UNIVERSIDAD DE COSTA RICA

SISTEMA DE ESTUDIOS DE POSGRADO

ANÁLISIS ESPACIAL JERÁRQUICO BAYESIANO SOBRE LA ASOCIACIÓN ENTRE LA LETALIDAD DE LAS PERSONAS DIAGNOSTICADAS CON CÁNCER Y LA EXPOSICIÓN AMBIENTAL A PLAGUICIDAS EN COSTA RICA, 2011-2015

Tesis sometida a la consideración de la Comisión del Programa de Estudios de Posgrado en Estadística para optar al grado y título de Maestría Académica en Estadística

HAZEL PAOLA QUESADA LEITÓN

Ciudad Universitaria Rodrigo Facio, Costa Rica

2024

DEDICATORIA

A todos los momentos de procrastinación que, de alguna manera, me llevaron a encontrar nuevas ideas.

A todas las veces que tomé kinocola y a los días en que mi computadora no decidió colapsar.

A los memes que me hicieron reír cuando todo parecía imposible y a las siestas estratégicas que me mantuvieron cuerda.

Y, por supuesto, a mi familia, amigos y a mí misma por no rendirme.

¡Lo logramos!

AGRADECIMIENTO

Agradezco sinceramente a la PhD. Carolina Santamaria por brindarme acceso a los datos que se encuentran enmarcados en el proyecto de investigación del Instituto de Investigaciones en Salud, específicamente el proyecto 742-B7-371 "Asociación entre la exposición ambiental a plaguicidas y el cáncer en Costa Rica: 1980-2014".

De igual manera agradezco a:

Guaner Rojas Rojas, por su invaluable guía y apoyo durante el desarrollo de mi tesis. Su enseñanza en el curso de Estadística Bayesiana ha sido fundamental para mi investigación. Además, agradezco sus valiosos aportes y sugerencias en el análisis de datos, así como su dedicación en la revisión y corrección del documento.

Gilbert Brenes Camacho, por su colaboración como lector de tesis y por los importantes aportes que ha brindado con sus correcciones y comentarios, los cuales han contribuido a la mejora del documento.

Carolina Santamaria Ulloa, por ser mi lectora de tesis, por facilitar el acceso al conjunto de datos y por sus significativos comentarios relacionados con el documento, aportando además su experiencia en el campo de la salud.

"Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar por el grado y título de Maestría Académica en Estadística"

CRISTINA BARBOZA SOLIS (FIRMA)

Signature numérique de CRISTINA BARBOZA SOLIS (FIRMA)
Date : 2024.08.20 08:13:35 -06'00'

Dra. Cristina Barboza Solís
Representante de la Decana
Sistema de Estudios de Posgrado

GUANER DAVID ROJAS ROJAS (FIRMA)
PERSONA FISICA, CPF-05-0328-0120.
Fecha declarada: 26/08/2024 01:07:42 PM
Esta representación visual no es fuente de confianza. Valide siempre la firma.

Dr. Guaner Rojas Rojas
Director de Tesis

GILBERT BRENES CAMACHO (FIRMA)

Firmado digitalmente por GILBERT BRENES CAMACHO (FIRMA)
Fecha: 2024.09.04 16:06:52 -06'00'

Dr. Gilbert Brenes Camacho
Asesor

CAROLINA SANTAMARIA ULLOA (FIRMA)

Firmado digitalmente por CAROLINA SANTAMARIA ULLOA (FIRMA)
Fecha: 2024.09.12 11:01:02 -06'00'

Dra. Carolina Santamaria Ulloa
Asesora

ALEJANDRA MARIA ARIAS SALAZAR (FIRMA)
PERSONA FISICA, CPF-01-1469-0581.
Fecha declarada: 20/09/2024 03:00:01 PM

Dra. Alejandra Arias Salazar
Representante Programa de Posgrado

HAZEL PAOLA QUESADA LEITON (FIRMA)

Firmado digitalmente por HAZEL PAOLA QUESADA LEITON (FIRMA)
Fecha: 2024.07.29 08:02:17 -06'00'

Hazel Paola Quesada Leitón
Candidata

Tabla de contenido

RESUMEN

El objetivo principal de esta investigación fue identificar un método unificado para analizar datos relacionados con mortalidad por cáncer de mama y próstata en Costa Rica, considerando aspectos jerárquicos, espaciales y ecológicos. Esto se basó en un modelo lineal jerárquico binomial, cuya eficacia se evaluó a través de simulaciones. Inicialmente, se llevó a cabo un estudio empírico para comprender la relación entre las variables y construir un índice de exposición a plaguicidas.

Los resultados mostraron que los modelos cuya unidad geográfica corresponde a los conglomerados (unión de distritos dentro de provincias) mejoraron el rendimiento, en comparación con el agrupamiento original por distritos. Se generaron varios escenarios de simulación para identificar el modelo óptimo, variando tanto la unidad geográfica como el coeficiente de regresión asociado a la exposición a plaguicidas.

La simulación demostró que los modelos jerárquicos y espaciales superaron al modelo simple en términos de precisión y ajuste. Aunque en los diferentes escenarios de simulación evaluados el modelo jerárquico fue ligeramente superior según la log verosimilitud marginal, los modelos jerárquicos espaciales fueron preferidos debido a su capacidad para controlar el sesgo en la estimación del error debido a la correlación de datos.

