuesta por los dos parámetros que indexan la distribución (Ecuación (2-23)) y no impone el supuesto de homoscedasticidad, aunque sí el de un coeficiente de variación constante (Salinas-Rodríguez et al., 2006). Este enfoque puede ser de utilidad si satisface los supuestos de los errores del modelo que este impone, el modelo gamma debe preferirse si se tienen sólo valores no negativos y la distribución de la variable de respuesta no es simétrica (Salinas-Rodríguez et al., 2006).

Para el modelo lineal generalizado con variable de respuesta Gamma, se asume que la distribución condicional de y_i dado x_i, se distribuye como una variable aleatoria Gamma con función de densidad (ecuación (2-23)), ya que por medio de esta forma se puede identificar más fácilmente la función canónica que sera empleada para la función de vínculo del modelo lineal generalizado.

$$f(y_i|x_i) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}} \tag{2-23}$$

La distribución Gamma, es una distribución adecuada para modelar el comportamiento de variables aleatorias continuas con asimetría positiva. Es decir, variables que presentan una mayor densidad de sucesos a la izquierda de la media que a la derecha. En su expresión se encuentran dos parámetros, siempre positivos, α y β, α determina la forma de la distribución situando la máxima intensidad de probabilidad y β es el parámetro de escala, el cual determina la forma o alcance de la simetría (Arroyo et al., 2014).

2.3.4. Estimación de los parámetros

Teniendo en cuenta que en el modelo lineal generalizado con variable de respuesta Gamma, se pueden utilizar tres funciones de vínculo como: la inversa, la logarítmica o la identidad (Cayuela, 2014); si se utiliza la función identidad, se estaría realizando una estimación por mínimos cuadrados ordinarios, en cambio con las otras funciones de enlace se pueden estimar los parámetros con la teoría clásica de los GLM, donde esta se realizan mediante los métodos de máxima verosimilitud con procedimientos iterativos como los de Newton Raphson o Fisher Scoring, la cuales se describen a continuación.

Maxima Verosimilitud:

Se debe de tener en cuenta que las variables aleatorias $Y_1, ..., Y_n$ se suponen independientes. Entonces el logaritmo natural de la verosimilitud del modelo Gamma para la observación

i-ésima se obtiene de la ecuación (2-24).

$$log[L(\alpha,\beta)] = -nlog[\Gamma(\alpha)] - n\alpha log[\beta] + (\alpha-1)\sum_{i=1}^{n} log[x_i] - \frac{1}{\beta}\sum_{i=1}^{n} x_i \tag{2-24}$$

Y este pertenece a una familia exponencial de la forma:

$$f_y(Y,\theta,\phi) = exp\left\{\frac{y(\theta)-b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Donde a partir de esta ecuación tenemos que $\mu = E(Y) = b'(\theta)$ y $Var(Y) = a(\phi)b''(\theta)$. Ademas si sabemos que $Y_1, ..., Y_n$ son variables aleatorias, donde se desea máximizar la verosimilitud respecto a $\beta = [\beta_1, ..., \beta_p]^T$ entonces se quiere resolver:

$$\frac{\partial}{\partial\beta_j}\ell(\beta, y) = 0 \qquad\qquad \forall j = 1, ..., p \tag{2-25}$$

La función de la ecuación (2-25) no es lineal en β, por lo que las estimaciones máximo verosímiles se deben hallar con un procedimiento iterativo. Para la estimación de los coeficientes de regresión y parámetros de dispersión, se usa el método de Newton-Rapshon o el de Fisher Scoring (Ayati and Abbasi, 2011):.

Newton Raphson y Fisher Scoring

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

$$\beta^{(t+1)} = \beta^{(t)} - [\ell''(\beta^{(t)})]^{-1}\ell'(\beta^{(t)}) \tag{2-26}$$

donde $\ell'(\beta^{(t)} = \frac{\partial_i}{\partial\beta}$, $\ell''(\beta^{(t)}) = \frac{\partial^2{}_i}{\partial\beta_k\partial\beta_j}$. De esta forma, la ecuación (2-26), se descompone de la siguiente manera:

$$\frac{\partial\ell}{\partial\beta_j} = \sum_{i=1}^{n} \frac{Y_i-\mu_i}{V_i} \;\; \frac{\partial\mu_i}{\partial\eta_i} \;\; x_{ij} = 0$$

$$\frac{\partial^2\ell}{\partial\beta_k\partial\beta_j} = \sum_{i=1}^{n} \frac{\partial}{\partial\beta_k}[Y_i-\mu_i]\frac{1}{V_i}\frac{\partial\mu_i}{\partial\eta_i}x_{ij} + \sum_{i=1}^{n}(Y_i-\mu_i)\frac{\partial}{\partial\beta_k}\left[\frac{1}{V_i}\frac{\partial\mu_i}{\partial\eta_i}x_{ij}\right]$$

De Acuerdo a las ecuaciones $V = a(\phi)b''(\theta)$ donde este depende de la función de enlace a utilizar, digamos que se utiliza la función identidad y se obtiene $V = a(\phi)b''(\eta)$. El método de Fisher Scoring se caracteriza por calcular $E\left(\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j}\right)$ que es más preciso y realiza una aproximación a los mínimos cuadrados, por lo tanto esta es igual a:

$$E\left(\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j}\right) = -\sum_{i=1}^{n} V_i^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 x_{ij} x_{ik} \tag{2-27}$$

La ecuación (2-27) la podemos expresar de manera matricial de la forma $X'WX$, donde W se encuentra en la ecuación (2-28), por último realizando un proceso algebraico a la ecuación (2-26) se obtiene la ecuación (2-29):

$$W = diag\left(V_i^{-1}\left[\frac{\partial \mu_i}{\partial \eta_i}\right]^2\right) \tag{2-28}$$

$$\begin{aligned} \beta^{(t+1)} &= \beta^{(t)} + (X'WX)^{-1}X'V^{-1}\frac{\partial \mu}{\partial \eta}(Y - \mu) \\ &= (X'WX)^{-1}X'Wz \end{aligned} \tag{2-29}$$

Donde $z = \eta + \frac{\partial \eta}{\partial \mu}(Y - \mu)$ y de esta manera se ve el método de Fisher-Scoring como Mínimos Cuadrados Iterativos, usando pseudo-observaciones z y pesos W que se actualizan en cada paso para actualizar β.

En este tipo de modelos no resulta posible interpretar directamente las estimaciones de los parámetros β, ya que son modelos no lineales. Lo que se hace en la práctica es mirar el signo de los estimadores, si el estimador es positivo, significa que incrementos en la variable asociada causan incrementos en $E(\hat{Y})$, por el contrario, si el estimador muestra signo negativo, indica que incrementos en la variable asociada causara disminuciones en $E(\hat{Y})$ (Ángel et al., 2015).

2.3.5. Pruebas de Bondad de Ajuste

Existen diversos tipos de estadísticos para los modelos lineales generalizados equivalentes a los utilizados en el modelo lineal múltiple, para esto inicialmente se definirá la desvianza residual (Dobson and Barnett, 2008). La desvianza residual del modelo corresponde a (Montgomery et al., 2006):

$$D(y : \mu(\hat{\beta})) = \sum_{i=1}^{n} w_i d(y_i; \hat{\mu}_i) \tag{2-30}$$

Con $d(y_i;\hat{\mu}_i)$ denotando la unidad de desvianza correspondiente a la observación y_i y el valor estimado $\hat{\mu}_i$, y donde w_i corresponde a los pesos asignados (si los presenta). Si el modelo incluye el parámetro de dispersión denotado σ^2 se escala la desvianza residual de la siguiente forma:

$$D^*(y:\mu(\hat{\beta})) = \frac{D(y:\mu(\hat{\beta}))}{\sigma^2} \tag{2-31}$$

Pruebas de bondad de ajuste del modelo GLM

La razón de Verosimilitud:

En esta prueba se compara 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. En este caso, en el modelo saturado, el valor máximo del logaritmo de la función de verosimilitud es cero (Montgomery et al., 2006).

$$\lambda(\vec{\beta}) = 2ln(modelo\ saturado) - 2ln(\vec{\beta}) \tag{2-32}$$

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

$$\lambda(\vec{\beta}) > \chi^2_{n-p} \quad se\ concluye\ que\ el\ modelo\ es\ bueno \tag{2-33}$$

$$\lambda(\vec{\beta}) < \chi^2_{n-p} \quad se\ concluye\ que\ el\ modelo\ no\ es\ adecuado \tag{2-34}$$

Tabla de análisis de Desvianza:

La prueba inicial de bondad de ajuste es un estadístico de una prueba correspondiente al modelo inicial, es a menudo representado en una tabla de análisis de desvianza de forma análoga que la tabla ANOVA de los modelos lineales (Tabla **2-5**). En la tabla de la bondad del ajuste el estadístico de prueba correspondiente al modelo inicial es $G^2(H_1) = D(y;\mu(\hat{\beta}))$ se muestra en la línea denominada “Error". El estadístico debe ser comparado con los percentiles de la distribución χ^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 2-5: Evaluación inicial de la bondad del ajuste de un modelo H_1, H_0 y $\hat{\mu_0}$ se refiere a modelo mínimo, es decir, un modelo con todas las observaciones que tienen el mismo valor medio.

Fuente	gl	Desvianza	Desvianza Media	Interpretación de la prueba de bondad de ajuste
Modelo H_0	k-1	$D(\mu(\hat{\beta});\hat{\mu_0})$	$\frac{D(\mu(\hat{\beta});\hat{\mu_0})}{k-1}$	$G^2(H_0\|H_1)$
Residuales (Error)	n-k	$D(y;\mu(\hat{\beta}))$	$\frac{D(y;\mu(\hat{\beta}))}{n-k}$	$G^2(H_1)$
Total	n-1	$D(y;\hat{\mu_0})$		$G^2(H_0)$

Hipótesis sobre los parámetros individuales:

Las hipótesis sobre los parámetros individuales β_j pueden hacerse utilizando el estadístico de Wald (Dobson and Barnett, 2008).

$H_0 : \beta_j = 0 \quad H_a : \beta_j \neq 0$

Estadístico: $u_j = \frac{\hat{\beta}-\beta_j}{s.e(\hat{\beta}_j)} \sim N(0,1)$ (2-35)

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

En particular un test equivalente seria $z_j = u_j^2$ y se rechazaría la hipótesis con una significancia α si $z_j > \chi^2_{i-\alpha}(1)$ (Madsen and Thyregod, 2010).

Criterios de información:

En esta investigación se utilizó el criterio de información de Akaike (AIC), ya que toman en consideración todos los parámetros p^* estimados. El AIC se define como (Montgomery et al., 2006):

$$AIC = -2ln(L) + 2p^* \tag{2-36}$$

Coeficiente de determinación R^2:

La analogía al R^2 de la regresión lineal múltiple, como una estadística utilizada en los GLM es (Dobson and Barnett, 2008):

$$R^2 = \frac{l(\bar{\pi};y) - l(\hat{\pi};y)}{l(\bar{\pi};y)} \tag{2-37}$$

Donde $l(\bar{\pi}; y)$ es la desvianza del modelo saturado, y $l(\hat{\pi}; y)$ es la desvianza 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.

Diagnostico del Modelo:

El análisis de residuales de los modelos de variable continua positiva, puede ser realizado con los residuales estandarizados (Agresti, 2007):

$$\text{Residuales Estandarizados} = d_i = \frac{y_i - \hat{\mu}_i}{SE} \tag{2-38}$$

Los residuales estandarizados tienen media cero y varianza aproximadamente unitaria, por esto cuando un residual $d_i > 3$ entonces se puede considerar como un posible valor atípico (Montgomery et al., 2006). Sin embargo se busca ubicar observaciones influyentes dentro de los residuales o una mala especificación del modelo. los supuestos a verificar son (Madsen and Thyregod, 2010):

- $V(d) = \sigma^2$ supuesto de homogeneidad de varianzas
- $cov(d_i, d_j) = 0; i \neq j$ supuesto de independencia en los errores

Sin embargo el supuesto de normalidad en los errores no necesariamente se cumple para este tipo de modelos.

- **Gráfico residuales frente a valores del predictor lineal:** Este gráfico permite la detección de errores de especificación.
- **Gráfico de la distancia de Cook:** La distancia de Cook permite medir la influencia de una observación particular en un modelo de regresión. Se define como Dobson and Barnett (2008):

$$D_i = \frac{1}{p}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \tag{2-39}$$

Donde h_{ii} es el elemento i de la diagonal de la matriz hat. e_i es el valor residual, p es el número de parámetros ajustados.

2.3.6. Multicolinealidad

La multicolinealidad es el término usualmente utilizado para referirse a la existencia de relaciones lineales o cuasilineales entre las variables predictoras en un modelo (Ramírez et al., 2005). También cuando tenemos modelos con un gran número de variables explicativas puede ocurrir que dichas variables sean redundantes o que muchas de estas variables estén correlacionadas entre sí. Al introducir variables correlacionadas en un modelo, este se vuelve inestable. Por otro lado, las estimaciones de los parámetros del modelo se vuelven imprecisas y los signos de los coeficientes pueden llegar incluso a ser opuestos a lo que la intuición nos sugiere. Además se inflan los errores estándar de dichos coeficientes por lo que los test estadísticos pueden fallar a la hora de revelar la significación de estas variables (Cayuela, 2014).

Una de las principales dificultades en el uso de estimaciones en presencia de este fenómeno es que si su propósito fundamental es evaluar la contribución individual de las variables explicativas, la estimación de los parámetros afecta la predicción del modelo (Ramírez et al., 2005). Esto es debido a que en presencia de multicolinealidad los coeficientes β_j tienden a ser inestables, es decir sus errores estándar presentan magnitudes indebidamente grandes. Esta falta de precisión afecta los contrastes parciales diseñados para evaluar la contribución individual de cada variable explicativa, corriéndose un alto riesgo de no encontrar significación en variables que realmente la tengan (Ramírez et al., 2005).

Existen dos tipos de Multicolinealidad (Gallo Gallón et al., 2013):

- **Multicolinealidad exacta:** se afirma que hay multicolinealidad exacta, cuando una o más variables regresoras X_i que son una combinación lineal de otra, es decir, que el coeficiente de correlación $r \cong \pm 1$. Esto hace que el determinante de la matriz $X'X$ sea igual a cero, lo que indica que existe dependencia lineal entre las variables (Gallo Gallón et al., 2013).

  La multicolinealidad exacta se da cuando el rango es menor al número de columnas, es decir cuando hay menos observaciones que variables predictoras. Cuando hay multicolinealidad exacta no se pueden estimar los parámetros del modelo de regresión múltiple; lo que se estima son combinaciones lineales de ellos que reciben el nombre de funciones estimables.

- **Multicolinealidad aproximada:**se dice que hay multicolinealidad aproximada, cuando una o más variables regresoras, no son exactamente una combinación lineal de la otra, pero su coeficiente de correlación entre estas variables es muy cercano a uno por lo tanto el determinante de la matriz $X'X$ es muy cercano a cero.

Identificación de la multicolinealidad

Las principales técnicas para poder detectar estas colinealidades son las siguientes:

Matriz de correlaciones:
Una forma muy práctica de determinar el grado de colinealidad es la construcción de una matriz de correlación. Las variables se colocan en filas y en columnas y sus intercepciones deben presentar el coeficiente de regresión lineal de Pearson. Se recomienda que sea eliminada una de las variables que tenga un coeficiente de correlación mayor a 0.8 con otras (González, 2010).

Factores de incremento de varianza (VIF)
Otra práctica muy usual consiste en modelar cada columna de X sobre las restantes. A partir de un R^2 muy elevado en una o más variables se evidencia una relación lineal entre la variable tomada como regresora y las tomadas como regresores (Tusell, 2011).

Llamemos $R^2(i)$ al R^2 resultante de modelar X_i sobre las restantes columnas de X. Se define el factor de incremento de varianza (variance inflation factor) $VIF(i)$ así (Tusell, 2011):

$$VIF(i) = \frac{1}{1 - R^2_{(i)}} \tag{2-40}$$

valores de $VIF(i)$ mayores que 10 (equivalentes a $R^2(i) > 0,90$) se consideran indicativos de multicolinealidad afectando a X_i junto a alguna de las restantes columnas de X. Y si los valores de la raíz cuadrada de $VIF(i)$ son superiores a 2 también se dice que hay presencia de multicolinealidad (Tusell, 2011).

3 Metodología

En este capitulo se presenta la metodología de la investigación, inicialmente se muestra las características de la población de estudio, la definición de caso, los criterios de inclusión y exclusión, y finalmente la base de datos. Después se muestra la espacialización de los casos de Cáncer infantil por tipo de enfermedad en Santiago de Cali (georeferenciación). Seguidamente se muestra la fase del análisis exploratorio de los datos. Por ultimo el análisis espacial y el ajuste del modelo lineal generalizado.

3.1. Población de estudio

Esta investigación se realizó en la ciudad de Santiago de Cali, la capital del departamento del Valle del Cauca, Colombia, específicamente esta ciudad se compone de 15 corregimientos en la zona rural, de 22 comunas y 249 barrios en el perímetro urbano. Cali es la tercera ciudad mas poblada seguida de Medellín (Alcaldia de Santiago de Cali, 2017), se encuentra entre la cordillera occidental y la cordillera central de los Andes, además alberga a 2.383.392 de habitantes en el área urbana (Alcaldia de Santiago de Cali, 2017).

En la Tabla (**3-1**) se encuentra la cantidad de habitantes en cada comuna, la cantidad de niños menores de 15 años y la densidad de niños en cada comuna.

3.2. Definición de Caso

Para el sistema de vigilancia "Vigicancer", se considera como casos a toda persona $\leq$ 18 años de edad con diagnostico de neoplasia invasora del comportamiento maligno. Ademas se consideran como casos los tumores de comportamientos inciertos o benignos del Sistema Nervioso Central (SNC). La Histiocitosis de células de Langerhans multisistematica clasificada como un comportamiento incierto, el cual se incluyo dados los hallazgos recientes de patología molecular que demuestran que es de origen neoplastico maligno.

Para los casos se tomo en cuenta el diagnóstico tumoral, el cual se utilizo el reporte histopatólogico y/o por inmunotipificación con el cual se toma la decisión de dar tratamiento especifico. También se tomo en cuenta la evidencia indirecta del compromiso en muestras citológicas de líquidos corporales; en el caso de no haber tomado muestras de

Tabla 3-1: Cantidad de Habitantes por comuna en Santiago de Cali (Alcaldia de Santiago de Cali, 2017)

Comuna	Población General	Población Niños	Densidad de Niños
1	88432,322	31835,63592	82,85691307
2	114650,5119	20637,09214	18,24186275
3	44088	9699,36	26,18290503
4	53369,16658	13875,98331	30,66527136
5	112088,8806	26901,33136	64,08703921
6	189837,3914	56951,21742	113,6375506
7	71334,15923	18546,8814	37,18628759
8	102387,5166	24573,00399	46,65692224
9	44994,3984	10798,65562	37,24433764
10	110853,9685	26604,95244	61,90466019
11	107339,0306	30054,92856	81,23829793
12	66881,49434	16051,55864	68,90861521
13	177641,0612	53292,31835	112,4960373
14	172695,7358	62170,46489	136,8392202
15	159369,1014	57372,87649	141,2974439
16	107170,4895	32151,14685	75,19269741
17	139665,3161	33519,67585	26,69629288
18	131452,8077	39435,84231	72,6444429
19	112947,2234	20330,50022	17,88570962
20	69330,64863	19412,58162	79,57612116
21	112335,6981	40440,85131	83,74707532
22	11159,97602	2455,194725	2,318600195

patología, se tomo el diagnostico clínico, lo que el grupo tratante se considera como el más probable diagnostico teniendo en cuenta la evidencia disponible incluyendo laboratorios e imágenes.

Para la fecha del diagnóstico, se tomo aquella que se encuentra consignada en el reporte de patología como la fecha de ingreso de la muestra o sino la fecha del reporte. En los casos de neoplasias con infiltración en médula osea o de algún liquido corporal, se tomo la fecha de lectura del examen. Si no hay histopatología y/o mielograma y/o tipificación inmunológica de líquidos, se tomo la fecha en que el grupo médico tratante considere que el diagnóstico este claro.