La selección del mejor modelo se basó en medidas de bondad de ajuste como la log verosimilitud marginal, criterio de información de devianza y criterio de información de Akaike Watanabe, donde el agrupamiento de distritos dentro de provincias mediante análisis de conglomerados mostró el mejor ajuste. Los modelos espaciales con criterio de vecindad de reina fueron identificados como los más adecuados para el análisis de los datos.

Además, se subraya la importancia de controlar el uso de plaguicidas, dada su posible relación con el cáncer. Se insta a las instituciones encargadas del registro de plaguicidas a implementar un control más estricto para mejorar la calidad de los modelos de análisis y, en última instancia, contribuir a la comprensión y prevención del cáncer.

ABSTRACT

The primary objective of this research was to identify a unified method for analyzing data related to breast and prostate cancer mortality in Costa Rica, considering hierarchical, spatial, and ecological aspects. This was based on a hierarchical binomial linear model, whose effectiveness was evaluated through simulations. Initially, an empirical study was conducted to understand the relationship between variables and to construct a pesticide exposure index.

The results showed that models using conglomerates as the geographic unit (combination of districts within provinces) performed better compared to the original district-level grouping. Several simulation scenarios were generated to identify the optimal model, varying both the geographic unit and the regression coefficient associated with pesticide exposure.

The simulation demonstrated that hierarchical and spatial models outperformed the simple model in terms of accuracy and fit. Although the hierarchical model was slightly superior according to the marginal log-likelihood in the different simulation scenarios evaluated, the hierarchical spatial models were preferred due to their ability to control for bias in error estimation due to data correlation.

The best model selection was based on goodness-of-fit measures such as marginal log-likelihood, deviance information criterion, and Watanabe-Akaike information criterion, where clustering districts within provinces through cluster analysis showed the best fit. Spatial models using the queen contiguity criterion were identified as the most suitable for data analysis.

Additionally, the importance of controlling pesticide use is emphasized, given its possible relationship with cancer. Institutions responsible for pesticide registration are urged to implement stricter control measures to improve the quality of analysis models and ultimately contribute to the understanding and prevention of cancer.

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE ANEXOS

CAPÍTULO 1: Introducción

La relación existente entre la exposición ambiental a plaguicidas y relación con el cáncer ha sido estudiada en distintas regiones por instancias tales como el Centro Internacional de Investigaciones sobre el Cáncer (IARC, por sus siglas en inglés) y la Agencia de Protección Ambiental de Estados Unidos (EPA, por sus siglas en inglés). Estudios observacionales previos han sugerido que la exposición a plaguicidas puede estar relacionado con que las personas tengan una mayor propensión de ser diagnosticadas con cáncer (Wesseling et al., 1996, Santamaría, 2009).

Ramírez et al. (2017) indican que en Costa Rica se encuentran autorizados 21 plaguicidas conocidos por ser extremada o altamente peligrosos (PAP) según la clasificación de la Organización Mundial de la Salud, además comentan que aproximadamente un 80% de los ingredientes activos de los plaguicidas utilizados en el país son considerados PAP (pp. 7). En la tabla 1 se muestran los PAP más frecuentemente usados en el país para los cultivos de café y piña, destacando la presencia del paraquat que es una sustancia cuya toxicidad ha sido advertida en el Convenio de Rotterdam (Organización Panamericana de la Salud, 2022).

Tabla 1. Costa Rica: Plaguicidas altamente peligrosos más frecuentemente usados en cultivos de café y piña, 2016.

Ejemplo de producto PAP	Ingrediente activo PAP
Soprano 25 SC	carbendazina + epoxiconazol
Opera 18.3 SE	epoxiconazol + pyraclostrobina
Opus 12.5 SC	epoxiconazol
Validacin 5 SL, Cepex 10 SL	validamicina A
Duett 25 SC	carbendazina + epoxiconazol
Karmex 80 WP	diuron
Gramuron X 30 SC	diuron + paraquat
Sevin 80 SP	carbaryl
Mocap 15 GR	etoprofos
Vydate 24 SL	oxamil
Gramoxone Super 20 SL, Quemante 20 SL	paraquat
Preglone 20 SL	paraquat + diquat
Roundup 35.6 SL, Biokil 35.6 SL	glifosato

Fuente: Ramírez et al. 2016.

En Costa Rica se han realizado estudios previos para determinar la relación entre algunos tipos de cáncer y el uso de los plaguicidas (Wesseling et al., 1996, Santamaría, 2009). Estos estudios han tomado en cuenta la relación espacial de los datos mediante coeficientes de asociación y regresión con pesos geográficos; sin embargo, no ha considerado la estructura jerárquica presente en los datos ni la asociación espacial existente.

El estudio propuesto en esta investigación incorpora la información brindada por el Registro Nacional de Tumores (RNT) en el período de 1980 al 2014. La información corresponde a personas que han sido diagnosticadas con cualquier tipo de cáncer. Por otra parte, se requiere información del registro de mortalidad del Tribunal Supremo de Elecciones (TSE) mediante la cual se pueda identificar quienes han padecido de cáncer y han fallecido. La información brindad por el TSE abarca el período desde el 21 de octubre del año 1863 hasta el 31 de mayo del año 2019, sin embargo, se utilizó el período del 2010 al 2019 para determinar la fecha de muerte de las personas diagnosticadas con cáncer.