Dado que el origen celular de los cánceres en niños es muy diversos, siendo en su mayoría no epiteliales (alrededor del 90 %), se utilizo una clasificación basada en la histología más

que en la topografía de los tumores. Para tal efecto se utilizo la clasificación internacional de cáncer infantil (ICCC) en su 3 versión, desarrollada por la IARC. La clasificación presenta los 12 grupos principalmente que son: I Leucemias, II linfomas y neoplasias reticoluendoteliales, III tumores del SNC, IV Neuroblastoma y otros tumores de células nerviosas perifericas, V Retinoblastoma, VI Renales, VII Hepáticos, VIII Óseas Malignas, IX Sarcomas de tejidos blandos y extraóseos, X Tumores Germinales, Trofablasticos y otros ganadales, XI Otras neoplasias epiteliales y melanoma, XII Otras neoplasias y neoplasias malignas no especificadas.

Para este proyecto se escogió estudiar la Leucemia Linfoblastica Aguda, linfoma de burkitt y Medulloblastoma, que corresponden a los grupos I, II y III del ICCC-3 respectivamente.

3.3. Criterios de Inclusión y Exclusión

En esta sección se describe los criterios que se tuvieron en cuenta para ingresar a la base de datos con la que se trabajo en esta investigación.

3.3.1. Criterios de Inclusión

- Pacientes menores e iguales a 15 años
- Diagnostico confirmado de Leucemia Linfoblastica Aguda, Linfoma de Burkitt o Meduloblastima
- Zona de nacimiento en Santiago de Cali
- Dirección conocida en el área urbana de Cali

3.3.2. Criterios de Exclusión

- Para el objetivo de definir el patrón espacial en los casos de cáncer de interés, se excluyeron a todos los pacientes que no se encuentren en el área urbana de Cali y los que la dirección registrada no sea conocida.
- Para el objetivo de modelar las tasas de casos de LLA se excluyeron todos los pacientes que no cumplían con los criterios de de inclusión.

3.4. Base de datos

La información sobre los casos de Cáncer en Cali es recogida de las historias clínicas. La base de datos de Vigicancer proporcionada para este estudio, esta conformada por las

siguientes variables: identificación de cada paciente, dirección, municipio, sexo, edad, mes y año del diagnóstico, el tipo de cáncer que padece cada paciente, comuna, etnia y células precursoras; del cual solo nos interesa estudiar si tienen Leucemia Linfoblastica Aguda (LLA), Linfoma de Burkitt o Meduloblastoma. La consolidación de los registros se realizó contando con el apoyo permanente del Registro Poblacional de Cáncer de Cali (RPCC) de la Universidad del Valle y están acorde con los requerimientos de calidad de la Agencia Internacional para la Investigación del Cáncer (IARC).

La información empleada en esta investigación pertenece al Sistema de Registro Poblacional de Cáncer de Cali (RPCC); este es un grupo de investigación de la Universidad del Valle, adscrito a la Facultad de Salud. RPCC es uno de los más importantes del mundo y el único de base poblacional de tan larga trascendencia en Colombia (Muñoz et al., 2014). El RPCC es considerado como la fuente de epidemiología descriptiva de cáncer más importante de Latinoamérica (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, producto de la búsqueda activa y permanente de datos en todas las fuentes de información, preservando siempre la confidencialidad (Registro Poblacional de Cáncer en Cali, 2017).

La Figura (**3-1**) muestra los pasos realizados de filtración de información para realizar la investigación de este proyecto. Se obtuvieron 781 registros, los cuales contenían 12 categorías de ICCC-3 (International Classification of Childhood Cancer), encontrando pacientes con direcciones en otros municipios del Valle del Cauca; por consiguiente se realizó una depuración teniendo en cuenta la ciudad de nacimiento; de igual manera a pacientes que no registraban dirección y que tuvieran alguna de las tres enfermedades de interés. Se obtuvo a partir de este filtro un archivo de 244 pacientes, cuya ubicación fueron georreferenciardas a coordenadas planas con el fin de trabajar los patrones puntuales. En este proceso se obtuvieron 232 coordenadas, ya que las otras 12 restantes no cumplían con los criterios de inclusión. Por ultimo, con el fin de contar con el rango de edad requerida que es entre los 0 y 15 años de edad resultando finalmente con 202 casos de cáncer infantil con los que se trabajo en este estudio.

Considerando lo mencionado anteriormente, la base de datos final implementada en esta investigación esta compuesta por las variables: comuna, sexo, edad, año, mes, código del paciente, dirección, comuna, etnia y células precursoras.

Definición operativa de las variables:

- **Sexo:** (Cualitativa - Nominal). Se determina el genero del paciente segun sus característica fenotípicas en el momento del examen físico. Se midió como una variable

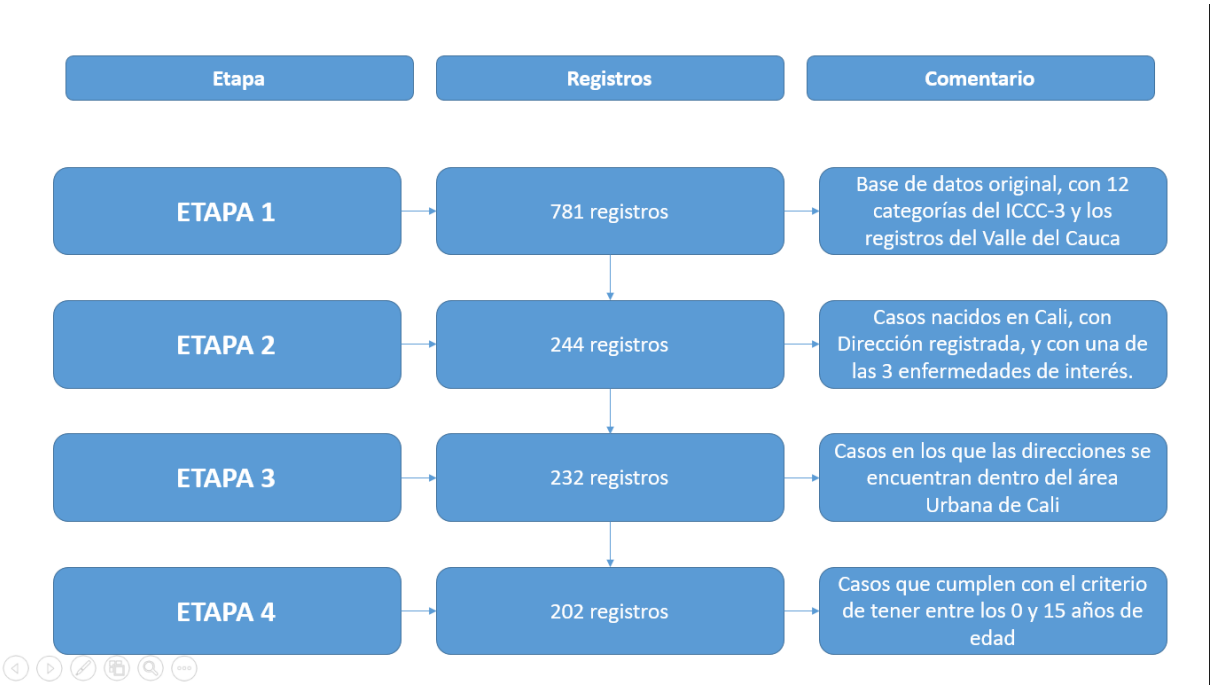

Figura 3-1: Diagrama de depuración de base de datos

categórica en Masculino, Femenino y Ambiguo (donde no se presento ningún caso de estos).

- **Edad:** (Cuantitativa - Razón). Se considera como el tiempo comprendido entre la fecha de nacimiento y la fecha de la realización del diagnostico. Se midió como una variable continua en años. Para el análisis la edad se categorizó en grupos de la siguiente forma: de 0 a 9.99 años y de 10 años a 15 años.

- **Año:** (Cuantitativa - Intervalo). Año en el que se diagnosticó al menor con uno de los tres tipos de cáncer infantil, tomó como valores los años 2009, 2010,..., 2016.

- **Mes:** (Cualitativa - Ordinal). Se define como el mes del año en el se realizó el diagnostico del cáncer al menor, posee doce categorías correspondientes a los meses de enero, febrero, ..., diciembre, codificadas del 1 al 12 de acuerdo al mes correspondiente.

- **Comuna:** (Cualitativa - Nominal). Comuna de la ciudad en la cual se ubicó la residencia de cada paciente diagnosticado con uno de los tres tipos de cáncer infantil estudiados en esta investigación, la ciudad se encuentra compuesta por 22 comunas que corresponden a las categorías de la variable.

- **Etnia:** (Cualitativa-Nominal). Determina si el individuo es afrocolombiano o no, según sus características fenotípicas al examen físico. Se midió por medio de observación

directa del medica tratante y se consigno como una variable dicotomica.

- **Linaje LLA:** (Cualitativa-Nominal). Esta variable distingue el linaje de la Leucemia Linfoblastica Aguda, tomando dos posibles valores, Células B o Células T.
- **Densidad de Niños:** (Cuantitativa-Razón) esta variable indica la cantidad de niños menores de 15 años por cada km^2 de a comuna.

3.5. Georeferenciación

El proceso de depuración y geocodicación, dando la ubicación espacial de los casos que presentan cáncer de Leucemia Linfoblastica Aguda, Linfoma de Burkitt y Meduloblastomas en menores de 15 años. Las direcciones de los pacientes se transformaron en coordenadas en el sistema Gauss Kruger (geográficas planas), que se representa en longitud y latitud.

3.6. Análisis Estadístico

En esta sección se presenta el análisis estadístico que se realizó en la investigación. La primera etapa la constituyó el análisis exploratorio, continuando con las pruebas de completa aleatoriedad espacial y la estimación de la intensidad. La segunda parte esta compuesta por la modelación estadística de la tasa de casos de cáncer en niños menores de 15 años por comuna en el perímetro urbano de Santiago Cali. Los resultados estadísticos se obtuvieron con el software libre R, utilizando principalmente los paquetes: ggplot2, spatstat y MASS, geoftest, rater, ncdf4, geoR, sp y ggmap.

3.6.1. Análisis Exploratorio

Inicialmente se realizó un análisis descriptivo univariado para las variables, obteniendo tablas de frecuencia y gráficos de barras para las variables cualitativas, mientras que para las variables cuantitativas se obtuvieron medidas de resumen compuestas por el mínimo, máximo, media, mediana y desviación estándar. También se realizó un análisis bivariado donde se obtuvieron tablas de frecuencias cruzando la información con las variables de interés.

3.6.2. Test de completa aleatoriedad espacial

Para iniciar el análisis estadístico a nivel espacial, se tuvieron en cuenta los registros georeferenciados dentro del perímetro urbano de Santiago de Cali y para esto se obtuvieron los mapas de la ciudad de Cali con los casos de Leucemia Linfoblastica Aguda, Linfoma de Burkitt y Meduloblastomas. Una vez creado el patrón puntual se realizaron las pruebas de aleatoriedad espacial basados en cuadrantes y en distancias.

Pruebas basadas en cuadrantes

Para las pruebas basadas en cuadrantes, se dividió la región tanto horizontal como verticalmente en 5 partes y se limitó con el borde del área de Santiago de Cali, obteniendo un total de 20 zonas para las pruebas chi-cuadrado y el test de Montecarlo. Inicialmente se plantea las hipótesis para la prueba Chi-cuadrado de la siguiente forma:

$H_0 : El\ patron\ es\ aleatorio \qquad H_1 : El\ patron\ es\ agrupado$

Sin embargo, para la prueba chi-cuadrado se necesita que todas las casillas observadas tengan una frecuencia mayor a 5 eventos. Debido al no cumplimiento de el número mínimo de eventos se realizo el test de Monte Carlo, que mediante 1000 simulaciones se solucionó el problema descrito.

En muchas ocasiones las pruebas basadas en cuadrantes tienen como falencia la definición de la cuadrática en el espacio geografico, ya que se define en una zona cuadrada, por tanto se pierde la noción de perímetro urbano de la ciudad, creando sesgo en la estimación de las pruebas (Cressie, 2015).

Pruebas basadas en distancias

Estas pruebas son de gran importancia, ya que no se basan en la cantidad de casos dentro de un área como en el método anterior, sino en la distancia promedio entre los registros, por esto se emplearon las diferentes metodologías de pruebas basadas como la Función G, la Función F, la Función K de Ripley y la Función K de L, descritas a continuación.

Función *G(d)*

Esta función se encarga de calcular la distancia del vecino más cercano $G(d)$. Por tanto sus parámetros varían para cada punto en el espacio. A manera de ejemplo de la construcción de la curva, se muestra en la ecuación (3-1) la estimación de la función empírica $\hat{G(D)}$ y en la ecuación (3-2) la estimación de la función teórica $G(D)$.

$$\hat{G}(D) = \frac{N(di \leq d)}{n} \tag{3-1}$$

$$G(D) = 1 - e^{-\lambda \pi d^2} \tag{3-2}$$

Obteniendo gráficas para cada tipo de patrón, como se muestra en la Figura (**3-2**).
Donde:

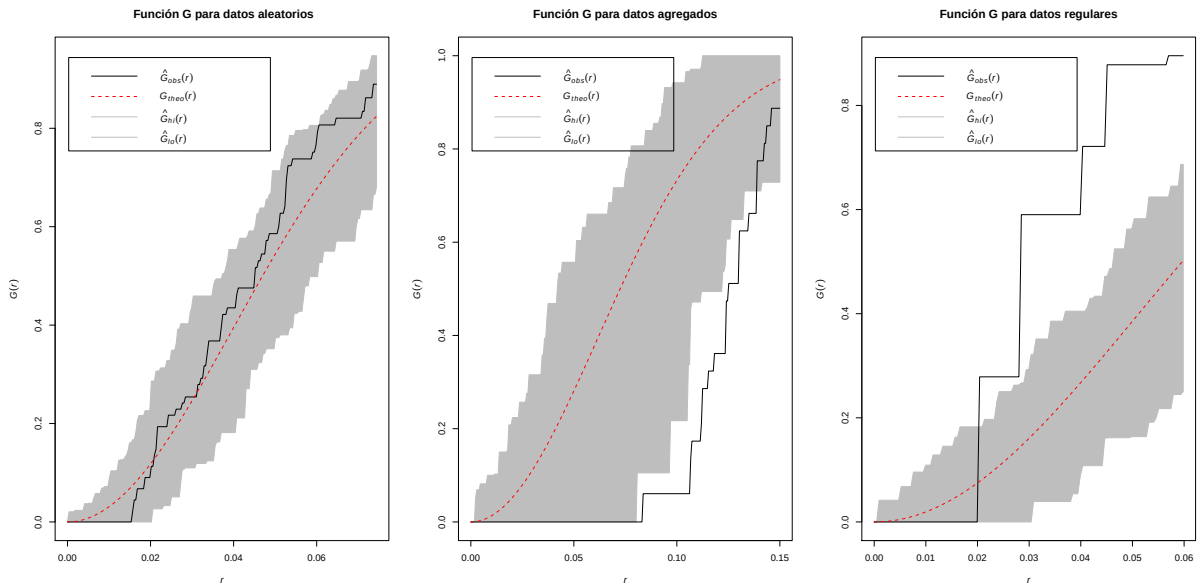

Figura 3-2: Función distancia del vecino más cercano *G(h)* para cada tipo de patrón

- Si $\hat{G(d)}$ crece rápidamente en distancias cortas, quiere decir que los eventos son regularmente espaciados.
- Si los eventos son agregados, $\hat{G(d)}$ crece lentamente hasta cierta distancia(espacio eventos) y después crece rápidamente.

Función *F(d)*

La función de distribución esférica $F(d)$, tiene en cuenta la mínima distancia entre punto-evento, es decir se elije un punto aleatoriamente dentro de una región de estudio y se mide la distancia con los registros. A manera de ejemplo se presenta las dos estimaciones de la función *F(D)*. En la ecuación (3-3) se encuentra la función empírica y en la ecuación (3-4) se encuentra la función teórica.

$$\hat{F}(D) = \frac{N(di \leq d)}{m} \tag{3-3}$$

$$F(D) = 1 - e^{-\lambda \pi d^2} \tag{3-4}$$

Obteniendo gráficas para cada tipo de patrón, como se muestra en la Figura (**3-3**).
Donde:

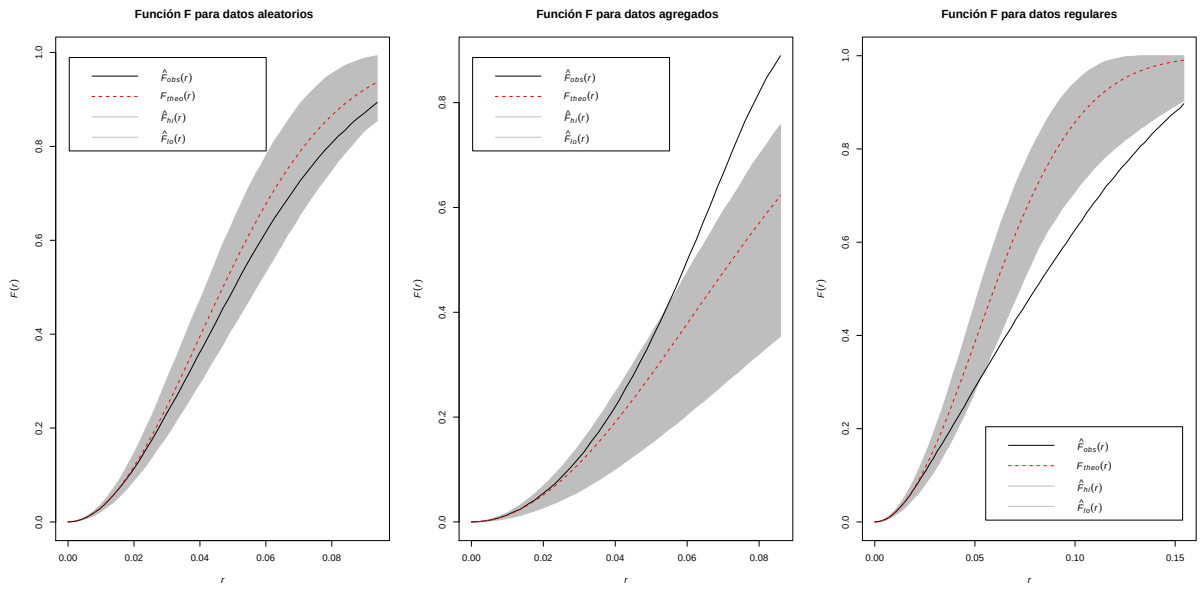

Figura 3-3: Función distancia del vecino más cercano *F(h)* para cada tipo de patrón

- Si $\hat{F(d)}$ crece lentamente al comienzo y rápidamente para distancia largas, el patrón espacial es agregado.
- Si $\hat{F(d)}$ crece rápido al comienzo y luego no, se dice que el patrón es regular.

Función K de Ripley y la Función K de L

La función K de Ripley, también es conocida como la medida reducida del segundo momento o la segunda función de momento reducido, el proceso de estimación para esta función es similar al descrito para la función $G(d)$ y $F(d)$, obteniéndose un d que varía desde el inicio del cálculo.

En cambio la Función K de L, es una modificación de la función K de Ripley, para poder probar la aleatoriedad espacial, el cual se dice que el patrón es aleatorio si no hay desviación de la recta creada, como se observa en la siguiente ecuación:

$$L(d) = \sqrt{\frac{K(d)}{\pi}} \text{ Bajo aleatoriedad } L(d) = \sqrt{\frac{\pi d^2}{\pi}} = d \tag{3-5}$$

En la Figura (**3-4**) se encuentran las gráficas de las funciones $K(d)$ para cada uno de los tipos de patrones espaciales.
En la Figura (**3-5**) se encuentran las gráficas de las funciones $L(d)$ para cada uno de los tipos de patrones espaciales.

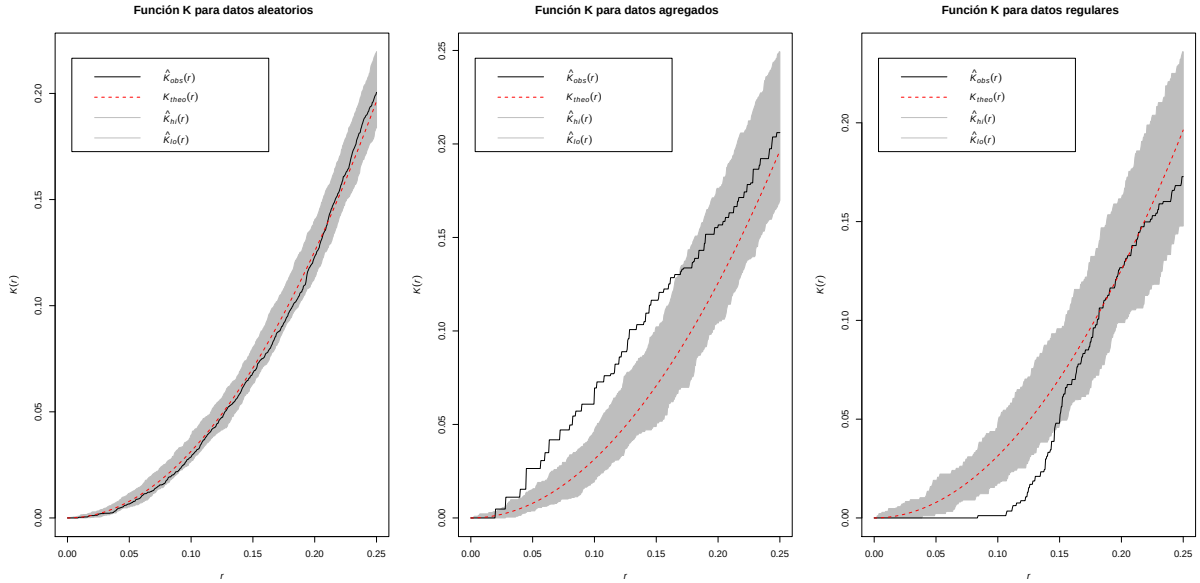

Figura 3-4: Función de la medida reducida *K(h)* para cada tipo de patrón

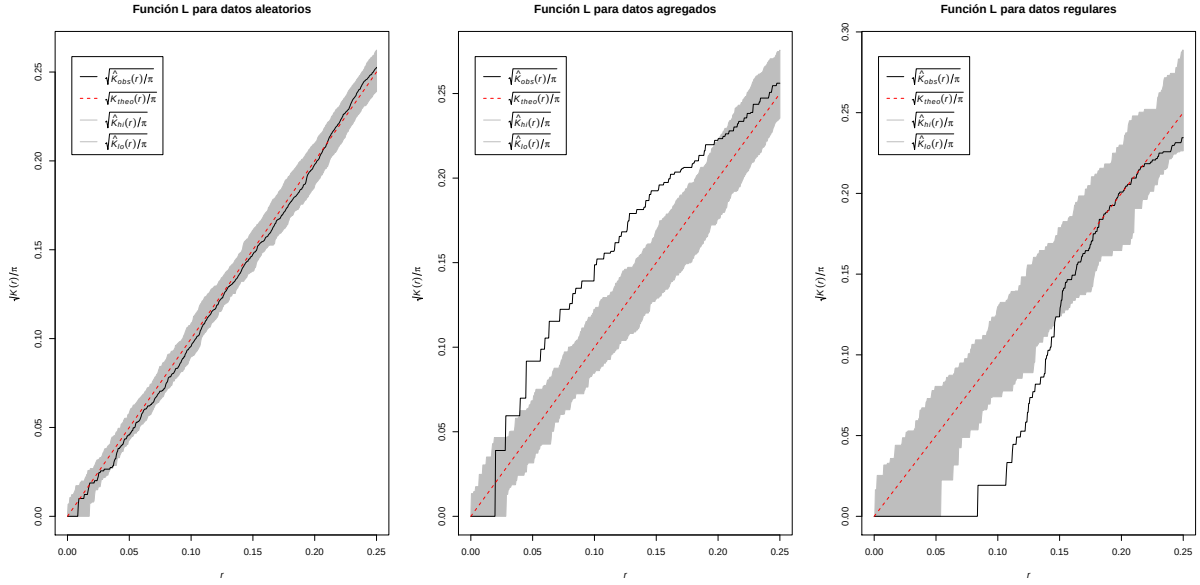

Figura 3-5: Función de la medida reducida modificada *L(h)* para cada tipo de patrón

3.6.3. Estimación de intensidad

En la estimación de la intensidad se realizó la conversión de los patrones de puntos a superficies lisas en el espacio, con el fin de mostrar los sectores del perímetro urbano de

la ciudad de Santiago de Cali, en las cuales se encuentran los casos de Cáncer infantil. La estimación de la función de intensidad se realizó con la densidad Kernel, utilizando un Kernel Gausiano (ecuación 3-6). La estimación de intensidad se limito con el perímetro urbano de la ciudad y de esta forma obtener mejor análisis de la intensidad de los casos registrados con algún tipo de cáncer infantil.

$$\hat{\lambda}(s) = \frac{1}{|A|}\sum_{i=1}^{n} K\left(\frac{s - s_i}{n}\right) \text{ donde } k = \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} \tag{3-6}$$

Para la estimación de intensidad se omitió que el espacio es de tipo red, por tanto se hizo necesario la optimización de los anchos de banda de la densidad kernel para que la intensidad estuviera lo más cercana posible a los casos registrados. Los anchos de banda se calcularon utilizando los criterios de validación cruzada (*bw.diggle*, ecuación (3-7)) y verosimilitud de validación cruzada(*bw.ppl*, ecuación (3-8)). Donde $MSE(\sigma)$ es el error cuadrático medio en el ancho de banda σ, $g(0)$ es la función de correlación de pares, λ es la intensidad media, $\lambda_i(x_i)$ es la estimación de suavizado de Kernel, x_i es el ancho de banda suavizado σ y $\lambda(u)$ es la estimación de suavizado de Kernel de la intensidad en una ubicación espacial u.

$$M(\sigma) = \frac{MSE(\sigma)}{\lambda^2} - g(0) \tag{3-7}$$

$$LCV(\sigma) = \sum_{i} log(\lambda_i(x_i)) - \int_W \lambda(u)du \tag{3-8}$$

De acuerdo a lo expuesto en el marco teórico sobre la selección del ancho de banda de la intensidad de Kernel por los métodos de validación cruzada y verosimilitud de validación cruzada, donde se define que el ancho de banda σ varia de acuerdo a una intensidad λ, como se logra ver en la ecuación (3-7) que si vamos aumentando el λ el estimador del ancho de banda σ va disminuyendo, lo que cumpliría el propósito de este método, ya que su principal objetivo es minimizar el error cuadrático medio. En cambio en la ecuación (3-8) vemos que si se va aumentando el λ el estimador del ancho de banda σ va aumentando por la integral y la sumatoria que abarca esta ecuación, por tal razón se cumple el propósito de este método que es maximizar la probabilidad del proceso puntual. La comparación gráfica de estos métodos se encuentran en los anexos, donde se diferencia la estimación de la intensidad. Pero como se busca es tener una intensidad mas aproximada a los casos estudiados, se opta por utilizar el el método de Validación Cruzada, ya que minimiza el error cuadrático medio y realiza mejores estimaciones de intensidad.

3.7. Modelación

Inicialmente para realizar la modelación se identificaron las variables socio-económicas disponibles que tuvieran mayor relevancia. Esto se tuvo en cuenta, ya que el objetivo de esta investigación es Modelar el fenómeno de casos de Leucemia Linfoblastica Aguda en niños menores de 15 años en la ciudad de Santiago de Cali a través de variables que lograran explicarlo. Entre estas se tuvieron en cuenta el genero del menor, la edad que presentaba el menor en el momento que fue diagnosticado, si son de descendencia Afrocolombiano, si la enfermedad es de Células B o Células T.

En la primera etapa de modelación se tuvieron los datos puros, es decir, el registro de cada paciente con sus respectivas características para cada variable. Se planteó para este estudio modelos de variable continua en el cual la variable de respuesta es la tasa de casos que presentan leucemia linfoblastica aguda por cada 1'000.000 de habitantes en las comunas (unidad de área) de Santiago de Cali. Las variables explicativas con las que se construyó el conteo son: El genero (Hombre o Mujer), La edad (Menores de 10 años o de 10 a 15 años), la etnia (Afrocolombiano o no afrocolombiano), linaje de LLA (Células B o Células T) y densidad de niños menores de 15 años. Para poder emplear estas variables en el modelo con la variable de respuesta (tasa de casos por comuna), se obtuvo una tabla de contingencia.

Por medio de la tabla de contingencia se pueden ver las subpoblaciones (comunas) con las que se trabajaron para realizar el modelo, de acuerdo a las categorías de las variables que caracterizan a los pacientes que serían las variables a aplicar en el modelo. En la Tabla (**3-2**) se encuentran los conteos de estas variables para las 22 comunas de la ciudad de Santiago de Cali.

Junto a las variables antes mencionadas en la Tabla (**3-1**), se encuentra la variable D̈ensidad de Niños"por comuna de la ciudad de Santiago de Cali, el cual ayuda a medir el riesgo de encontrar más casos de Leucemia Linfoblastica Aguda en cada comuna. Teniendo un total de 9 variables explicativas para ajustar el modelo lineal generalizado.

Antes de ajustar el modelo, es de gran importancia mirar el comportamiento de cada una de las variables explicativas entre ellas, para observar mejor esto, se encuentra la Tabla (**4-11**); donde se muestra las cuatro variables categóricas y su asociación con las demás variables; el valor p marcado en esta tabla, es de chi-cuadrado de Pearson, el cual muestra la asociación entre dos variables, donde se encontró que existe una asociación alta entre las variables genero vs edad y etnia vs genero; entre la variable edad vs linaje existe una relación no tan baja, con un valor p=0.4239. En cambio existe una relación baja entre las variables genero vs linaje y edad vs etnia, por último se encontró que entre las variables etnia y linaje hay

Tabla 3-2: Tabla de Contingencia para cuantificar las variables a emplear en el modelo

Comuna	Hombre	Mujer	Edad1	Edad2	SiAfro	NoAfro	CelulasB	CelulasT
1	5	1	5	1	0	5	5	1
2	2	5	5	2	2	5	7	0
3	13	3	10	6	2	14	12	4
4	1	2	2	1	0	3	3	0
5	2	3	3	2	0	5	5	0
6	7	4	8	3	1	10	8	3
7	4	3	4	3	2	5	6	1
8	4	7	5	6	0	8	9	1
9	3	3	4	2	0	6	6	0
10	3	3	4	2	1	5	6	0
11	5	6	9	2	2	9	10	0
12	2	0	2	0	0	2	2	0
13	2	3	5	0	1	4	3	1
14	4	7	8	3	4	7	10	1
15	6	0	5	1	1	4	3	2
16	4	5	7	2	2	7	8	1
17	9	4	9	4	1	12	12	0
18	5	2	7	0	1	6	7	0
19	5	5	6	4	1	7	9	0
20	0	1	0	1	0	1	1	0
21	4	4	5	3	1	7	7	1
22	0	1	1	0	0	1	1	0

Tabla 3-3: Tabla de asociación entre las variables categóricas

		Edad		valor p	Etnia		Valor p	Linaje		Valor p
		<10	10 >x<15		Afro	No Afro		Células B	Celulas T	
Genero	Hombre	64	25	0.6802	13	74	0.7524	74	72	0.1736
	Mujer	49	23		9	59		62	4	
Edad	<10				19	91	0.1713	98	9	0.4239
	10 >x <15				3	42		38	7	
Etnia	Afro							18	3	0.0000
	No Afro							117	13	

una independencia.

Una vez realizado el agrupamiento de los datos y la relación entre las variables, se identifico la distribución de la tasa de casos de Leucemia, para esto se hizo uso de la prueba Anderson Darling para Bondad de ajuste. Las pruebas se realizaron teniendo en cuenta que son datos continuos, para esto se utilizaron como distribuciones de referencia la distribución Normal y la Gamma, donde en la ecuación 3-9 se plantea las hipótesis.

$$H_0 : F_n(x) = F_t(x) \qquad H_1 : F_n(x) \neq F_t(x) \tag{3-9}$$

siendo $F_n(x)$ la función de distribución empírica y $F_t(x)$ la función de densidad teórica (Normal y Gamma). Para realizar las pruebas Anderson Darling, inicialmente se hizo mediante métodos gráficos el ajuste de la distribución, revisando como se ajustan las distribuciones teóricas a los datos, además se realizó una estimación previa de los parámetros

de la distribución Normal y Gamma ($F_t(x)$), con la cual se compararon los datos. La estimación de los parámetros de las distribuciones se obtuvo mediante el método de máxima verosimilitud, y con ellos realizar la prueba Anderson Darling.

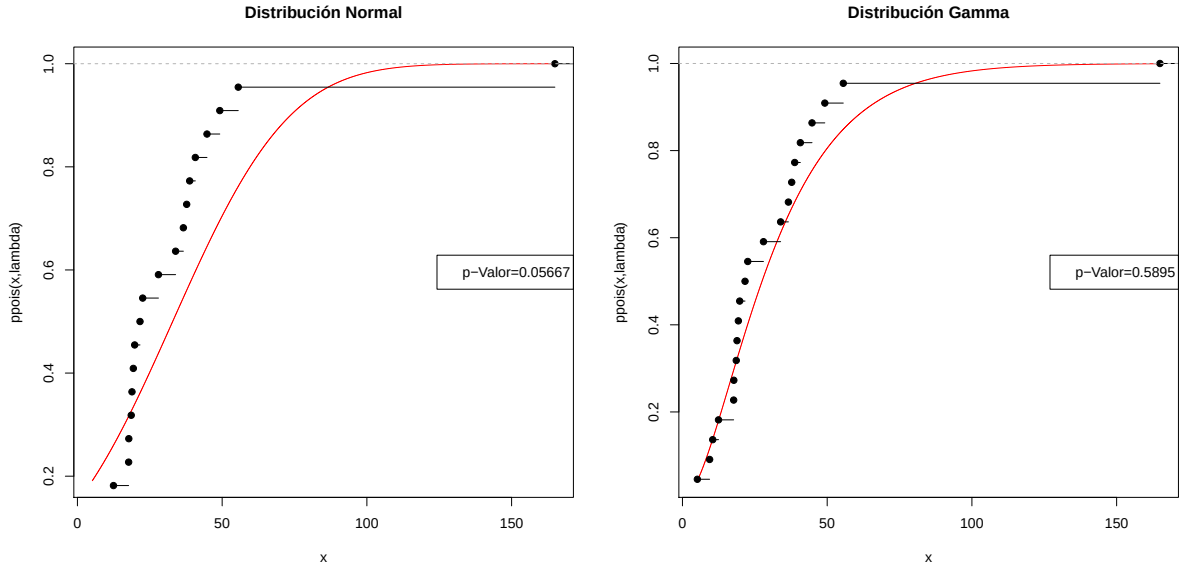

Figura 3-6: Prueba Anderson Darling para la distribución de la tasa de casos de Leucemia

En la Figura (**3-6**) se observa que la distribución que mejor se ajusta a la tasa de casos de Leucemia en Santiago de Cali es la Gamma, no solo por su valor p, sino porque los datos se ajustan mejor a su distribución teórica. Por lo tanto la variable de respuesta para los modelos que se van a ajustar en el estudio siguen una de Distribución Gamma y la función de enlace apropiada para este tipo de distribución seria una Logarítmica (Cayuela, 2014).

Modelo Lineal Generalizado con Variable de Respuesta con Distribución Gamma

Para modelar la tasa de casos de Leucemia Linfoblastica Aguda, en Santiago de Cali, Colombia, se plantea un modelo lineal generalizado de la forma (Montgomery et al., 2006):

$$E(\vec{y}|X) = \vec{\mu} \qquad \vec{\eta} = X'\beta$$

Como se menciono anteriormente, la función de vínculo que se va a utilizar es la logarítmica, ya que realiza mejores estimaciones (Cayuela et al., 2016). La matriz de diseño X se compuso de las variables independientes Genero, Edad, Afrodecendencia, linaje de LLA, y densidad de niños por comuna, que están asociadas a los siguientes niveles:

$$\text{Genero} \begin{cases} \text{Mujer (Control)} \\ \textit{Hombre} \end{cases} \qquad \text{Edad} \begin{cases} \text{Mayores e iguales a 10 años (Control)} \\ \text{Menores de 10 años} \end{cases}$$

$$\text{Afrodecendencia} \begin{cases} \text{Si es Afrocolombiano (Control)} \\ \text{No es Afrocolombiano} \end{cases} \quad \text{Células de Leucemia} \begin{cases} \text{Células T (Control)} \\ \text{Células B} \end{cases}$$

Cada una de las variables categóricas tendrá una variable de control, asociada a las variables que se indicaron anteriormente, por lo que que cada variable estimará un solo parámetro, para cada una de las variables agregadas en el modelo. La variable Densidad de niños menores de 15 años, no tiene categorías, sino que es una variable continua.

Inicialmente para evaluar el aporte individual de cada variable explicativa, se realizaron 5 modelos simples como se muestra en la Tabla (**3-4**) y se compararon con un modelo nulo (sin ninguna covariable) por medio del AIC y de la prueba χ^2 de comparación de modelos.

Tabla 3-4: Expresiones de los modelos simples

Modelo	**Expresión matemática**
Modelo 1	$ln(\mu) = \beta_0 + \beta_1(Hombre)$
Modelo 2	$ln(\mu) = \beta_0 + \beta_1(<$ 10 años)
Modelo 3	$ln(\mu) = \beta_0 + \beta_1$(No son afrocolombianos)
Modelo 4	$ln(\mu) = \beta_0 + \beta_1$(Células B)
Modelo 5	$ln(\mu) = \beta_0 + \beta_1$(Densidad niños)
Modelo nulo	$ln(\mu) = \beta_0$

De igual manera el modelo completo queda expresado de la siguiente manera:

$$ln(\mu) = \beta_0 + \beta_1(Hombre) + \beta_2(< \text{ 10 años}) + \beta_3(\text{No son Afro}) + \beta_4(\text{Células B}) + \beta_5(\text{Densidad niños}) \tag{3-10}$$

Para las tasas de los casos de leucemia en las comunas se tiene:

$$\vec{y} = \begin{bmatrix} 18{,}846 \\ 33{,}919 \\ \vdots \\ 40{,}729 \end{bmatrix}_{22x1} \qquad \vec{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_5 \end{bmatrix}_{22x1}$$

$$\vec{X} = \Big[\underbrace{Intercepto}_{\vec{1},} \quad \underbrace{Genero}_{\vec{Mujer},\vec{Hombre}} \quad \underbrace{Edad}_{\text{Mayor a } \vec{10} \text{ años,Menor a } \vec{10} \text{ años}} \quad \underbrace{Afrodecendencia}_{\text{Si } \vec{\text{Afro}},\text{No } \vec{\text{Afro}}} \quad \underbrace{Celulas}_{\text{Célul}\vec{\text{as}} \text{ T,Célul}\vec{\text{as}} \text{ B Densidad}\vec{\text{ Niños}}} \Big]_{22}$$

Estimación de los parámetros y bondad de ajuste del modelo

Los parámetros de regresión $\vec{\beta}$ se estiman mediante el proceso iterativo de Fisher Scoring. El método de Fisher Scoring, es un proceso iterativo que a partir de una secuencia de estimaciones repetitivas se logra conseguir que ciertos parámetros converjan y se estabilicen en un solo valor. La estimación de $\hat{\beta}$ por este método es la siguiente:

$$\beta^{(t+1)} = \beta^{(t)} - [\ell''(\beta^{(t)})]^{-1}\ell'(\beta^{(t)}) \tag{3-11}$$

donde $\ell'(\beta^{(t)} = \frac{\partial_i}{\partial\beta}$, $\ell''(\beta^{(t)}) = \frac{\partial^2 L_i}{\partial\beta_k\partial\beta_j}$. De esta forma, la ecuación (3-11), se descompone de la siguiente manera:

$$\frac{\partial\ell}{\partial\beta_j} = \sum_{i=1}^{n} \frac{Y_i - \mu_i}{V_i} \quad \frac{\partial\mu_i}{\partial\eta_i} \quad x_{ij} = 0$$

$$\frac{\partial^2\ell}{\partial\beta_k\partial\beta_j} = \sum_{i=1}^{n} \frac{\partial}{\partial\beta_k}[Y_i - \mu_i]\frac{1}{V_i}\frac{\partial\mu_i}{\partial\eta_i}x_{ij} + \sum_{i=1}^{n}(Y_i - \mu_i)\frac{\partial}{\partial\beta_k}\left[\frac{1}{V_i}\frac{\partial\mu_i}{\partial\eta_i}x_{ij}\right]$$

De acuerdo con las ecuaciones $V = a(\phi)b''(\theta)$ donde esta depende de la función de enlace a utilizar, si se utiliza la función identidad se obtiene $V = a(\phi)b''(\eta)$.