La implementación de la presente investigación necesita tener datos relacionados a con la exposición a plaguicidas para calcular el Índice de Exposición a Plaguicidas (IEP), tal índice será descrito en secciones posteriores. La información de distintos registros permite determinar las cifras más importantes para el IEP; tal como la cantidad utilizada de plaguicidas y el área sembrada para los distintos cultivos, así como la población que habita por cantón. También, la información de los Censos Agropecuarios de 1984 y 2014 que permiten obtener la cantidad de hectáreas sembradas por cultivo y tipo de producto utilizado.

El presente estudio muestra un corte de tipo ecológico puesto que se tiene información agregada geográficamente; lo cual genera que se deba analizar la estructura multinivel para poder unir esta información con cada uno de los individuos diagnosticados con cáncer. La dificultad asociada a la construcción del IEP consiste en que la información no es de fácil acceso. La menor desagregación que se puede obtener de la misma es a nivel de cantón o distrito. Se trabaja con la exposición ambiental, es decir, cualquier persona que resida en

los cantones o distritos del país; las personas no tienen que estar directamente expuestas a los plaguicidas.

Para realizar el análisis de datos se cuenta con un archivo de datos brindado por el Registro Nacional de Tumores (RNT), el cual incluye información respecto al sexo de las personas, la edad, el tipo de cáncer; información que se puede utilizar como variables independientes en un modelo jerárquico.

1.1 Antecedentes

En Costa Rica, desde el año 1941 se creó el sistema de seguridad social, lo cual ha generado que el país continúe destacando entre los países Latinoamericanos e inclusive entre algunos países desarrollados en cuestiones de salud. Entre los aspectos a resaltar se tiene el alto porcentaje de cobertura del seguro de salud con la que cuenta el país (90,3%) y diversas políticas públicas que han generado logros en el área de salud, (Rosero, 1991; OECD, 2017; ENAHO 2023).

Los avances en aspectos relacionados con salud, la transición demográfica y epidemiológica en las últimas décadas han generado que las causas de morbilidad y mortalidad sufrieran una transición de enfermedades infecciosas ,reducción de la tasa de muerte por enfermedades prevenibles por vacunación de 90 por cada cien mil en 1940 a 2 por cada cien mil en 1980, teniendo en cuenta que además en el 2012 un 83% de las muertes en Costa Rica fue producto de enfermedades no transmisibles, a enfermedades crónicas no transmisibles y degenerativas como enfermedades cardiovasculares y cáncer, (Rosero, B. 1991; OECD, 2017).

Cabe destacar que no es un patrón exclusivo de Costa Rica, ya que, a nivel mundial las muertes cardiovasculares ocupan el primer puesto, seguidas por las muertes ocasionadas por cáncer; las cuales para el año 2060 se espera que ocupen el primer puesto. (Mattiuzzi & Lippi 2019). En los países de bajos ingresos, 6 de las 10 principales causas de muerte son enfermedades transmisibles; en contraste, cuando se analiza la situación en países de ingresos medianos y altos, el panorama cambia significativamente: 8 de las 10 principales

causas de muerte en los países de ingresos medianos y 9 de las 10 en los países de ingresos altos están asociadas a enfermedades no transmisibles (OMS, 2020).

Previo a ahondar en la relación que puede existir entre el cáncer y los diversos compuestos químicos utilizados en la agricultura, es importante resaltar como se puede definir un plaguicida, Dich et al. (1997) los definen como:

"Cualquier sustancia o mezcla de sustancias (i) destinadas a prevenir, destruir, o controlar cualquier peste, incluyendo vector de enfermedades humanas o de animales, inesperadas especies de plantas o animales que causan daño durante la producción, procesamiento, almacenamiento, transporte o venta de comida, productos agrícolas, madera y productos de madera, o alimentos para animales; o (ii) administrado a animales para el control de insectos u otras plagas en o sobre su cuerpo" pp.420

Los estudios de Partanen et al. (2009) puntualizan las posibles causas de cáncer: contaminación de aire, agua y comida, factores de dieta, obesidad, inactividad física, tabaquismo, alcohol, radiación solar, factores hormonales, exposiciones tempranas en la vida, virus, herencia, drogas y ocupación.

Una posible causa de cáncer se relaciona con la ocupación o tipo de trabajo de las personas. Según Corella et al. (2000), hay algunos trabajos que aumentan las probabilidades de desarrollar cáncer de vejiga, de igual manera indica que se puede observar una relación entre asbesto y el cáncer de pulmón; una relación entre trabajadores expuestos al benceno con linfomas y leucemia; y que hay un leve aumento en el riesgo de cáncer cuando las personas trabajadoras están expuestas a algunos herbicidas. Adicionalmente, Wesseling et al. (1996) comentan que las personas que trabajan en agricultura han reportado altos riesgos de padecer cáncer. Además, exponen que en Costa Rica se presentaron casos de cáncer en trabajadores en plantaciones de banano, quienes estuvieron expuestos a sustancias químicas que han estado relacionadas con esterilización y son posibles carcinógenos humanos.