El método de Fisher Scoring se caracteriza por calcular $E\left(\frac{\partial^2\ell}{\partial\beta_k\partial\beta_j}\right)$ que es mas precisa y realiza una aproximación a los mínimos cuadrados, por lo tanto esta es igual a:

$$E\left(\frac{\partial^2\ell}{\partial\beta_k\partial\beta_j}\right) = -\sum_{i=1}^{n} V_i^{-1}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 x_{ij}x_{ik} \tag{3-12}$$

La ecuación (3-12) se puede expresar de forma matricial como $-X'WX$. En la ecuación (3-13) se expresa la matriz W; por último realizando un proceso algebraico a la ecuación (3-11) se obtiene la ecuación (3-14):

$$W = diag\left(V_i^{-1}\left[\frac{\partial\mu_i}{\partial\eta_i}\right]^2\right) \tag{3-13}$$

$$\begin{aligned}\beta^{(t+1)} &= \beta^{(t)} + (X'WX)^{-1}X'V^{-1}\frac{\partial\mu}{\partial\eta}(Y - \mu) \\ &= (X'WX)^{-1}X'Wz\end{aligned} \tag{3-14}$$

Donde $z = \eta + \frac{\partial \eta}{\partial \mu}(Y - \mu)$ y de esta manera se ve el método iterativo de Fisher Scoring, usando pseudo-observaciones z y pesos W que se actualizan en cada paso para actualizar β.

Por lo que en este tipo de modelos no resulta posible interpretar directamente las estimaciones de los parámetros β, ya que son modelos no lineales. Lo que se hace en la práctica es mirar el signo de los estimadores, si el estimador es positivo, significa que los incrementos en la variable asociada causan incrementos en $\hat{Y}$, por el contrario, si el estimador muestra signo negativo, indica que los incrementos en la variable asociada causará disminuciones en $\hat{Y}$ (Ángel et al., 2015).
Una vez estimado el modelo de regresión se procede a calcular los indicadores de desvianza nula, desvianza residual, y las pruebas de Bondad de Ajuste (coeficiente de determinación R^2 y AIC) y por la prueba VIF de multicolinealidad.

Para probar si el modelo se ajustó adecuadamente se realizó la prueba de bondad de ajuste, cuyo criterio es:

$$(Null.desviance - residual.desviance) > x^2_{n-p} \tag{3-15}$$

Si se cumple el criterio, se dice que el modelo presenta un buen ajuste, de lo contrario el modelo no se encuentra bien ajustado.

Para conocer el valor-p asociado a esta prueba se calcula por medio de $P(x > (null.deviance - residual.desviance))$ utilizando la distribución chi-cuadrado.

Usando las pruebas de bondad de ajuste del modelo conforme se mostró en el marco teórico, se tiene que el coeficiente de determinación R^2 se calcula como se muestra en la ecuación (3-16).

$$R^2 = \frac{l(\bar{\pi}; y) - l(\hat{\pi}; y)}{l(\bar{\pi}; y)} = \frac{(null.deviance - residual.deviance)}{null.deviance} \tag{3-16}$$

Y el criterio de Información de Akaike se calcula como se muestra en la ecuación)3-17), donde L es la función de verosimilitud y p^* son los parámetros estimados en tal modelo.

$$AIC = -2ln(L) + 2p* \tag{3-17}$$

Además se obtiene un resumen del modelo con la estimación de los parámetros, el error estándar (SE), el valor estadístico de Wald ($Z_i = \hat{\beta}/SE$), con el cual se desea probar la hipótesis nula $H_0 : \beta = 0$ versus la hipótesis alternativa $H_a : \beta \neq 0$, y $z^2 \sim \chi_{(1)}$ (Agresti, 2007). Finalmente se encuentra $Pr(> |z|)$, si esta probabilidad se encuentra por debajo

del 0.05, se dice que la categoría presenta diferencias significativas con la categoría de control.

Por último, se utilizan métodos gráficos para validar el supuesto homogeneidad la varianza en los errores. En la Figura (**3-7**) se muestra los valores predichos versus los residuales del modelo lineal generalizado con las variables incluidas, observando que los residuales no tienen una tendencia lineal y se encuentran al rededor de cero, concluyendo que se cumple el supuesto de homogeneidad de varianza para este modelo.

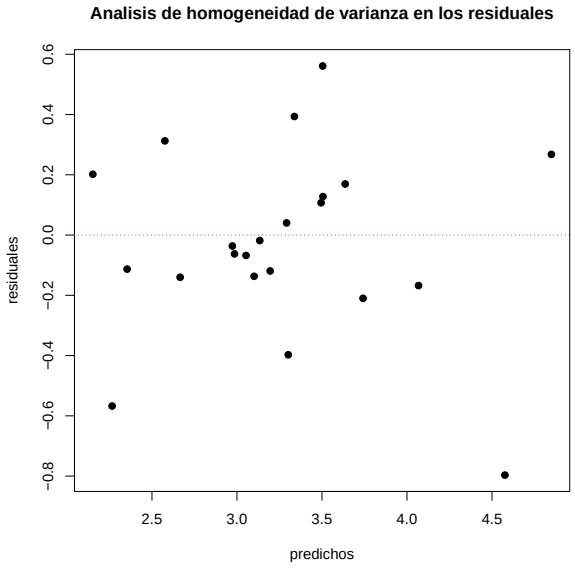

Figura 3-7: Validación de los supuestos para el modelo completo

Como se mostró antes de ajustar el modelo, donde existe una alta correlación entre las variables, para determinar si existe problemas de colinealidad, se utilizó el Factor de Inflación de Varianza (VIF), que se calcula como se muestra en la ecuación (3-18), donde se calcula el coeficiente de determinación R^2 para cada uno de los parámetros, es decir, que cada parámetro es modelado con respecto a las demás variables explicativas, generando de esta forma el R^2 de cada uno de los 6 parámetros que conforman el modelo completo. Si la raíz cuadrada de este indicador es superior a 2, se dice que la variable tiene una colinealidad alta y puede causar problemas al modelo.

$$VIF(i) = \frac{1}{1 - R_i^2} \tag{3-18}$$

Si el modelo completo no presenta un buen ajuste, pocas variables significativas y presenta problemas de multicolinealidad entre las variables; se expone un modelo que permiten

explicar las tasas de casos de Leucemia Linfoblastica Aguda en Santiago de Cali por medio de las variables que la explican. Por medio del método *Stepwise* se realizó una selección de variables con el criterio AIC, donde el método opera a partir de un modelo completo o saturado, y comienza a quitar una variable en cada iteracción. A través el criterio AIC, se obtiene el modelo que tenga aquella combinación de variables con el menor valor de este criterio (Montgomery et al., 2006).

De esta forma se llega a un modelo que logre explicar el fenómeno de tasas de casos de Leucemia linfoblastica aguda, por medio de variables socio-demográficas por comunas en la ciudad de Santiago de Cali.

4 Resultados

En este capítulo, inicialmente se presenta la fase exploratoria de los registros de cáncer infantil teniendo en cuenta las tres enfermedades de interés. Después se presenta un análisis espacial a partir de patrones puntuales de los casos. Por último, se muestran los resultados de la modelación estadística de las tasas de casos de Leucemia Linfoblastica Aguda en niños menores de 15 años de Santiago de Cali.

4.1. Análisis Exploratorio

Las variables socio-demográficas con las que se trabajó y con las que se conformó la base de datos fueron: Comuna, Sexo, Edad, etnia del paciente y el linaje de las células del LLA. Por medio de un análisis exploratorio se encontró que hay 117 niños y 85 niñas diagnosticados con algún tipo de cáncer infantil de interés, con una edad promedio de aproximadamente 7 años, teniendo en cuenta que se encuentran casos entre los 0.1 meses de nacidos hasta los 15 años.

En la Tabla (**4-1**), se muestra que la mayor cantidad de casos se presentan en niños que en niñas, ademas los diagnosticados con leucemia se encontraron 90 (55,6 %) niños y 72 niñas; el linfoma de burkitt se encontraron 14 (73,7 %) niños y 5 niñas y con meduloblastoma se encontraron 18 (61,9 %) de niños y 8 de niñas.

En la Tabla (**4-2**) se muestra que en todas las enfermedades hay una edad promedio similar en ambos sexos, de la misma forma el comportamiento de las estadística descriptivas no se diferencian por genero. En cuanto a los tipos de cáncer, el meduloblastoma tiene una edad promedio por debajo a la edad promedio de los casos de leucemia y del linfoma burkitt, siendo de 5,7 años la edad promedio de los casos de meduloblastoma y de aproximadamente de 7 años la edad promedio de las otras enfermedades.

Para este estudio se tuvieron en cuenta todos los registros entre el periodo 2009 al 2016, por lo que en la Tabla (**4-3**), muestra, que el año con la mayor cantidad de casos diagnosticados fue el 2016 con 69 casos, seguido de los años 2009 y 2012 de los cuales 30 y 27 pacientes fueron diagnosticados en cada año, respectivamente.

Tabla 4-1: Número de casos para el tipo de cáncer por genero

Género	Leucemia	Burkit	Meduloblastoma	Total Genero
Niño	90	14	13	117
Niña	72	5	8	85
Total Enfermedad	162	19	21	202

Tabla 4-2: Estadísticas descriptivas para la edad por genero

Indicador	Niño	Niña	Totales	Leucemia	Burkit	Meduloblastoma
Promedio	7.1140	7.2910	7.1890	7.3870	7.0430	5.7880
Desviación estándar	4.2424	3.9017	4.0935	4.1920	3.2534	3.8743
Mediana	6.2060	6.4330	6.2310	7.0680	6.0000	4.6000
Mínimo	0.1067	0.6868	0.1067	0.3448	3.659	0.1067
Máximo	15.05	15.52	15.52	15.52	14.02	13.39
Coeficiente de Variación	0.5963	0.5351	0.5694	0.5675	0.4619	0.6693

Tabla 4-3: Numero de casos para los periodos por Tipo de cáncer

Tipo de Cáncer	2009	2010	2011	2012	2013	2014	2015	2016
Linfoma	20	17	17	23	16	2	9	58
Burkit	6	2	1	0	4	0	0	6
Meduloblastoma	4	4	1	4	3	0	0	5
Total	30	23	19	27	23	2	9	69

La ciudad de Santiago de Cali cuenta con 22 comunas, por lo que en la Tabla (**4-4**) se muestra la cantidad de casos registrados en cada una de las comunas por enfermedad y género. De acuerdo a la Tabla (**4-4**), la mayor cantidad de registros se encontró en las comunas 3 y 17; de igual manera, hay comunas en las que se registra menos casos de Linfoma Burkit y Meduloblastomas que los casos de Leucemia, donde como mínimo se registra un caso por comuna. Además se observa que en las comunas 3, 17, 6, 8, 11 y 14 presentan una mayor cantidad de casos con leucemia linfoblastica aguda en comparación de las otras comunas.

En la Tabla (**4-5**), se muestra el número de pacientes por etnia y por cada una de las enfermedades. Se observa que la mayoria de los pacientes con Leucemia Linfoblastica Aguda, con Linfoma de Burkitt y con Meduloblastomas no son afrocolombianos, sin embargo una minoría de 22 y 4 pacientes que son afroamericanos padecen de LLA y Meduloblastomas, respectivamente.

En la Tabla (**4-6**) se encuentra el linaje de células características del cáncer de Leucemia

Tabla 4-4: Estadísticas descriptivas para las Comunas por genero

Comuna	Niño	Niña	Casos	Leucemia	Burkit	Meduloblastoma
comuna 1	6	1	7	6	1	
comuna 2	3	5	8	7	1	
comuna 3	16	3	19	16	2	1
comuna 4	2	3	5	3	2	
comuna 5	3	3	6	5		1
comuna 6	8	4	12	11	1	
comuna 7	5	4	9	7	2	
comuna 8	5	8	13	11	1	1
comuna 9	3	3	6	6		
comuna 10	3	3	6	6		
comuna 11	6	6	12	11	1	
comuna 12	4	0	4	2	1	1
comuna 13	7	4	11	5	2	4
comuna 14	4	8	12	11	1	
comuna 15	8	2	10	6	1	3
comuna 16	5	6	11	9		2
comuna 17	10	5	15	13	2	
comuna 18	5	5	10	7		3
comuna 19	6	5	11	10		1
comuna 20	2	1	3	1	1	1
comuna 21	5	5	10	8		2
comuna 22	1	1	2	1		1

Tabla 4-5: Estadísticas descriptivas para etnia por enfermedad

Etnia	Leucemia	Burkirt	Meduloblastoma
Afrocolombiano	22	0	4
No Afrocolombiano	133	17	13

Tabla 4-6: Estadísticas descriptivas para tipo de células de Leucemia

Tipo de células	Leucemia
Células B	136
Células T	16

Linfoblastica Aguda, ya que en solo esta enfermedad se puede presentar Células B o Células T, mostrando que en la Tabla (**4-6**) existe una mayor cantidad de pacientes que padecen de Leucemia de Células B que de Leucemia de Células T.

4.1.1. Características de las variables, por comunas

Como se observo en la Tabla (**4-4**), los casos de Linfoma de Burkitt y Meduloblastomas no tienen muchos registros, la cantidad de sucesos por comuna para estas dos enfermedades no supera los cuatro casos, por lo que estos dos tipos de cáncer no se tendrán en cuenta para realizar la modelación, ni en el análisis descriptivo de esta sección.

Para los casos de leucemia linfoblastica aguda condicionadas a las comunas de Santiago de Cali, se ve de manera gráfica la cantidad estos datos con respecto al genero, la etnia, el linaje de las células y la edad en dos categorías.

En la Figura (**4-1**), se muestra la cantidad de casos de leucemia con respecto al genero, observando que la comuna con mayor casos de niños es la comuna 3, las comunas 20 y 22 tienen pocos registros de niños; además, las comunas 12 y 15 presentan pocos registros de niñas pero las comunas 8 y 14 presentan la mayor cantidad de niñas que padecen LLA.

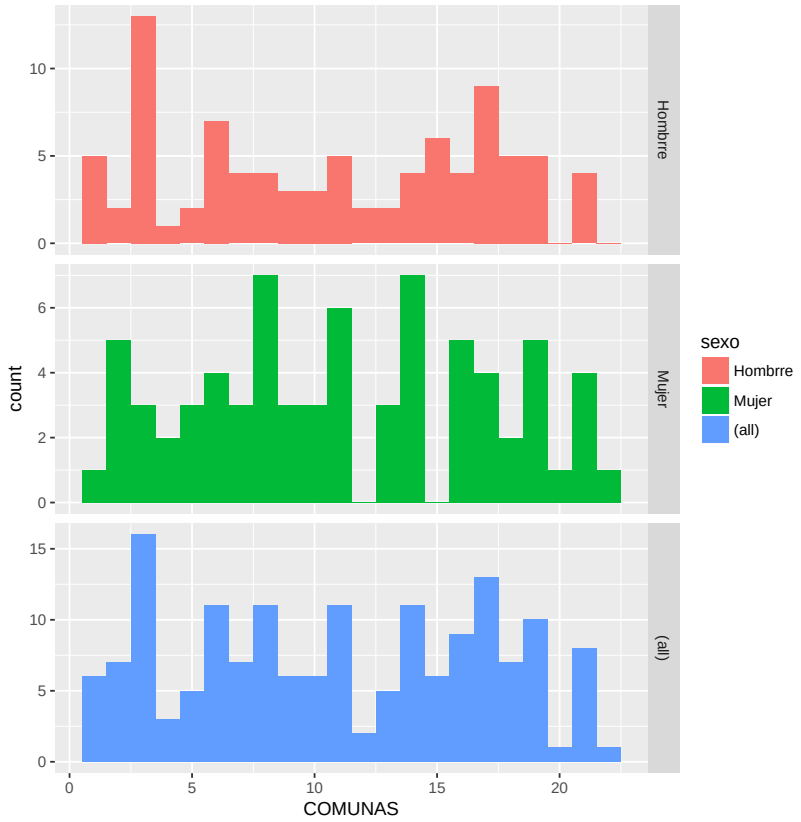

Figura 4-1: Frecuencias del conteo por comuna, de acuerdo al genero de los registros de Leucemia

Como se mencionó anteriormente, la edad se agrupó en dos características: menores de 10 años y mayores e iguales a 10 años hasta los 15 años, este corte se realizó debido a que partir de los 10 años se presentan diferencias biológicas. En la Figura (**4-2**) se muestra la dinámica de la edad por cada comuna. Inicialmente, se muestra que no existen registros de niños menores de 10 años en la comuna 20, mientras que esta frecuencia es la más alta en la comuna 3. Para los niños mayores de 10 años y menores de 15, no se presentaron registros en las comunas 12, 13, 18 y 22 y la mayor cantidad de registros se presentaron en la comuna 3 y 8.

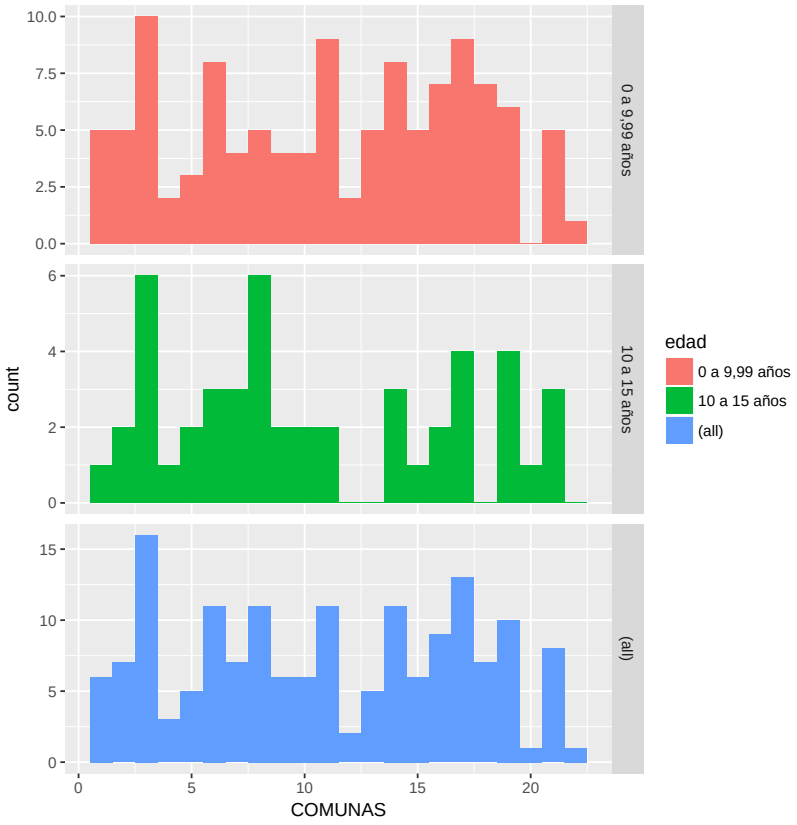

Figura 4-2: Frecuencias del conteo por comuna, de acuerdo a la edad categorizada de los registros de leucemia

Como se mencionó en la Tabla (**4-5**), existen más pacientes que no son de descendencia afrocolombiana; en la Figura (**4-3**) se observa que hay registros que no tienen reconocida una de estas dos características y son en total 7 registros de 162 (4,3 % de datos perdidos). Como se muestra en la Figura (**4-3**) existen muy pocos pacientes afrocolombianos que padecen de LLA y son más las comunas de Cali que no tienen pacientes con esta etnia como las comunas 1, 4, 5, 8, 12, 20 y 22. Las comunas 20 y 22 tienen menos población afrocolombiana, por lo tanto en esta comunas es sensato no encontrar estos casos en particular.