El uso de plaguicidas no puede considerarse como menor, ya que para el año 1985 se producían 3 millones de toneladas de plaguicidas en el mundo (Dich et al., 1996). Bravo et al (2015) indican que, para finales del siglo XX, la producción mundial se estimó en 2 800 millones de kilogramos, mientras que un dato más reciente indica que para el 2011 la producción era de aproximadamente de 2 700 millones de kilogramos.

Bravo et al. (2015) mencionan que en Centroamérica para el año 1994 se importaron 34 millones de kilogramos de plaguicidas, 45 millones de kilogramos a finales de la década de los noventa y casi 33 millones de kg por año en el periodo 2000- 2004. Finalmente, en el contexto costarricense, Bravo et al. (2015) indica que para el período del 2005 al 2009 en Costa Rica se importaron en promedio entre 11 496 y 12 291 toneladas.

En el estudio realizado por Bravo et al. (2015) se puede observar que, en general, Costa Rica es de los países Centroamericanos que tiene la mayor cantidad de kilogramos por habitante. Sin importar si las personas están directamente expuestas a plaguicidas o no, se importó 2.8 kg de plaguicidas por persona durante el período del 2005 al 2009, lo cual es un indicador que Costa Rica no es un país ajeno a las consecuencias que puedan causar las plaguicidas, ya que a nivel de Centro América era el de mayor consumo.

Teniendo en cuenta la información relacionada con la importación de plaguicidas, se reanudará la discusión sobre su vínculo con el cáncer. Una de las causas frecuentes del cáncer está relacionada con la ocupación. Según la Organización Mundial de la Salud (OMS, 2018), aquellos que trabajan en la agricultura son los más susceptibles a padecer cáncer, al igual que aquellos que están expuestos durante la aplicación de plaguicidas. Sin embargo, la OMS también señala que la población en general está expuesta, aunque en menor medida que estos grupos, debido a la presencia de plaguicidas en los alimentos o en el agua que consumen.

Howard (2014) señala que la aparición del cáncer no es inmediata después de la exposición a un agente causal; por el contrario, puede tomar varios años e incluso décadas para manifestarse clínicamente, lo que se conoce como el período de latencia. Howard, J. (2014)

citando a Nadler y Zurbenko (2013) indica que este período puede variar desde los 2.2 años (para la leucemia linfocítica crónica) hasta 57 años (para el cáncer de colon transverso).

Mostafalou y Abdollahi (2013) seleccionaron artículos que incluyeran las palabras plaguicidas y cáncer (varios tipos de cáncer) y algún otro tipo de enfermedades que se puedan relacionar a los plaguicidas durante el período de 1980 a 2013.

En el cuadro 1.1 se muestra la cantidad de estudios que se han realizado en relación con la exposición a plaguicidas con diversos tipos de cáncer. Se destaca que hay 23 estudios que relacionan la exposición con leucemia en niños, lo cual evidencia que los efectos se producen en personas de todas las edades. En total hay 213 estudios que asocian la exposición a plaguicidas con algún tipo de cáncer.

Cuadro 1.1. Número de estudios publicados que asocian enfermedades crónicas a la exposición a plaguicidas (1980-2013) *

Enfermedad	Número de estudios
Cáncer	
Leucemia niños	23
Leucemia adultos	16
Linfoma Hodgkins	7
Linfoma no Hodgkins	30
Meloma múltiple	7
Sarcoma tejido blanco	5
Cerebro niños	14
Cerebro adultos	13
Huesos	5
Próstata	19
Colo-rectal	11
Páncreas	9
Riñones	11
Pulmón	11
Estómago	3
Hígado	3
Testicular	3
Vejiga	3
Tiroides	2
Melanoma	5
Ojos	3
Labios	3
Boca	1
Laringe	1
Naso-sinusal	1
Ovarios	2
Útero	1
Cervical	1
Defectos nacimiento	20
Des. Reproductivos	17
Neuro-degenerativos	
Parkinson	45
Alzheimer	4
Esclerosis lateral	9
Cardiovasculares	4
Respiratorias	
Asma	13
Obst. pulmonar crónica	8
Diabetes	6

* Tabla adaptada de Mostafalou y Abdollahi (2013).
Fuente: March, G. (2014). Agricultura y plaguicidas: un análisis global. - 1a ed. - Rio Cuarto: FADA - Fundación Agropecuaria para el Desarrollo de Argentina.

En el cuadro 1.2 se muestran los plaguicidas que según menciona March (2014) están asociados con una alta incidencia de cáncer.

Cuadro 1.2. Plaguicidas asociados con elevada incidencia de cáncer en estudios epidemiológicos (1980-2013)*

Enfermedad	Plaguicidas
Leucemia	Clordano, heptacloro, clopirifos, diazinón, EPTC, fonofos
Linfoma no-Hodgkin	Lindano, clordano
Mieloma Múltiple	Permetrina
Cáncer de cerebro	Clorpirifos
Cáncer de próstata	Fonofos, bromuro de metilo, butilato, clordecone, DDT, lindano, simazina
Cáncer de colon	Aldicarb, dicamba, EPTC, imazethapyr, trifuralina
Cáncer de recto	Clordano, clorpirifos, pendimentalin
Cáncer de páncreas	EPTC, pendimentalin, DDT.
Cáncer de pulmón	Clorpirifos, diazinon, dicamba, dieldrin, metolacloro, pendimentalin
Cáncer de vejiga	Imazethapyr
Melanoma	Carvaril, toxafeno, parathion, maneb, amcozeb

* Tabla adaptada de Mostafalou y Abdollahi (2013).

Fuente: March, G. (2014). Agricultura y plaguicidas: un análisis global. - 1a ed. - Rio Cuarto : FADA - Fundación Agropecuaria para el Desarrollo de Argentina.

1.2 Justificación

Costa Rica es uno de los países en Centroamérica que más importa plaguicidas, lo cual resulta ser de especial interés para el estudio de la relación entre la mortalidad de personas diagnosticadas con cáncer y la exposición ambiental a plaguicidas, pues los estudios de EPA y la IARC han determinado que los plaguicidas pueden tener diferentes grados de carcinogenicidad, es decir, que pueden causar cáncer en los humanos (Ramírez et al, 2016).

El desarrollo de una enfermedad como en el cáncer no es de poca relevancia, ya que en la actualidad a nivel mundial esta enfermedad ocupa el segundo puesto de causas de muerte y se proyecta que para el 2060 será la causa que cobre más vidas. Es de gran importancia realizar un estudio a nivel país donde, mediante un análisis espacial jerárquico bayesiano, se pueda analizar a nivel distrital (cantonal o para otra unidad geográfica) la asociación entre la exposición ambiental a plaguicidas y la mortalidad de personas diagnosticadas con distintos tipos de cáncer.

Debido a que modelos más simples no consideran estas estructuras jerárquicas y espaciales, la metodología empleada permite capturar de manera más precisa las variaciones en los datos. Los modelos jerárquicos bayesianos permiten incorporar jerarquías de los datos, lo cual es esencial para captar la variabilidad geográfica y el componente espacial permite capturar las correlaciones espaciales que afectan la incidencia y mortalidad por cáncer. Estos modelos pueden ajustar de manera más efectiva los sesgos y errores en las estimaciones al considerar la estructura dependiente de los datos y la correlación entre áreas geográficas.

En Costa Rica se ha realizado un estudio en que se relaciona la exposición a plaguicidas con la incidencia del cáncer de seno (Santamaría, 2009); así como un estudio que analiza la inequidad en Costa Rica de la incidencia del cáncer de cérvix (Santamaría & Valverde, 2019), ambos estudios comparten el análisis espacial realizado mediante la regresión ponderada geográficamente, destacando que se utiliza información medida toda a un mismo nivel. Respecto al cáncer de seno se determinó que existía una asociación entre la exposición a plaguicidas y la incidencia de cáncer.

Las variables independientes utilizadas por Santamaría & Valverde (2019) para el análisis del cáncer de cérvix fueron el Índice de Desarrollo Social y el Índice de Densidad de Acceso a los Servicios de Salud y la Subutilización de la Prueba de Papanicolaou. Santamaría (2009) utilizó en el estudio de cáncer de seno las variables independientes del Índice de Exposición de Plaguicidas en 1984, Índice de Acceso a los Servicios de Salud en

el año 2000, la tasa global de fecundidad de cohorte, proporción de mujeres que tuvieron su primer embarazo a término a los 30 años de edad o más y el Índice de rezago social.

Muir et al. (2004) realizaron un estudió en Inglaterra para analizar la incidencia del cáncer de seno y su posible asociación con la exposición a plaguicidas. Para cuantificar la correlación espacial utilizaron el coeficiente de correlación espacial I de Moran, además realizaron una regresión lineal para analizar la relación espacial entre el cáncer y los plaguicidas. Obtuvieron que en la zona rural de Inglaterra existía una relación entre la incidencia de cáncer y la exposición a Aldicarb, Atrazine y Lindane.

El alcance de esta investigación radica en que los resultados obtenidos de la misma puedan utilizarse como evidencia para la creación de políticas en salud tendientes a disminuir la exposición ambiental a plaguicidas en la población costarricense. Este estudio permite abordar a nivel metodológico los problemas de naturaleza espacial con una estructura jerárquica, es decir, con diferentes niveles de agregación.

La información de exposición a plaguicidas de las personas se mide a nivel cantón debido a la falta de información de uso de ciertos productos en unidades más pequeñas. Debido a esta agrupación de las personas residentes en estas unidades geográficas es que no son independientes entre sí; por lo que es de esperar que personas que residen en un mismo lugar tengan la misma exposición ambiental a estas sustancias, lo cual viola uno de los principales supuestos como lo es la independencia de las observaciones que se tiene al utilizar una técnica tal como los modelos lineales generalizados. Destacando que en los estudios previamente mencionados no se tiene la violación de este supuesto ya que se tiene información medida solamente para un nivel.

En la presente investigación se cuenta con datos agrupados e individuales y es de esperar que distritos o cantones cercanos tengan comportamientos similares en cuanto a la incidencia de cáncer, así como en la exposición ambiental a plaguicidas. Por lo que para el desarrollo de la presente investigación se considera utilizar un modelo espacial jerárquico que permita tomar en cuenta información medida en diferentes niveles de agregación

(modelos jerárquicos), además que tome en cuenta que las unidades geográficas a analizar tampoco son del todo independientes entre sí.