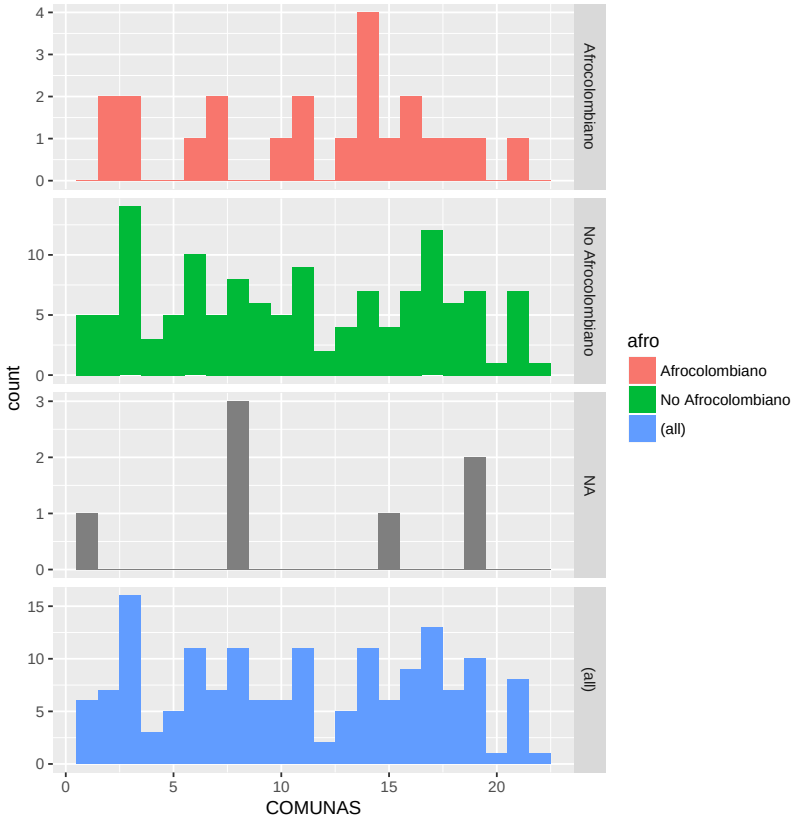

Figura 4-3: Frecuencias del conteo por comuna, de acuerdo a la característica de afrocolombiano de los registros de leucemia

Por último, la variable del linaje de blastos del LLA. La Figura (**4-4**) muestra lo que se encuentra en la Tabla (**4-6**), donde se presentan más registros de cáncer de Leucemia de Células B que de Células T. Los resultados también muestran que son más las comunas que no tiene pacientes con esta característica como las comunas 2, 4, 5, 9, 10, 11, 12, 17, 18, 19, 20 y 22, pero ademas existen 6 pacientes que no tienen este registro del diagnóstico, es decir, el 3.7 % de los datos perdidos en los pacientes con leucemia para esta característica. La mayoría de registros de Células B se encuentran en las comunas 3 y 17, siendo la comuna 3 la que más registros de cáncer de Leucemia tiene. Esta comuna se caracteriza por ser una de las comunas con menos población en la ciudad (Alonso et al., 2007). La comuna 3 se encuentra conformada por los barrios: El Calvario, El Hoyo, El Nacional, El Peñón, El Piloto, La Merced, Los Libertadores, Navarro-La Chanca, San Antonio, San Cayetano, San Juan Bosco, San Nicolás, San Pascual, San Pedro y Santa Rosa.

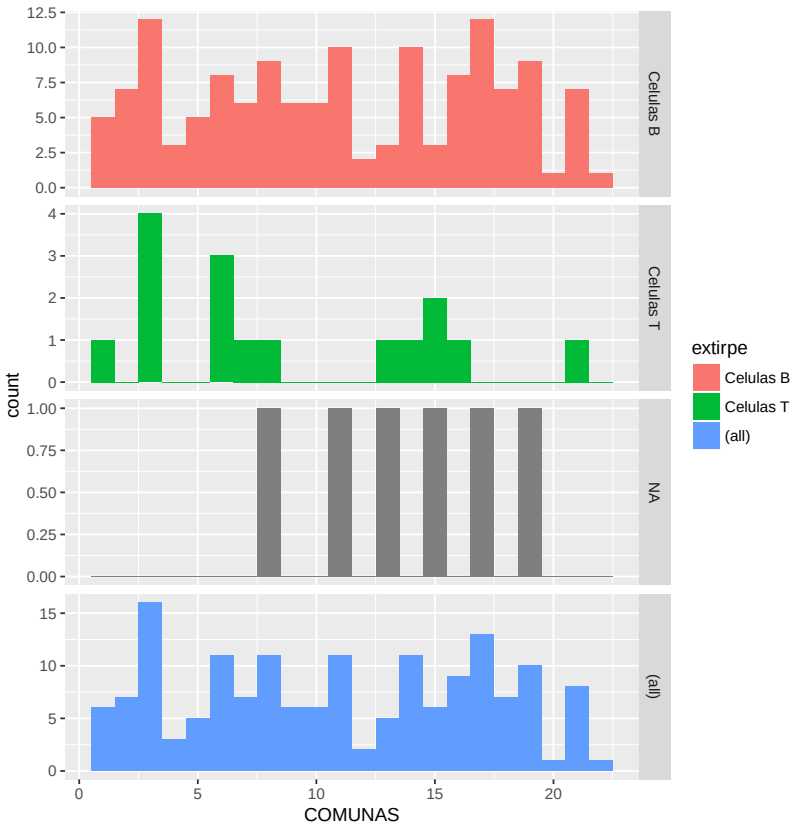

Figura 4-4: Frecuencias del conteo por comuna, de acuerdo a la característica de células B o T de los registros de leucemia

4.2. Análisis Espacial

En la Figura (**4-5**), se muestran los casos de cáncer infantil de interés en el área urbana de Santiago de Cali. A través de patrones puntuales; inicialmente, se realiza una prueba de aleatoriedad basada en cuadrantes y distancias para identificar si existe una aleatoriedad espacial en la ocurrencia de los eventos.

La Figura (**4-5**), muestra que los 162 casos de leucemia se encuentran distribuidos por toda la ciudad, en cambio los 19 registros de Burkit se encuentran más distanciados los unos de los otros presentando una pequeña agrupación en el lado Este de la ciudad y los 21 pacientes que presentan Meduloblastomas se encuentran distribuidas en dos sectores de la ciudad de Santiago de Cali, una gran cantidad se encuentran en el lado Este de la ciudad y una minoría se localizá en el Sur, hacia las laderas de la ciudad.

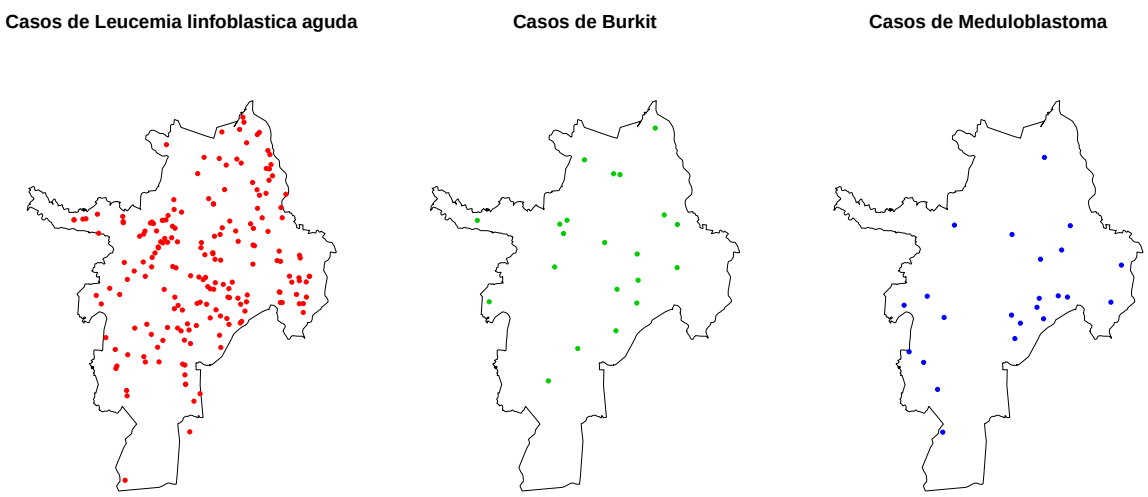

Figura 4-5: Área urbana de la ciudad de Santiago de Cali y los casos de cáncer infantil en el período 2009-2016

4.2.1. Pruebas de Aleatoriedad Espacial

Las pruebas de aleatoriedad espacial se realizan con el fin de conocer el tipo de patrón espacial de los casos de cáncer infantil en el perímetro urbano de Santiago de Cali. En las Tablas (**4-7**), (**4-8**) y (**4-9**), se presenta un resumen de los resultados de todas las pruebas que se realizaron para cada enfermedad estudiada en esta investigación. Los resultados muestran una distribución aleatoria para las enfermedades de Burkitt y Meduloblastomas; para los casos de Leucemia las pruebas de aleatoriedad espacial basada en cuadrantes y en distancia resultaron diferentes. Las pruebas basadas en distancias no presentan un valor-p, ya que son pruebas gráficas en la que se compara la función teórica de aleatoriedad con la función empírica. Las gráficas de las pruebas de aleatoriedad espacial basada en distancias se encuentran en los Anexos.

En la Tabla (**4-7**), se muestra que la mayoría de las pruebas señalan que los casos de Leucemia Linfoblastica Aguda tienen un patrón agrupado en Santiago de Cali.

En cambio, para las enfermedades de Linfoma de Burkitt (Tabla **4-8**) y Meduloblastoma (Tabla **4-9**) se encontró que todas las pruebas de aleatoriedad espacial apuntan a un patrón aleatorio.

Tabla 4-7: Prueba de Aleatoriedad para los casos de Leucemia Linfoblastica Aguda

Método	Prueba	Resultado	Valor-p
En cuadrantes	Test χ^2	Agrupado	0.002729
	Test de Montecarlo	Aleatorio	0.173
En distancias	Función G	Aleatorio	-
	Función F	Agrupado	-
	Función K de Ripley	Agrupado	-
	Función K de L	Agrupado	-

Tabla 4-8: Prueba de Aleatoriedad para los casos de Linfoma de Burkitt

Método	Prueba	Resultado	Valor-p
En cuadrantes	Test χ^2	Aleatorio	0.3445
	Test de Montecarlo	Aleatorio	0.164
En distancias	Función G	Aleatorio	-
	Función F	Aleatorio	-
	Función K de Ripley	Aleatorio	-
	Función K de L	Aleatorio	-

Tabla 4-9: Prueba de Aleatoriedad para los casos de Meduloblastoma

Método	Prueba	Resultado	Valor-p
En cuadrantes	Test χ^2	Aleatorio	0.5615
	Test de Montecarlo	Aleatorio	0.171
En distancias	Función G	Aleatorio	-
	Función F	Aleatorio	-
	Función K de Ripley	Aleatorio	-
	Función K de L	Aleatorio	-

4.2.2. Estimación de la intensidad

La función de intensidad calculada por medio del método Kernel (Figura **4-6**), ayuda a identificar de una manera más rápida y espacial la agrupación de los casos (Espinal and Aruneri, 2014). Los resultados muestran que existe una agrupación de casos de Meduloblastomas en el lado Este de la ciudad. Así mismo, se observa que los casos de Linfoma Burkitt se encuentran más agrupados hacia el Centro de la ciudad y los casos de Leucemia Linfoblastica Aguda se encuentran mas agrupados hacia el Centro, Este y Oeste de la ciudad.

Para tener una estimación de intensidad mas precisa, se calcularon los anchos de banda de

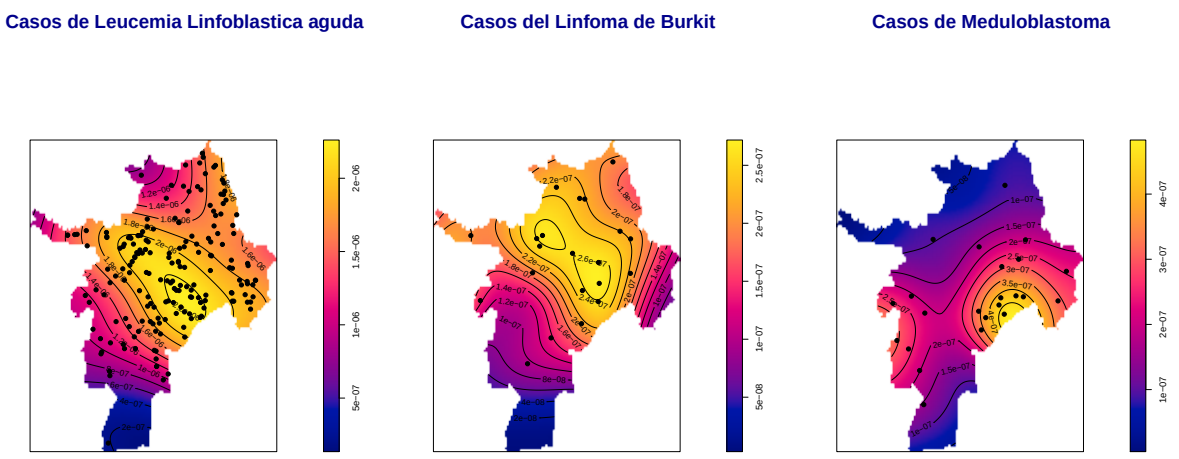

Figura 4-6: Estimación Kernel de la intensidad para los casos de cáncer infantil en Cali

la densidad Kernel por medio del criterio de Validación Cruzada y la Verosimilitud de la Validación Cruzada. En la Tabla (**4-10**), se encuentra el ancho de banda de cada criterio para las enfermedades de interés, identificando que existe un ancho de banda mínimo y un ancho de banda máximo. De acuerdo a los resultados de la Tabla (**4-10**), el criterio de validación cruzada es el mejor estimador de intensidad, ya que maneja un área de intensidad más preciso a los casos presentados. Las gráficas de cada estimación de intensidad para las enfermedades de interés se encuentra en los Anexos.

Tabla 4-10: Ancho de Banda para la estimación de intensidad por medio de los criterios

Método	**Leucemia**	**Burkitt**	**Meduloblastoma**
Validación Cruzada	693.7232	862.4667	806.2189
Verosimilitud de Validación Cruzada	1149.999	4059.314	2404.084

En la Figura (**4-7**), se encuentran las estimaciones de intensidad por el criterio de Validación Cruzada para cada una de las enfermedades de interés. En el lado izquierdo de la Figura (**4-7**), se encuentran los casos de Leucemia Linfoblastica Aguda y se observan altas intensidades en las zonas del Este, Oeste y Noroeste de la ciudad; en el centro de la Figura (**4-7**) se encuentran los registros de Linfoma de Burkitt, la cual se observan en las zonas del Este y Oeste de la ciudad varias intensidades altas; por ultimo en el lado derecho de la Figura (**4-7**) están los casos de Meduloblastoma y presenta en el lado Oeste varias zonas

con intensidades altas.

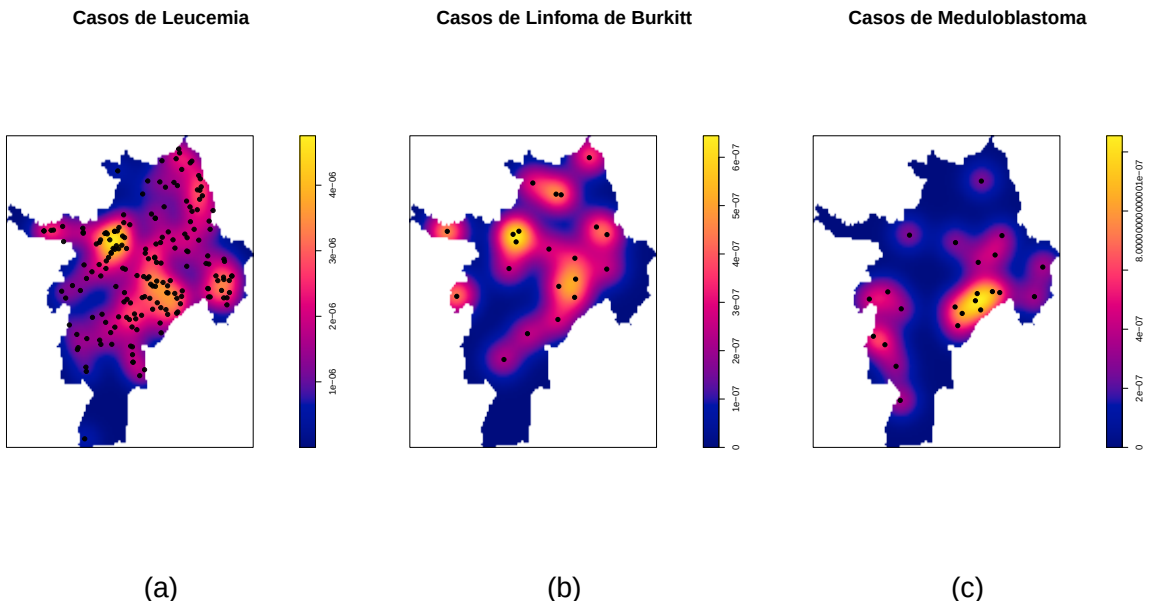

Figura 4-7: Estimación de intensidad por medio del criterio de validación cruzada para los casos de cáncer infantil en Cali

Al realizar un análisis más detallado de las estimaciones de intensidad por medio del criterio de Validación Cruzada. la Figura (**4-7**)(a), muestra las estimaciones de los registros de Leucemia Linfoblastica Aguda, encontrando varias zonas con intensidades altas, en las comunas 3 y 9, específicamente en los barrios Granada, La Merced, Santa Teresita, El Piloto, Santa Rosa, Versalles, El Calvario, El Peñón, Juananbú, San Juan Bosco, San Nicolas, San Pedro y Miraflores. Por otra parte se presentaron intensidades altas en la zona del Oriente y del Distrito de Agua Blanca, donde los casos registrados con esta intensidad pertenecen a las comunas 11, 13, 15 y 16, específicamente a los barrios La Fortaleza, Los Conquistadores, San Cristobal, El Recuerdo, La Independencia, Maracaibo, Agua Blanca, Los Robles, El Diamante, San Pedro Claver, Urbanización Nueva Base, El Poblado I, Ciudad Cordoba, El Retiro, Antonio Nariño, Los Sauces, Mariano Ramos, Cañaverales-Los Samanes, San Judas Tadeo I y Brisas del Limonar. Por último, una intensidad alta en las comunas 14 y 21 los barrios Alfonso Bonilla Aragon, Manuela Beltran, Puerta del Sol, Planta de Tratamiento, José Manuel Marroquin II, Los Naranjos II, Compartir, Calimio Desepaz y Valle Grande.

En la Figura (**4-7**)(b), se muestra la estimación de intensidad por el método de Validación Cruzada para los casos de Linfoma de Burkitt, observando una intensidad alta en la comuna 3 específicamente en los barrios La Merced y Granada; ademas, se logra ver una intensidad alta en los barrios La Fortaleza, Doce de Octubre y El Diamante que son barrios que

corresponden a las comunas 11, 12 y 13 respectivamente. Por último, se encuentran dos intensidades bajas, una esta comprendidas por las comunas 2 y 4, en los barrios La Paz, Vipasa, Olaya Herrera y la segunda intensidad baja se encuentra en la comuna 7 en los barrios La Base Aerea y Alfonzo Lopez III.

La Figura (**4-7**)(c), se muestra la estimación de la intensidad para los casos de Meduloblastomas. Se observa una gran cantidad de casos con esta enfermedad en las comunas 13 y 15 y específicamente en los barrios El Diamante, Los Robles, Nueva Floresta, Antonio Nariño, Mariano Ramos y el Poblado II. Ademas, se observo una baja intensidad en la zona Sur de la ciudad, en los barrios Lourdes, Napoles, El Lido y Brisas de Mayo, las cuales se encuentran ubicados en las comunas 18, 19 y 20, de la ciudad de Santiago de Cali.

4.3. Modelación

En esta sección se presentan los resultados de la modelación de las tasas de casos de Cáncer infantil en Santiago de Cali, en este proceso no se realizó la modelación para los casos de Linfoma de Burkitt y Meduloblastomas ya que son muy pocos los registros que se encuentran en toda la ciudad de Santiago.

En este estudio se trabaja con la tasa de casos de cáncer infantil y por medio de una prueba de bondad de ajuste, se encuentra que la mejor distribución de la familia exponencial que se puede empleara a los errores es una distribución Gamma.

Las variables de este modelo fueron el género (Hombre o Mujer), la edad categorizada (en menores de 10 años y mayores e iguales a 10 años), la etnia, y el linaje de LLA (células B o células T). Como cada variable posee varias categorías, entonces en cada una se elije un grupo base (conocido generalmente como ̈grupo no expuesto"). Para el género fue la categoría mujer, para la edad fue los mayores e iguales a 10 años, para la etnia la categoría base son los afrocolombianos y finalmente para las células precursoras fueron las células T. Junto a estas variables se incorporó la variable de densidad de niños menores de 15 años por cada comuna, logrando caracterizar a la comuna en riesgo.

Al realizar un modelo simple con cada una de las variables como se muestra en la Tabla (**4-11**), donde cada variable tiene un aporte individual para el modelo, siendo estas variables significativas de acuerdo al valor p, en el momento de realizar la comparación de estos modelos con el nullo. Por medio del AIC se identifica que hay dos modelos que tiene el mismo valor del criterio, que corresponde a los modelos 3 y 5, el cual consta de las variables de Etnia y Densidad de Niños, siendo estas dos variables una opción para lograr explicar la tasa de casos de Leucemia Linfoblastica Aguda.

Tabla 4-11: Aporte individual de cada variable a la tasa de casos de LLA

Modelo	Estimación del modelo	AIC	Modelo Nulo(valor p)
Modelo 1	$ln(\hat{\mu}) = 2{,}8660 + 0{,}1271(\text{Hombres})$	187.4	0.0003072
Modelo 2	$ln(\hat{\mu}) = 2{,}7211 + 0{,}1347(< 10\text{años})$	191.6	0.01074
Modelo 3	$ln(\hat{\mu}) = 2{,}5759 + 0{,}1317(\text{NoAfrocolombiano})$	185.4	0.0002065
Modelo 4	$ln(\hat{\mu}) = 2{,}5189 + 0{,}1355(\text{CelulasB})$	186.7	0.001261
Modelo 5	$ln(\hat{\mu}) = 4{,}23474 - 0{,}01336(\text{DensidadNiños})$	185.4	0.0008161

La interpretación de las estimaciones de los $\vec{\beta}$, es que por ejemplo en el modelo 1, donde a medida que aumenta la cantidad de hombres en una unidad, el logaritmo de la tasa de casos aumenta en un 0.1271 unidades, y de esta misma forma se interpretan los modelos 2, 3 y 4; la estimación del modelo 5 encontramos que a medida que aumenta la densidad de niños en una unidad, el logaritmo de la tasa de casos de LLA, disminuye en un 0.013336 unidades, indicando de esta forma que a mayor densidad de niños menores de 15 años, la tasa de la enfermedad va disminuyendo, lo cual seria algo incoherente, ya que la población en riesgo incrementa.

Al ver de que todas las variables aportaban de manera individual al modelo, se planteo el modelo completo, para la tasa de casos de LLA como se muestra en la ecuación (4-1).

$$\eta = ln(\vec{\mu}) = \beta_0 + \beta_1(\text{Hombre}) + \beta_2(< 10\text{años}) + \beta_3(\text{NoAfro}) + \beta_4(\text{CelulasB}) + \beta_5(\text{DensidadNiños}) \tag{4-1}$$

Es decir que en modelo de Leucemia se utilizó la función de enlace logarítmica, como se muestra en la ecuación (4-1). El modelo queda expresado en la ecuación (4-2).

$$\begin{aligned} Ln(\hat{\mu}) = {} & 3{,}249289 + 0{,}044744(\text{Hombre}) + 0{,}050567(< \text{de 10 años}) + 0{,}044158(\text{No afrocolombianos}) \\ & + 0{,}024331(\text{Células B}) - 0{,}013221(\text{Densidad Niños}) \end{aligned} \tag{4-2}$$

Una vez estimado el modelo de regresión por máxima verosimilitud a través del método iterativo de Fischer Scoring, donde en 8 interacciones se logro la convergencia, se realizarón las pruebas de Bondad de Ajuste, En la Tabla (**4-12**) se muestran las estimaciones, los resultados muestran que el modelo completo no presenta un buen ajuste, el valor p de esta prueba resultó ser superior a un $\alpha = 0{,}05$, respecto al R^2 vemos que el modelo explica el 83.6 % de la variabilidad de las tasas de casos de Leucemia Linfoblastica Aguda en Santiago

Tabla 4-12: Análisis de Desvianzas para el modelo completo

Dif. desvianza	chi-cuadrado	valor p	conclusión	R^2	AIC
10.1885	26.29623	0.1434	mal ajuste	0.835418	165.54

de Cali.

En la Tabla (**4-13**) se encuentra el análisis de desvianzas para el modelo completo, y en la Tabla (**4-14**) se encuentra las estimaciones del modelo completo. Se concluye que ninguna de las variables categóricas son significativas para el modelo con un $\alpha = 0,05$, en cambio la variable densidad de Niños, es significativa para el modelo con un $\alpha = 0{,}05$, siendo esta la variable que más aporta en la explicación de las tasas de casos de Leucemia infantil.

Tabla 4-13: Análisis de Desvianzas para el modelo completo

	Df	Deviance Resid.	Df	Resid. Dev
NULL			21	12.1884
Genero	2	4.7258	20	7.4626
Edad	2	0.0767	19	7.3859
Afro	2	1.5925	18	5.7933
Celulas	2	0.5001	17	5.2932
DenNinos	1	3.2933	16	1.9999

Tabla 4-14: Estimaciones del modelo completo

	Estimate	Std. Error	Z	Pr($>\|z\|$)	
(Intercept)	3.249289	0.221821	14.648	1.09e-10	***
Hombres	0.044744	0.066592	0.672	0.511	
Edad1	0.050567	0.092791	0.545	0.593	
NoAfro	0.044158	0.100018	0.441	0.665	
CelulasB	0.024331	0.084328	0.289	0.777	
DenNinos	-0.013221	0.002495	-5.299	7.20e-05	***

En la Tabla (**4-15**), se observan los resultados del indicador VIF. Los resultados muestran que existe multicolinealiedad entre las variables explicativas.

Al buscar un modelo que nos permita explicar la variabilidad de las tasas de casos de Leucemia Linfoblastica Aguda en Santiago de Cali, se realizá una selección de variables utilizando el método *Stepwise* y el criterio AIC.

Tabla 4-15: Raíz cuadrada del factor de inflación de la varianza (VIF)

Hombres	Edad1	NoAfro	CelulasB	DenNinos
2.680701	3.384586	4.438534	3.777086	1.311895

Tabla 4-16: Proceso de selección de variables *Srepwise*

Variables del Modelo	AIC
Hombres+Edad1+NoAfro+CelulasB+DenNinos	165.54
Hombres+Edad1+NoAfro+DenNinos	163.65
Edad1+NoAfro+DenNinos	162.15
NoAfro+DenNinos	161.03

En la Tabla (**4-16**), se muestra los resultados del proceso de Stepwise y los AIC obtenidos. El Menor AIC indica que el modelo con las variables no afrocolombiano y densidad e niños tiene el menor valor de este indicador. El modelo final queda expresado como:

$$\eta = ln(\vec{\mu}) = \beta_0 + \beta_1(\text{NoAfro}) + \beta_2(\text{DensidadNiños}) \tag{4-3}$$

Al igual que el modelo completo, a este modelo se le realizó las pruebas de Bondad de Ajuste. En la Tabla (**4-17**) se muestran los resultados.

Tabla 4-17: Análisis de Desvianzas para el modelo seleccionado

Dif. desvianza	chi-cuadrado	valor p	conclusión	R^2	AIC
10.05	30.14353	0.0482984	buen ajuste	0.824838	161.03

De acuerdo al valor p de la Tabla (**4-17**), se puede observar que este modelo presenta un mejor ajuste. El valor del R^2 es más pequeño que el del modelo completo, pero logra explicar un 82.5 % de la variabilidad de las tasas de casos de Leucemia. El AIC como era de esperarse, es más pequeño que el del modelo completo. Siendo este el modelo elegido para predecir las tasas de los casos de Leucemia Linfoblastica Aguda. La ecuación (4-4), muestra el modelo ajustado. Teniendo en cuenta que estos casos presentan una agrupación espacial, el cual no se tomo en cuenta durante este proceso, se considera que es un tema importante y queda para futuras investigaciones.

$$ln(\hat{\mu}) = 3{,}195575 + 0{,}140042(\text{No Afrocolombianos}) - 0{,}012027(\text{Densidad Niños}) \tag{4-4}$$

En la Tabla (**4-18**) se observa el análisis de desvianza para el modelo de la ecuación (4-4) y en la Tabla (**4-19**) están las estimaciones de los parámetros del modelo, donde se concluye que la variable categorizada de Afrodecendencia y la variable densidad de niños menores de 15 años, son significativas para el modelo a un $\alpha = 0{,}05$. Siendo estas variables las que más aportan en la explicación de la tasa de casos de Leucemia Linfoblastica Aguda en niños menores de 15 años de la ciudad de Santiago de Cali.

Tabla 4-18: Análisis de Desvianzas para el modelo completo

	Df	Deviance Resid.	Df	Resid. Dev
NULL			21	12.1884
Afro	2	5.3441	20	6.8443
DenNinos	1	4.7059	19	2.1384

Tabla 4-19: Resumen del modelo completo

	Estimate	Std. Error	Z	Pr($>\|z\|$)	
(Intercept)	3.195575	0.189153	16.894	6.68e-13	***
NoAfro	0.140042	0.021278	6.581	2.67e-06	***
DenNinos	-0.012027	0.001796	-6.697	2.11e-06	***

En la Tabla (**4-20**) se muestra que las variables ingresadas en el último modelo planteado no presentan el problema de multicolinealidad, siendo este el modelo con AIC más bajo

Tabla 4-20: Raíz cuadrada del factor de inflación de la varianza (VIF)

No Afrocolombiano	Dencidad Niños
1.000847	1.000847

Para validar el supuesto de homogeneidad de varianza, se realizó el gráfico de los residuales versus los valores predichos para las tasas de casos de leucemia. En la Figura (**4-8**), se observa que no existe ningun tipo de comportamiento que permita evidenciar un problema de heterogeneidad.

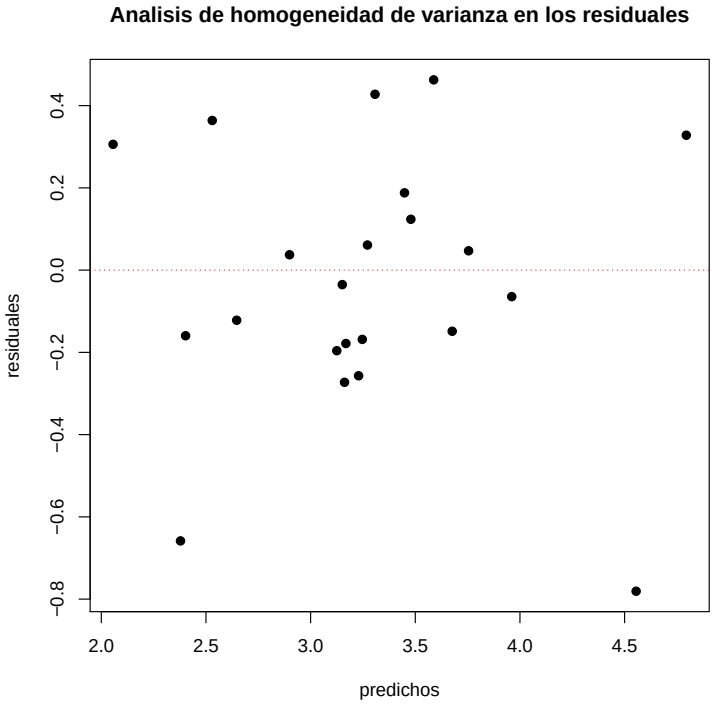

Figura 4-8: Validación de los supuestos para el modelo seleccionado

5 Conclusiones y Recomendaciones

5.1. Conclusiones

Según Bravo et al. (2013), se identificaron 1548 casos nuevos de cáncer infantil en Santiago de Cali entre los años 1992-2011, lo que en promedio equivale a 77 casos nuevos por año. Durante este estudio se identificaron 302 casos de cáncer de Leucemia Linfoblastica Aguda, Linfoma de Burkitt y Meduloblastomas durante los años 2009-2016, obteniendo en promedio de 25 casos por año.

Este estudio se desarrolló usando la base de datos del Registro Poblacional de Cáncer de Cali (RPCC) y trabajando con los registros de la ciudad de Santiago de Cali, donde se encontraron 162 casos de Leucemia Linfoblastica Aguda, 19 casos de Linfoma de Burkitt y 21 casos de Meduloblastoma. Donde además se encontró que 57,92 % son niños y el 42,08 % son niñas, concluyendo que la mayoría de los menores que presenta algún tipo de cáncer infantil son hombres.

Con respecto a la incidencia espacial de los casos de cáncer infantil estudiados en este proyecto y en la ciudad de Santiago de Cali; se encontró que para los casos de Leucemia Linfoblastica Aguda existe una gran cantidad de registros en las comunas 3, 11, 13, 14, 15, 16 y 21, abarcando principalmente los barrios: Granada, La Merced, El Piloto, El Valvario, El Peñon, Juananbu, San Juan Bosco, San Nicolas, La Fortaleza, El Recuerdo, Agua Blanca, Los Robles, El Diamante, San Pedro Claver, Urbanización Nueva Base, Manuela Beltran, José Manuel Marroquin, El Poblado, Ciudad Cordoba, Mariano Ramos, Cañaveralejo-Los Samanes, San Judas Tadeo I y Brisas del Limonar; pero en la comuna 3 existe una mayor cantidad (n=16) de estos casos, la cual se caracteriza por ser una de las comunas con menos población en la ciudad (Alonso et al., 2007), existe la sospecha de que esta gran cantidad de casos de LLA en la comuna 3 deba a que se registran las direcciones de las casas en las que se encuentran de paso. Para los casos de Linfoma de Burkitt, la incidencia se encuentro en las comunas 3, 4, 7, 11 y 13 que abarca principalmente los barrios: La Merced, Granada, La Fortaleza, Doce de Octubre, Diamante, La Paz, Vipasa, Olaya Herrera, Base Aerea y Alfonzo Lopéz III. Por ultimo para los registros de Meduloblastomas se tiene una gran cantidad de casos en las comunas 13, 15 y 18 que abarca principalmente los barrios: El Diamante, Los Robles, Nueva Floresta, Antonio Nariño, Mariano Ramos, Antonio Nariño, El Poblado II, Lourdes y Napoles.

A nivel espacial, se encuentra que el patrón espacial que siguen los casos de Leucemia Linfoblastica Aguda son agrupadas en sectores específicos de la ciudad, como se mencionó anteriormente, los casos de Linfoma de Burkitt siguen un patrón aleatorio y por último para los casos de Meduloblastomas siguen un patrón aleatorio, es decir, al azar en el espacio; y este puede deberse a los pocos casos registrados de estas enfermedades, pero aun así estas dos enfermedades tienen algunas prevalencias en algunas comunas como se mencionó anteriormente.

La Estimación de la función de intensidad de los casos de cáncer infantil analizados en esta investigación, mostró que el ancho de banda por medio del criterio de la validación cruzada son más precisas en el momento de estimar la intensidad de los casos. La comuna 3 es la que mÃ¡s registros tiene de casos de Leucemia y Linfoma de Burkitt; esto puede deberse a las condiciones socioeconómicas o ambientales en las que se encuentran algunos barrios. De acuerdo a la Organizacón Panamerica de la Salud (2016) la tasa de sobrevida son mas bajas, para niños que viven en entornos de bajos recursos, donde aproximadamente uno de cada dos niños fallecen por el cáncer en America, pero no hay muchos estudios que muestren las aréas de zona urbana donde se presenten mucha ocurrencia de cáncer infantil.

Durante el proceso de esta investigación se encontró que hay muy pocos registros de Linfoma de Burkitt y MeduloBlastomas como para realizar un modelo. Por lo tanto, en la modelación realizada para los registros de Leucemia, se determinó que los modelos lineales generalizados con variable de respuesta Gamma y función de enlace Logaritmica, presentan un buen desempeño para modelar las tasas de Casos de Leucemia Linfoblastica Aguda, a través de variables que la explicaran, mostrando un buen ajuste del modelo con las variables: la etnia de los pacientes y de la densidad de niños menores de 15 años. Logrando explicar un alto porcentaje de la variabilidad de las Tasas de estos casos por área.

5.2. Limitaciones del Estudio

La principal limitación que se presento es la dirección, ya que 42 registros no cumplián con el criterio de inclusión de tener una dirección clara y dentro del área urbana de Santiago de Cali, ademas, se asume que la dirección que dan los pacientes es la misma donde nacieron, por lo que estas enfermedades se originan durante toda la infancia. Por ultimo, muchas de las personas registran la dirección de los hogares de paso en los que se encuentran y no la dirección de donde nacieron o residen.

Otra limitación que se presento en este estudio, fue toda la informaición que se perdió (Los casos de Linfoma de Burkitt y Meduloblastoma) en el momento de realizar la depuración de los datos, obstaculizando la realización de los modelos de cáncer de interés, por lo que se

sugiere realizar otro tipo de estudio para estas enfermedades con tan poca prevalencia.

5.3. Recomendaciones

Se recomienda realizar un modelo con una agregación espacial para los errores, ya que los casos de Leucemia Linfoblastica Aguda presentan agrupaciones espaciales. También se puede realizar un modelo con variables etiológicas que logren explicar este fenómeno de cáncer infantil.