La utilidad de esta investigación radica en que se brinda una metodología a seguir cuando se tiene información medida en diferentes niveles y cuando se tiene observaciones que no se pueden considerar independientes, las cuales además se encuentran correlacionadas espacialmente.

1.3 Diseño de la investigación

El objetivo de esta investigación es identificar un método unificado para el análisis de datos con información jerárquica, espacial y ecológica sobre cáncer a partir de un modelo lineal jerárquico binomial en Costa Rica. Para ello, se realizó el análisis transversal de los datos empíricos haciendo uso de las técnicas que serán descritas en capítulos posteriores. A partir de los resultados obtenidos se simularon datos para analizar criterios como la log verosimilitud marginal, el criterio de información de Akaike Watanabe y el criterio de información de devianza, error cuadrático medio, raíz del error cuadrático medio y error absoluto medio respecto a los modelos utilizados, y de esta manera poder determinar cuál método permite capturar de mejor manera los coeficientes con los que se simularon los datos.

La presente investigación se dividió en dos partes:

- El estudio de los datos empíricos, comparando diversos modelos según qué tan adecuados son para la estimación de los diversos parámetros considerando la autocorrelación espacial y la estructura jerárquica de los datos.
- Un estudio de simulación, el cual se va a diseñar de acuerdo con los resultados de del estudio empírico.

1.4 Planteamiento de los objetivos

A continuación, se presentan los objetivos asociados que pretenden contestar la pregunta referente a cuál es la mejor aproximación para el análisis de datos con información jerárquica, espacial y ecológica en el contexto de casos de cáncer de mama y próstata en Costa Rica.

Objetivo general.

Identificar un método unificado para el análisis de datos con información jerárquica, espacial y ecológica sobre cáncer de mama y próstata a partir de un modelo lineal jerárquico bayesiano binomial en Costa Rica.

Objetivos específicos.

- Comparar modelos de regresión para datos jerárquicos con datos medidos a un nivel más bajo (nivel individual).
- Generar potenciales conglomerados espaciales de unidades geográficas para el cáncer de mama y próstata para la comparación de modelos jerárquicos.
- Analizar las características de los modelos estadísticos para datos jerárquicos en los que uno de los niveles representa unidades geográficas que pueden presentar autocorrelación espacial en los errores.
- Establecer la relación espacial que existe entre las tasas de incidencia de cáncer y la exposición a plaguicidas.

CAPÍTULO 2: Marco teórico

En el presente capítulo se presentará un desarrollo teórico de las técnicas utilizadas, partiendo de lo más general a casos más específicos; con un enfoque en un modelo de regresión que permita analizar una variable respuesta que sigue una distribución binomial. La descripción de tales técnicas corresponde al objetivo de integrar diversos conceptos necesarios para la comprensión de la técnica estadística a utilizar para el análisis de los datos de la presenta investigación.

2.1 Modelo de regresión

Cuando se realiza una investigación cuantitativa se deben de analizar las características o variables recolectadas que busquen cumplir con los objetivos planteados de la investigación. En este caso las variables recolectadas $Y_1, Y_2, \dots, Y_n$ asociadas a las unidades de estudio se puede suponer que son independientes e idénticamente distribuidas.

El valor esperado de cada una de estas variables es $E(Y_i) = \mu_y$, asumiendo que ninguna variable se relaciona con otra (Mendenhall et al, 2010) no es válido en muchos problemas inferenciales. Para ilustrar la afirmación Mendenhall et al. (2010) plantean el siguiente ejemplo: la potencia media de un antibiótico depende del tiempo que éste haya estado almacenado. Es decir, el comportamiento de una determinada variable Y depende del comportamiento de una variable X.

Mendenhall et al. (2010) comentan que la variable Y es conocida como variable dependiente y su promedio es una función de una variable X o varias variables X conocidas como variables independientes.

Relacionado al funcionamiento de los modelos de regresión, Gujarati (2010) afirma lo siguiente:

"*El análisis de regresión trata del estudio de la dependencia de una variable (variable respuesta) respecto de una o más variables (variables explicativas) con el objetivo de estimar o predecir la media o valor promedio poblacional de la primera en términos de los valores conocidos o fijos (en muestras repetidas) de las segundas.*" (pp. 15)

Neter et al., (1990) indican respecto a un modelo de regresión que es la relación entre 2 o más variables "*existe una distribución de probabilidad de (Y) para cada nivel de (X) y que los promedios de estas distribuciones varían de una manera sistemática con (X).*" (pág. 6)

Mendenhall et al. (2010) comentan que esta relación entre variables puede ser determinística o probabilística. La diferencia radica en que la primera relación no toma en cuenta ningún error para hacer pronósticos de la variable respuesta (Y) con respecto a las variables independientes (X). En el caso de los modelos probabilísticos, sí se toma en cuenta este componente de error, por lo que la relación entre las variables antes mencionadas puede verse representada de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i; \; i = 1, \dots, n \quad (1)$$

La ecuación 1 es conocida como regresión lineal simple donde:

n: representa la totalidad de las observaciones i.

Y_i: representa la variable respuesta.

β_0: Es el coeficiente que representa la intersección de la ecuación en el eje de las coordenadas.

β_1: Es el coeficiente de regresión asociado a la variable X_1.

X_1: Representa la variable independiente.

ε_i: Es el termino de error presente en los modelos probabilísticos, donde en este caso se espera que tenga una distribución normal con promedio 0 y varianza constante σ^2.

Una generalización del caso de la ecuación 1 es el modelo de regresión lineal múltiple donde se tiene más de una variable predictora y de manera general se puede representar de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i \quad (2)$$

Donde los β_k representan los coeficientes de regresión de las k variables independientes.

2.1.1 Supuestos de la regresión lineal

Los modelos de las ecuaciones 1 y 2 se basan en los siguientes supuestos para poder realizar predicciones y pruebas de hipótesis:

Normalidad

Uno de los principales supuestos en los que se basa la regresión lineal tanto simple como múltiple es que, los residuales o bien la variable respuesta condicionada por las variables independientes sigan una distribución normal.

Relación lineal entre la variable respuesta y los predictores

Las ecuaciones 1 y 2 planteadas hacen referencia a una relación lineal entre las variables a utilizar. Holmes et al. (2014) indican que si esta relación no se puede mantener se tendrá un mal ajuste de la ecuación a los datos y el modelo va a estar mal especificado.

Homocedasticidad

Holmes et al. (2014) comentan que otro de los supuestos es que la varianza de los valores residuales es constante para todos los valores de las variables independientes. A la igualdad de varianzas se le conoce como homocedasticidad.

No multicolinealidad

La multicolinealidad hace referencia a la relación existente entre las variables independientes. El supuesto en sí mismo como mencionan Holmes et al. (2014) es que una de las variables independientes no se pueda expresar como combinación lineal de otras de las variables predictoras, ya que de suceder tanto los coeficientes como sus errores estándar asociados son inestables y por lo tanto las conclusiones que se generen a partir de tal modelo no son válidas.

Independencia de los errores

Holmes et al. (2014) indican que bajo este supuesto los residuos de dos individuos deben de ser independientes, específicamente indican lo siguiente "*este supuesto implica que factores no medidos que influencian a la variable Y no están relacionados de un individuo a otro, esta temática se aborda en los modelos multinivel*" (pp. 4)

Holmes et al. (2014) indican sobre los modelos multinivel que se debe de tener en cuenta que los datos de los individuos pueden ser recolectados de clusters o aglomeraciones. Teniendo en cuenta que forman parte de un mismo grupo se espera que los comportamientos de tales individuos sean más similares entre ellos que con respecto al comportamiento de individuos de otra agrupación por lo tanto no se puede esperar que los residuos sean independientes.

2.2 Modelos lineales generalizados

Los modelos de regresión lineal simple o múltiple asumen que la variable respuesta condicional a las variables independientes se distribuye normalmente.

En el caso de que la variable respuesta condicionada por las variables independientes no siga una distribución normal se puede trabajar con modelos lineales generalizados (GLM, por sus siglas en inglés), los cuales son una generalización de los modelos de las ecuaciones 1 y 2. Relacionado a esto Nelder y Wedderburn (1972) comentan que los modelos lineales generalizados se caracterizan por tener una variable respuesta que puede seguir una distribución como la exponencial, gamma, normal entre otras; además se cuenta con una función de enlace cuyo propósito es linealizar la relación entre la variable dependiente y las variables independientes.

Faraway (2016) comenta que los modelos lineales generalizados se especifican mediante dos componentes: una variable respuesta que pertenezca a la familia exponencial de distribuciones y una función de enlace que describe la manera en que el promedio de la variable respuesta y las variables independientes se relacionan.

De manera general los GLM toman la siguiente forma:

$$f(Y\ \theta, \Phi) = exp\ \left[\frac{Y\theta - b(\theta)}{a(\Phi)} + c(Y, \Phi)\right] \quad (3)$$

θ: parámetro canónico, representa la ubicación.

Φ: parámetro de dispersión, representa la escala.

A partir de la especificación que se realice de las funciones a, b y c se obtienen diversas distribuciones de probabilidad, donde algunas de las más usadas son:

- Normal o Gaussiana
- Poisson
- Binomial
- Gamma

Existen muchas más distribuciones que forman parte de la familia exponencial además de las antes mencionadas y por ende forman parte de los GLM; tales como la Gaussiana inversa, Tobit, QuasiPoisson, QuasiBinomial, entre otras.

Respecto a las funciones de enlace que linealizan la relación entre el promedio de la variable dependiente con las variables independientes, Faraway (2016, pp.155) indica que las siguientes son las más comunes a utilizar:

- Normal, en este caso la selección de la función de enlace corresponde a la identidad.
- Poisson, debido a que se necesita que el parámetro μ sea positivo la función de identidad no es una buena opción, por lo que en este caso se utiliza el logaritmo natural.
- Binomial, el parámetro que especifica tal distribución es π, el cuál corresponde a una probabilidad de éxito. En este caso se pueden utilizar diversas funciones de enlace, sin embargo, entre las más populares se encuentran la logit y probit.
- Gamma, en este caso se utiliza como función de enlace a la función inversa.

Para poder estimar los coeficientes de regresión asociados a cada una de las variables independientes, así como la intersección de la ecuación, se puede utilizar la estimación por máxima verosimilitud.