6 Anexos

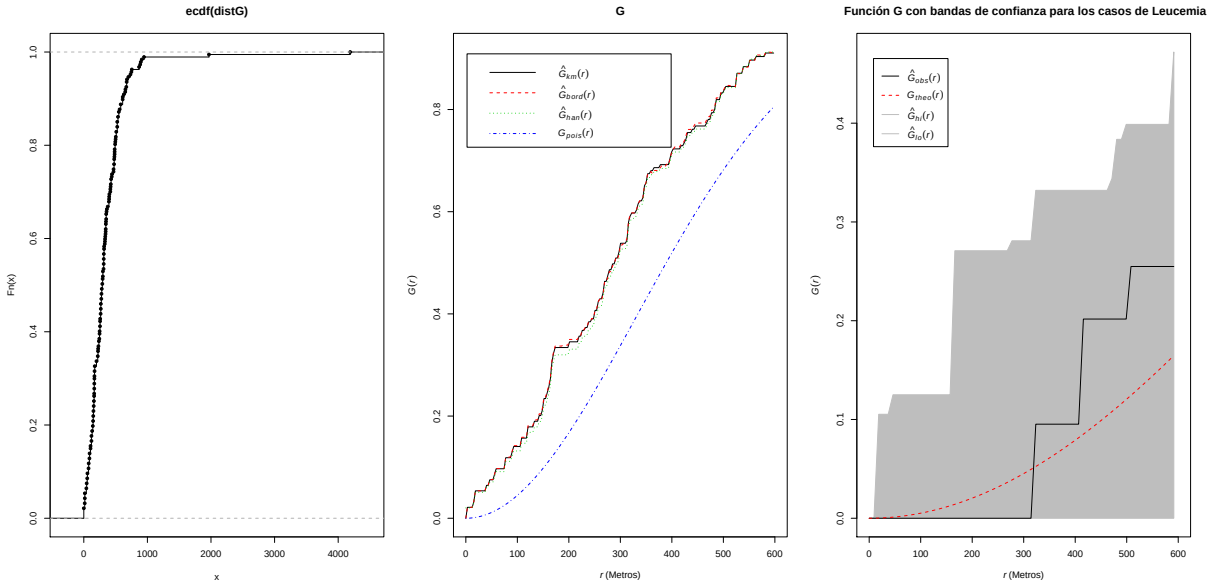

Figura 6-1: Función G Leucemia

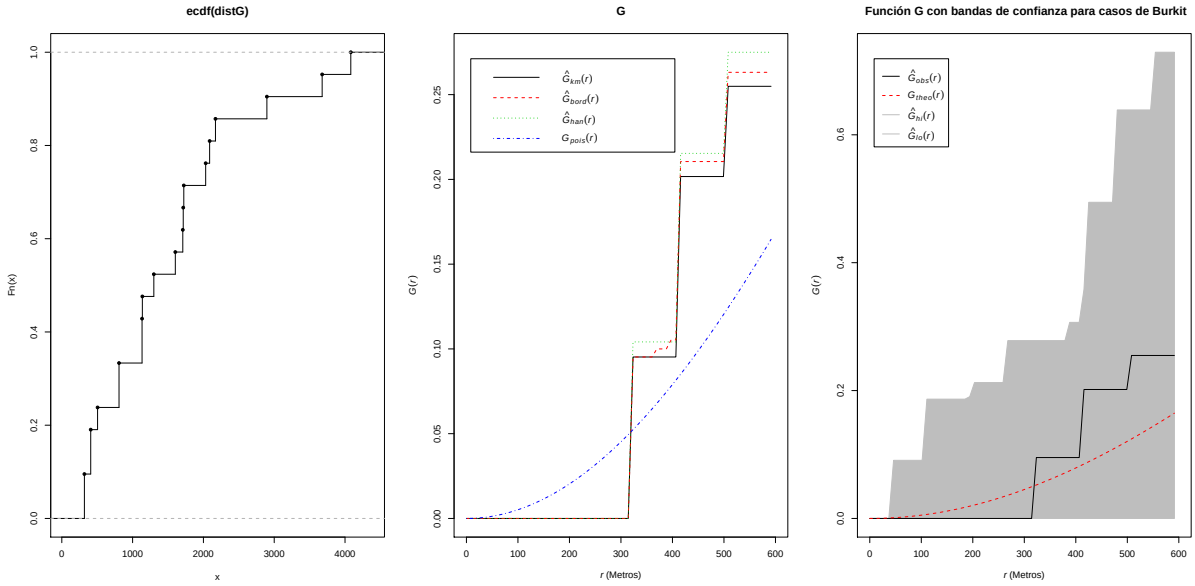

Figura 6-2: Función G Burkitt

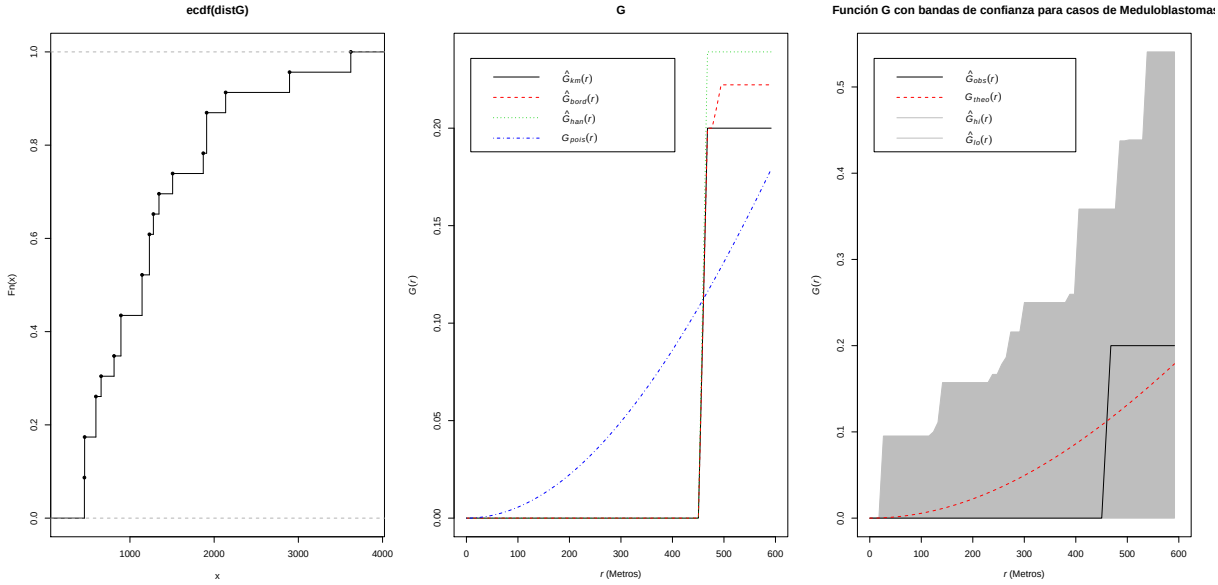

Figura 6-3: Función G Meduloblastoma

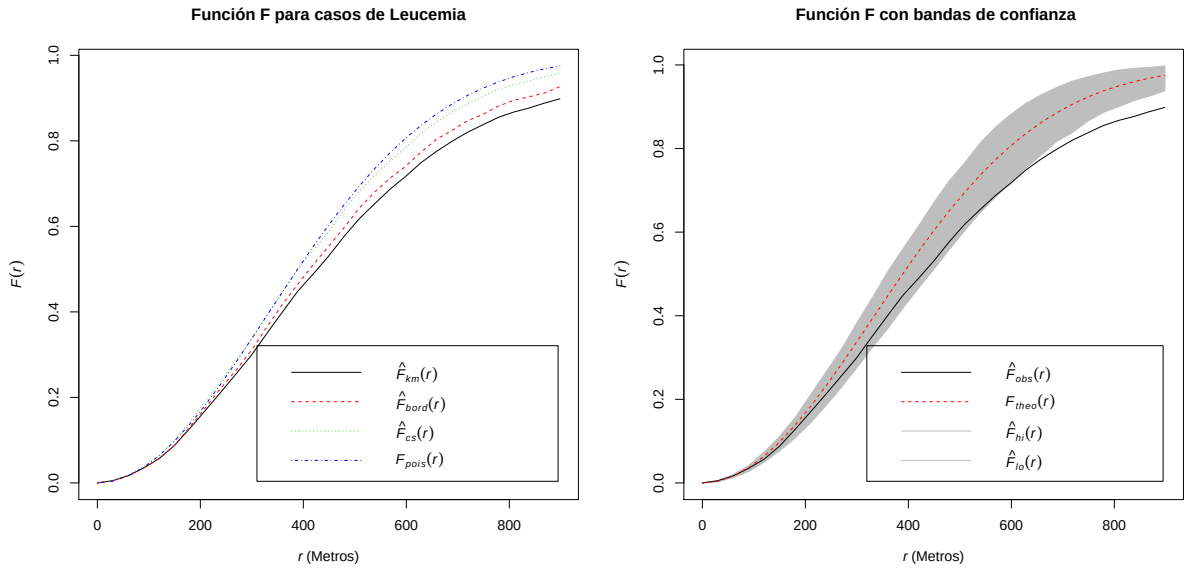

Figura 6-4: Función F Leucemia

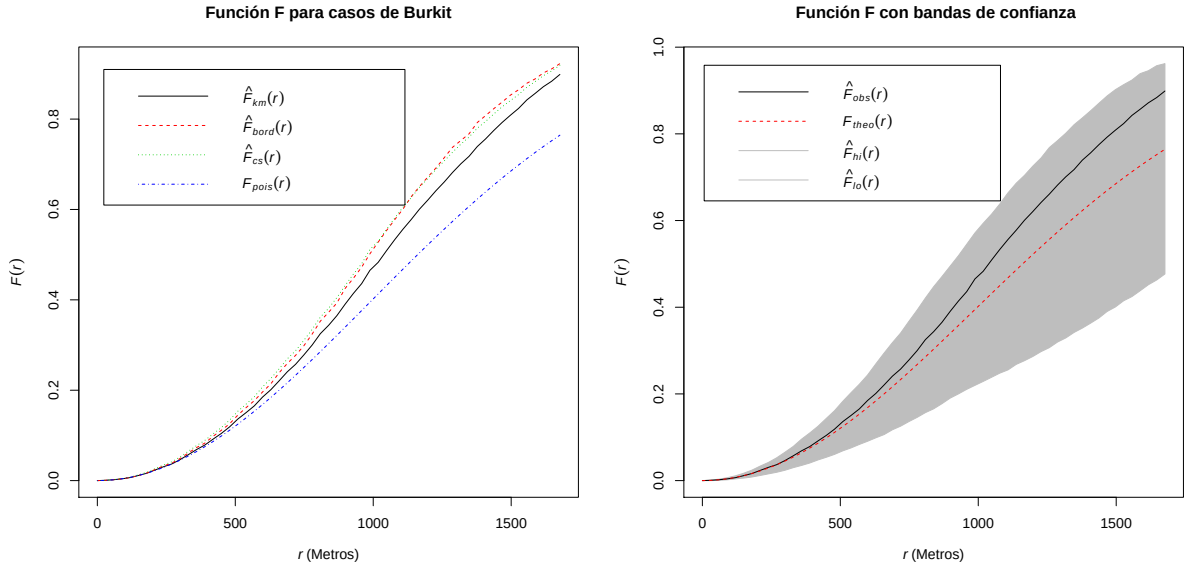

Figura 6-5: Función F Burkitt

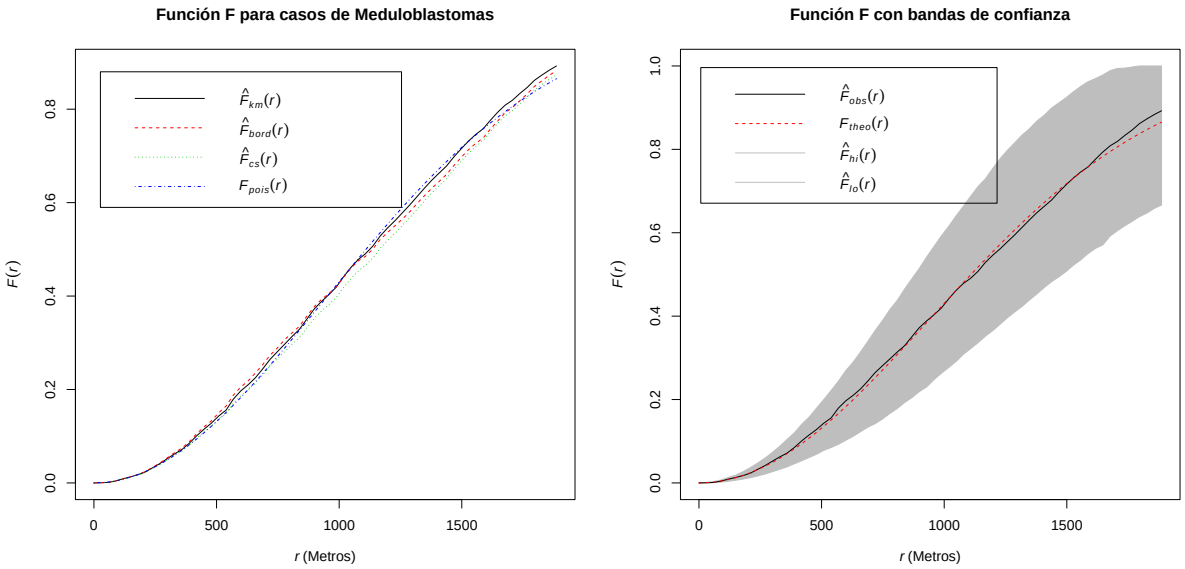

Figura 6-6: Función F Meduloblastoma

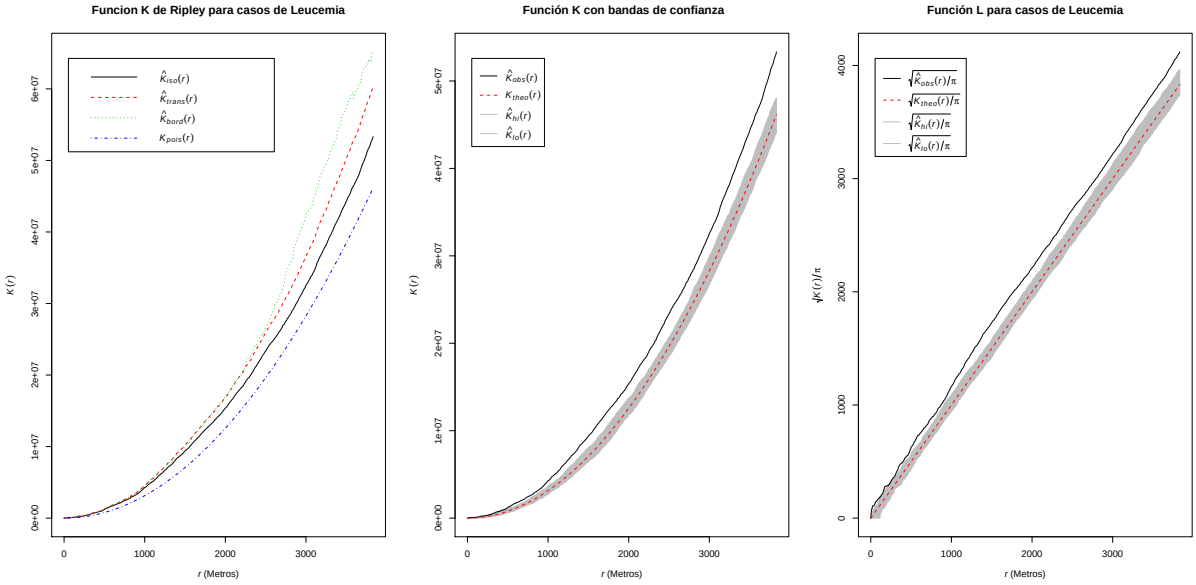

Figura 6-7: Función K Leucemia

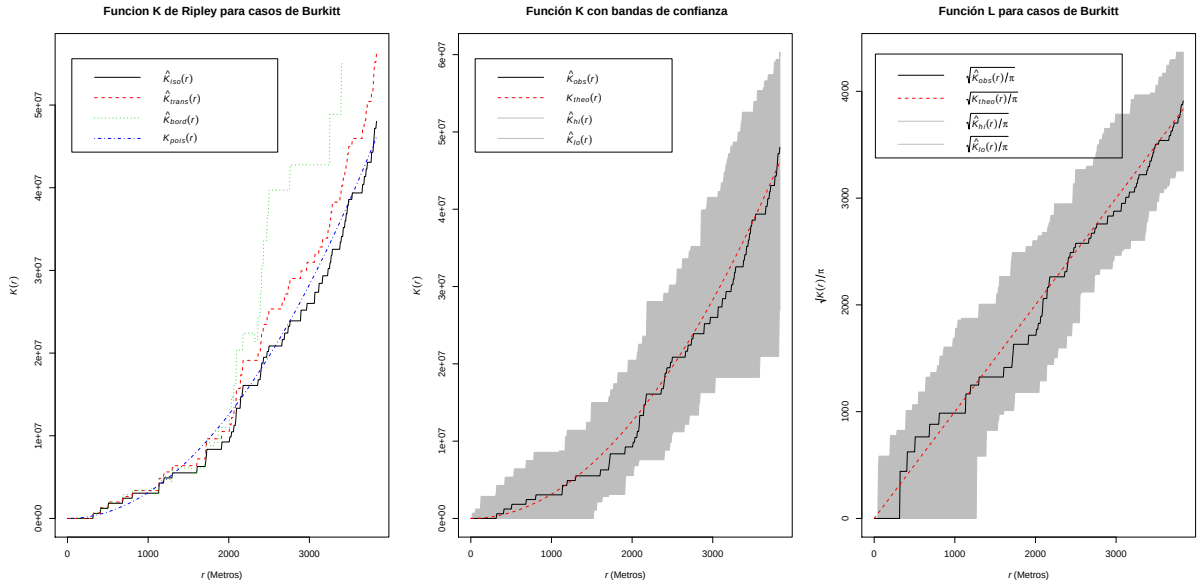

Figura 6-8: Función K Burkitt

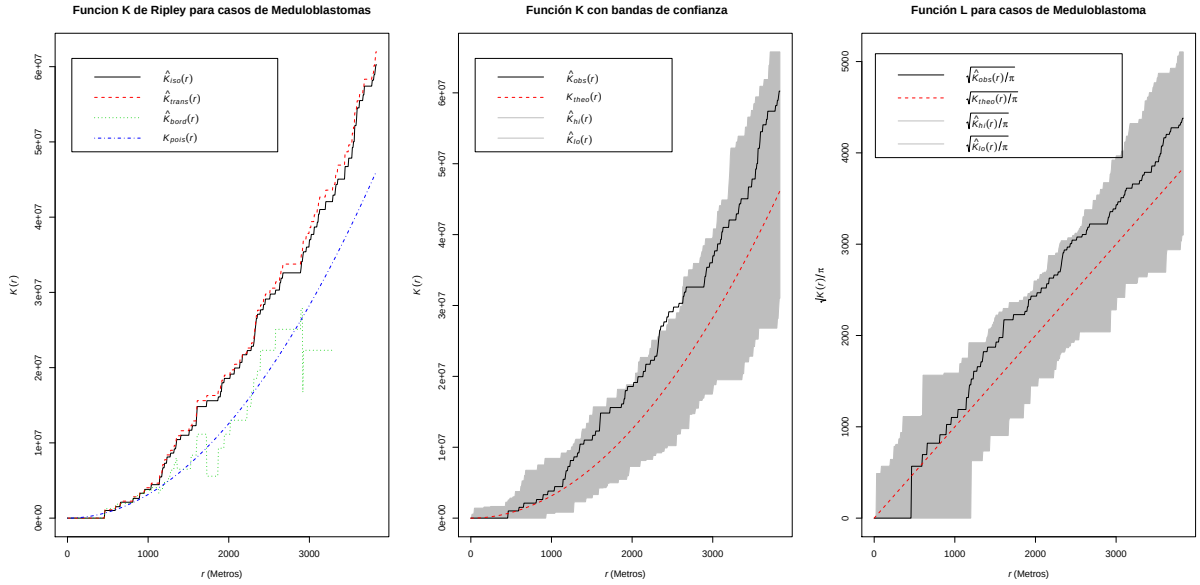

Figura 6-9: Función K Meduloblastoma

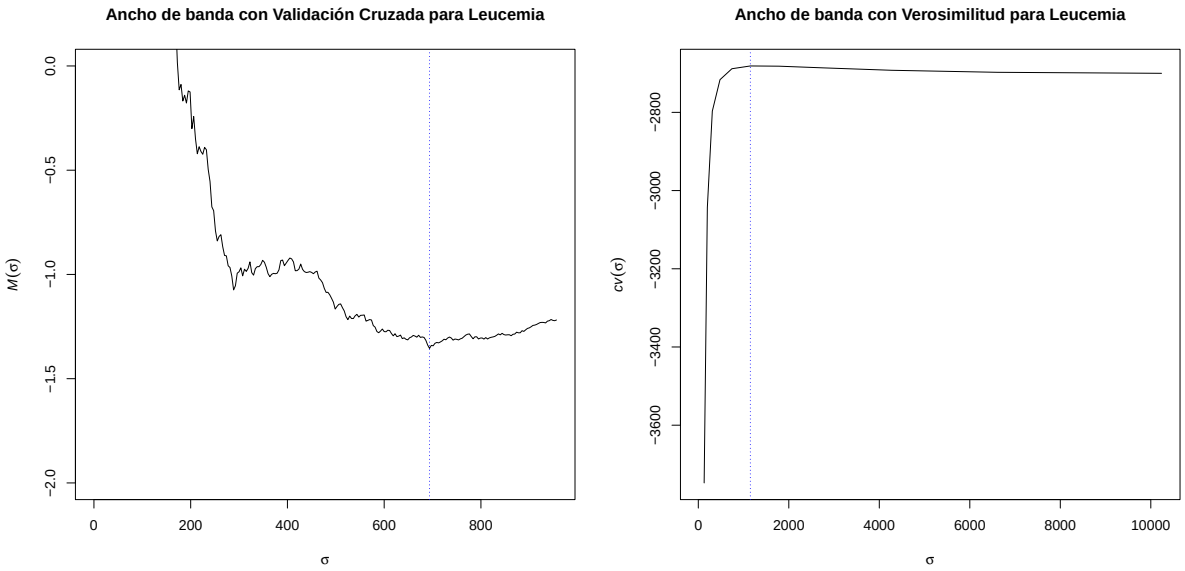

Figura 6-10: Ancho de Banda para los casos de Leucemia

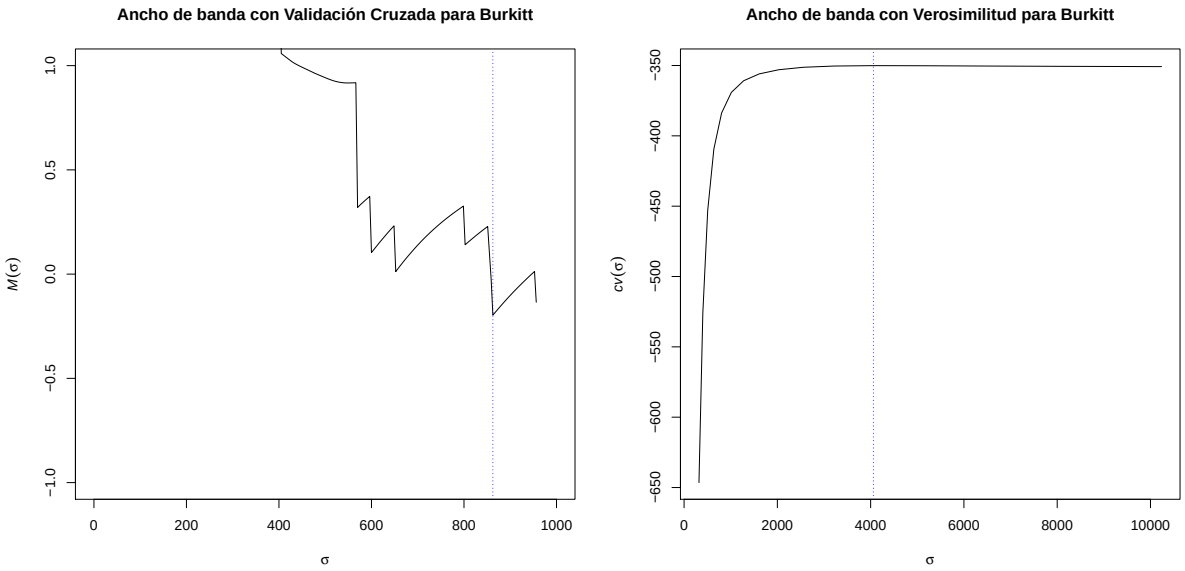

Figura 6-11: Ancho de Banda para los casos de Burkitt

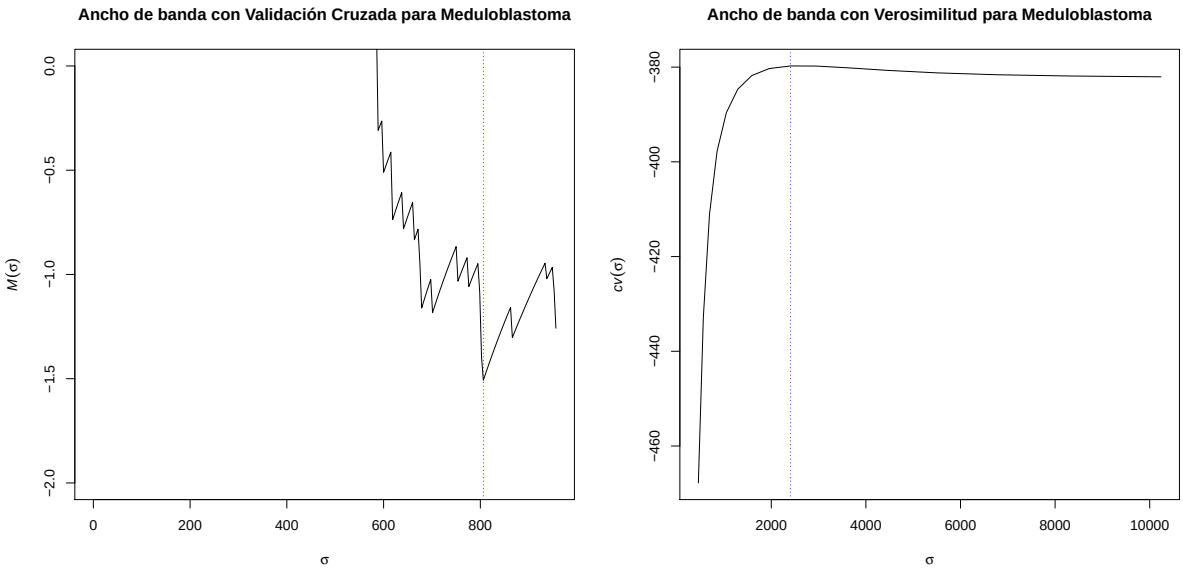

Figura 6-12: Ancho de Banda para los casos de Meduloblastoma

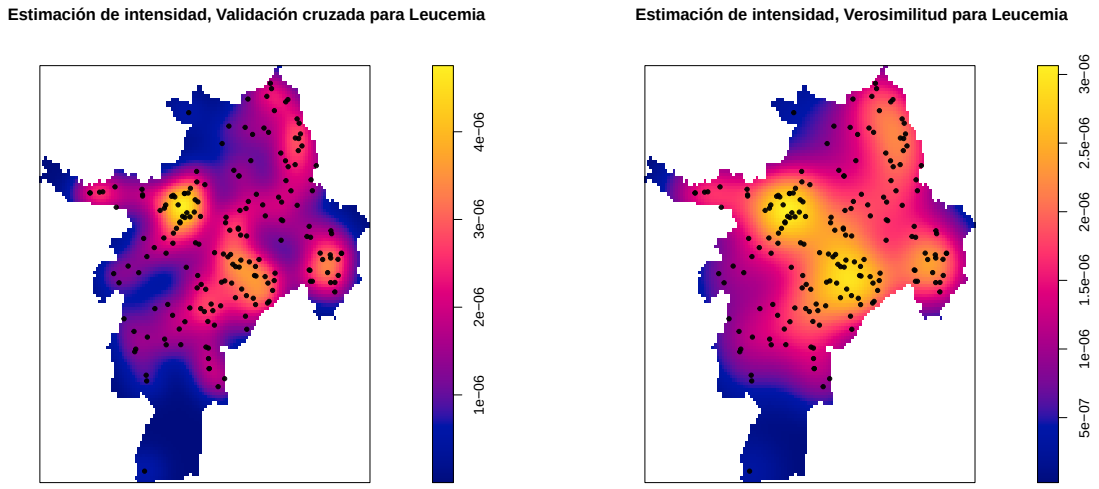

Figura 6-13: Estamación de intensidad por ambos métodos para los casos de Leucemia

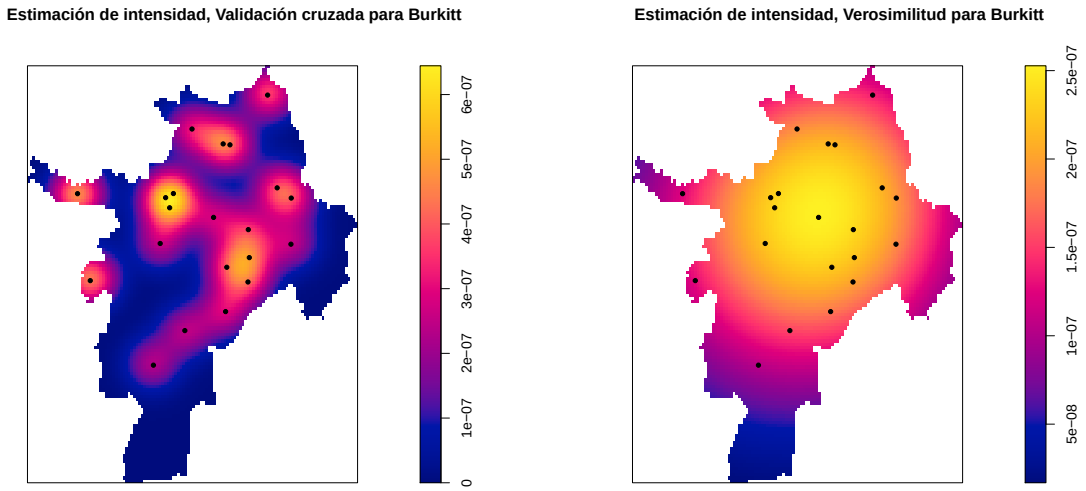

Figura 6-14: Estimación de Intensidad por ambos métodos para los casos de Linfoma de Burkitt

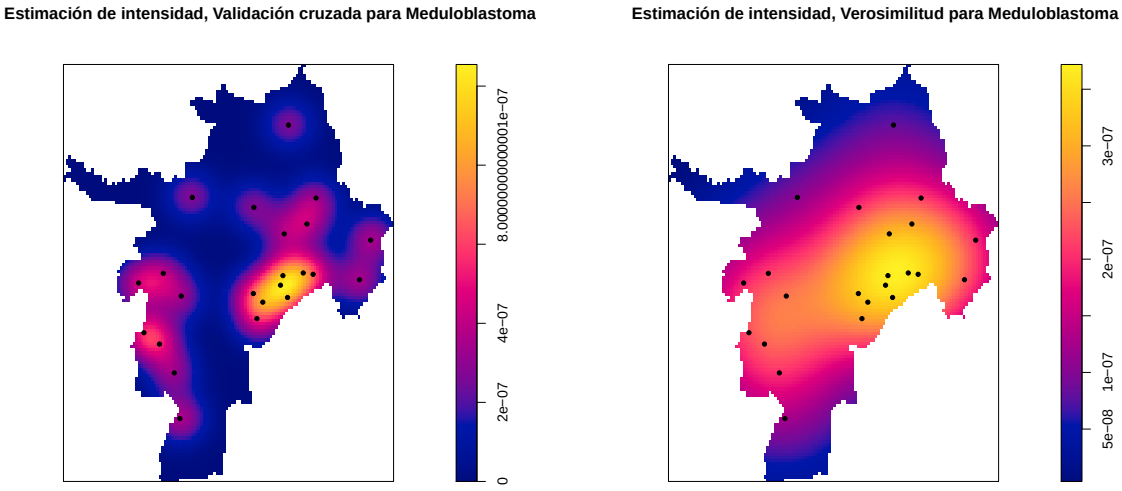

Figura 6-15: Estimación de Intensidad por ambos métodos para los casos de Meduloblastomas

Bibliografía

Agresti, A. (2007). An introduction to categorical data analysis, 2nd edn. Hoboken.

Alcaldia de Santiago de Cali (2017). Datos basicos de santiago de cali, `www.cali.gov.co`, 3 de Agosto de 2017.

Alonso, J. C., Solano, M., Vera, R., and Gallego, A. I. (2007). *Una mirada descriptiva a las comunas de Cali*. Cali: Universidad Icesi.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769.

Ángel, J., Kizys, R., and Manzanedo, L. (2015). Regresión logística binaria. *Universitat Oberta de Catalunya: 1-17*.

Arroyo, I., Bravo, L. C., Llinás, H., and Muñoz, F. L. (2014). Distribuciones poisson y gamma: Una discreta y continua relación. *Prospectiva*, 12(1):99–107.

Ayati, E. and Abbasi, E. (2011). Investigation on the role of traffic volume in accidents on urban highways. *Journal of safety research*, 42(3):209–214.

Bianco, A. M. (2010). Modelo Lineal Generalizado. *FCEN, Notas de Clase*, pages 1–42.

Borja-Aburto, V. H. (2000). Estudios ecológicos. *Salud Pública de México*, 42(6):533–538.

Bravo, L. E., García, L. S., Collazos, P., Aristizabal, P., and Ramirez, O. (2013). Epidemiología descriptiva de cáncer infantil en Cali, Colombia 1977-2011. *Colombia Médica*, 44(3):155–164.

Carroll, W. L. and Bhatla, T. (2016). *Lanzkowsky's manual of pediatric hematology and oncology*, chapter Acute Lymphoblastic Leukemia, pages 389–411. Academic Press.

Cayuela, L. (2010). Modelos lineales generalizados (glm). *Materiales de un curso del R del IREC, Notas de clase*.

Cayuela, L. (2014). Modelos lineales: Regresión, anova y ancova. *Eco Lab, Centro Andaluz de Medio Ambiente, Universidad de Granada. Notas de clase*, pages 1–57.

Cayuela, L., Guillen, M., and Bolancé, C. (2016). introducción GLMs- Función vínculo. *Universidad de Valencia, Notas de Clase*, pages 1–24.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Densidad de población (2017). Densidad de población, sielocal transparencia económica, `http://www.sielocal.com`, 30 de Octubre de 2017.

Dobson, A. J. and Barnett, A. (2008). *An introduction to generalized linear models*. CRC press.

Durá, N., Ramárez, E. B., Cedrá, T., et al. (2016). Exploración espaciotemporal del riesgo de enfermar de leucemia aguda en niños. *Revista Cubana de Salud Pública*, 42(4).

Elandt-Johnson, R. C. (1997). La definición de tasas: Algunas precisiones acerca de su correcta e incorrecta utilización. *Salud pública de México*, 39(5):474–479.

Espinal, E. K. and Aruneri, 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 cartographic modelling using functions kernel for the optimal location of mental health centers, requiring limeñas psychologically assaulted by your partner. *Revista ECIPerú*, 11(1):1–8.

Fajardo-Gutiérrez, A., Mejía-Aranguré, J. M., Hernández-Cruz, L., Mendoza-Sánchez, H. F., Garduño-Espinosa, J., and Martínez-García, M. d. C. (1999). Epidemiología descriptiva de las neoplasias malignas en niños. *Revista Panamericana de Salud Pública*, 6(2):75–88.

Fernández, P. (1995). Tipos de estudios clínico epidemiológicos. *Tratado de Epidemiología Clínica. Madrid: 23-47*.

Gallo Gallón, J. D. et al. (2013). Diseño y construcción de un toolbox en ambiente scilab que apoye la enseñanza-aprendizaje de la regresión lineal y además ofrezca alternativas de solución al problema de multicolinealidad. Master's thesis, Pereira: Universidad Tecnológica de Pereira.

Garzón F, B. W. (2014). Protocolo de vigilancia en salud pública: Cáncer infantil. *MinSalud e Instituto Nacional de Salud*, 1(1):1–28.

Giraldo, R. (2002). Introducción a la geoestadística: Teoría y aplicación. *Bogotá: Universidad Nacional de Colombia*.

González, A. P. (2010). *Analisis predictivo de datos mediante tecnicas de regresion estadistica*. PhD thesis, Universidad Complutense de Madrid.

Granados, J. A. T. (1994). Incidencia: concepto, terminología y análisis dimensional. *Med Clin (Barc)*, 103:140–142.

Hanson, D. R. and Atlas, M. P. (2016). *Lanzkowsky's manual of pediatric hematology and oncology*, chapter Central Nervous System Malignancies, pages 475–494. Academic Press.

Hurtado González, K. L. and Ramírez Rodriguez, A. F. (2015). *IMPLEMENTACIÓ N DE GEÓPÓRTAL EPIDEMIÓLÓ GICÓ DEL CA NCER Sistema de Vigilancia de Cáncer Infantil*. PhD thesis, Universidad del Valle.

Imbach, P., Kühne, T., and Arceci, R. J. (2011a). *Pediatric oncology*, chapter Non-Hodgkin Lymphoma, pages 61–69. Springer.

Imbach, P., Kühne, T., and Arceci, R. J. (2011b). *Pediatric oncology*, chapter Acute Lymphoblastic Leukemia, pages 11–27. Springer.

Imbach, P., Kühne, T., and Arceci, R. J. (2011c). *Pediatric oncology*, chapter Brain Tumors, pages 95–117. Springer.

Instituto Nacional de Cancerologia (2016). `www.cancer.gov.co`, 21 de Febrero de 2016.

Loader, C. (1999). *Local regression and likelihood*. Springer, New York, USA.

Madsen, H. and Thyregod, P. (2010). *Introduction to general and generalized linear models*. CRC Press.

Marsaglia, G., Marsaglia, J., et al. (2004). Evaluating the anderson-darling distribution. *Journal of Statistical Software*, 9(2):1–5.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability*. Chapman & Hall,.

Montgomery, D. C. D. C., Peck, E. A., and Vining, G. G. (2006). *Introducción al análisis de regresión lineal*. Number Tercera Edición.

Muñoz, N., Knaul, F., and Lazcano, E. (2014). 50 años del Registro Poblacional de Cáncer de Cali, Colombia. *Salud Pública de México*, 56:421 – 422.

Nelder, J. A. and Baker, R. J. (1972). *Generalized linear models*. Wiley Online Library.

Organizacón Mundial de la Salud (2016). `www.who.int`, 21 de Febrero de 2016.

Organizacón Panamerica de la Salud (2016). `www.paho.org`, 21 de Febrero de 2016.

Ortega-García, J., López-Hernández, F., Sobrino-Najul, E., Febo, I., and Fuster-Soler, J. (2011). Medio ambiente y cáncer pediátrico en la región de Murcia (España): integrando la historia clínica medioambiental en un sistema de información geográfica. 74(4):255–260.

Pinilla, R., López, S., Quintana, J. C., and Al-Malahy, A. A.-E. (2009). Linfoma de burkitt de localización abdominal: dos casos operados en el hospital al-wahdah, maabar, yemen. *Revista Colombiana de Cirugía*, 24(2):106–113.

Rainey, J. J., Mwanda, W. O., Wairiumu, P., Moormann, A. M., Wilson, M. L., and Rochford, R. (2007). Spatial distribution of burkitt?s lymphoma in kenya and association with malaria risk. *Tropical Medicine & International Health*, 12(8):936–943.

Ramírez, G., Vasquez, M., Camardiel, A., Perez, B., Galindo, P., et al. (2005). Detección gráfica de la multicolinealidad mediante el h-plot de la inversa de la matriz de correlaciones. *Revista Colombiana de Estadística*, 28(2):207–219.

Registro Poblacional de Cáncer en Cali (2017). `rpcc.univalle.eud.co`, 2 de Agosto de 2017.

Salinas-Rodríguez, A., Pérez-Núñez, R., and Ávila-Burgos, L. (2006). Modelos de regresión para variables expresadas como una proporción continua. *Salud pública de México*, 48(5):395–404.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical methods for spatial data analysis*. CRC press.

Thompson, J. A., Zhu, L., and Carozza, S. E. (2008). Geographic risk modeling of childhood cancer relative to county-level crops, hazardous air pollutants and population density characteristics in texas. *Environmental Health*, 7(1):45.

Tovar, C., Rafael, J., and Gómez, G. A. (2016). Incidencia de cáncer infantil en una ciudad colombiana. *Revista Ciencias de la Salud*, 14(3):315–328.

Tusell, F. (2011). *Análisis de regresión. Introducción teórica y práctica basada en R*. Author New York.

Wartenberg, D., Groves, F. D., and Adelman, A. S. (2014). *Hematologic Malignancies: Acute Leukemias*, chapter Acute Luymphoblastic Leukemia: Epidemiology and Etiology, pages 77–93. Springer.

Wheeler, D. C. (2007). A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in ohio, 1996–2003. *International Journal of Health Geographics*, 6(1):13.

Wurttemberger, O. R. (2016). Information and childhood cancer. *Colombia Médica*, 47(2).