Teniendo presente los objetivos de investigación y el tipo de variable respuesta a utilizar, se va a ahondar en las funciones de enlace asociadas a la distribución binomial. Se tienen dos funciones de enlace que son muy conocidas y utilizadas. La función de enlace logit tiene la siguiente forma:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + B_1 X_1 \qquad i = 1, \dots, n \quad (4)$$

Donde:

n: representa la totalidad de las observaciones.

π_i: corresponde a la probabilidad de éxito.

β_0: corresponde al coeficiente de regresión asociado a la intersección.

β_1: corresponde al coeficiente de regresión asociado a la variable independiente X_1.

Dobson (2008) indica que esta función de enlace tiene una interpretación natural como el logaritmo de los "odds". Dunn y Smyth (2018) define los "odds" como una razón de probabilidades donde se compara la probabilidad de que ocurra el evento A contra la probabilidad que de no ocurra tal evento.

De manera general en el caso de los GLM los supuestos en los que se basa son *(*Dunn, Smyth, 2018, pp.312):

- *"Falta de valores extremos: todas las respuestas fueron generadas del mismo proceso, por lo tanto, el mismo modelo es apropiado para todas las observaciones.*
- *Función de enlace: La función de enlace g() utilizada es la correcta.*
- *Linealidad: Todas las variables explicativas importantes son incluidas y cada una de las variables independientes son incluidas en el predictor lineal en la escala correcta.*
- *Función de varianza: La función de varianza V(μ) correcta es utilizada.*
- *Parámetro de dispersión: El parámetro de dispersión ϕ es constante.*
- *Independencia: Las respuestas y_i son independientes las unas de las otras."*

2.3 Modelos lineales mixtos

Una vez analizados los modelos lineales generalizados se debe de tener en cuenta la existencia de los modelos lineales mixtos, en cuyo caso la variable respuesta condicionada a las variables independientes tiene una distribución normal.

Los modelos lineales mixtos (LMM por sus siglas en inglés) dan cabida a los modelos jerárquicos, a los experimentos con bloques aleatorios y de igual manera a los modelos longitudinales o estudios de medidas repetidas. Estos últimos caracterizados por tener varias mediciones de las unidades estadísticas a través del tiempo. (Dobson, 2007)

Los LMM asumen que la relación entre la variable dependiente y las independientes es lineal; en este tipo de modelo se tienen efectos fijos y efectos aleatorios. Los modelos jerárquicos forman parte de los modelos lineales mixtos, por lo que es importante aclarar que en estos se tienen dos partes importantes a la hora de hacer su especificación: los factores fijos y los aleatorios, y sus respectivos efectos.

West et al. (2006) indican que un factor fijo se define como una variable categórica o de clasificación, respecto a la cual el investigador incluye determinados niveles que son de interés para el estudio, de esta manera se pueden generar contrastes entre los diferentes niveles. Por otra parte, estos mismos autores definen los factores aleatorios como una variable de clasificación, con la distinción de que los niveles incluidos representan una muestra de todos los niveles que la variable o factor puede tomar, en este caso el interés no se basa en hacer comparaciones de niveles; aclaran además que al tener un modelo multinivel con 2 o 3 niveles, las variables de clasificación incluidas en ambos casos se toman como factores aleatorios.

Este tipo de factores tienen asociados sus respectivos efectos. En el caso de los efectos fijos West et al. (2006) indican que los parámetros de efectos fijos, permiten describir la relación entre la variable respuesta y las variables independientes, y se pueden usar para hacer contrastes entre distintos niveles de la variable. Respecto a los efectos aleatorios, señalan que son valores aleatorios asociados a factores aleatorios, los cuales representan desviaciones aleatorias de las relaciones descritas por los efectos fijos.

Teniendo en cuenta los dos tipos de efectos presentes en modelo lineal mixto, se va a proceder ahora a analizar la estructura básica de un modelo jerárquico, el cual es definido como modelo nulo de dos niveles, es decir, sin covariables.

Acevedo, R. (2008) plantea el modelo jerárquico nulo mediante las siguientes ecuaciones:

$$Y_{ij} = \beta_{0j} + e_{ij} \quad (5)$$

$$\beta_{0j} = \beta_{00} + \mu_{0j} \quad (6)$$

La primera ecuación (5) representa el primer nivel, donde Y_{ij} representa la variable respuesta del modelo, medida para un determinado sujeto i dentro del grupo j (segundo nivel). β_{0j} corresponde al intercepto, es el promedio de la variable respuesta para el grupo j del segundo nivel. Finalmente, de esta primera ecuación, e_{ij} constituye el residuo o varianza residual, cuya media es cero y tiene una varianza σ_e^2; el mismo representa las diferencias en la variable respuesta de los individuos, entre el valor estimado por la

regresión y el valor observado. Acevedo (2008) de igual manera menciona que en su modelo más simple (de los modelos jerárquicos) se asume la varianza de error al azar, es la misma para todos los sujetos.

La ecuación 6 del segundo nivel, donde β_{00} es la media general de todos los grupos j y μ_{0j} es el residuo o varianza residual del nivel dos, la desviación del valor estimado para el grupo j de su valor real. Tiene media cero y varianza $\sigma^2_{\mu_0}$. (Acevedo, 2008)

Las ecuaciones 3 y 4 se pueden reescribir como se muestra en la siguiente ecuación:

$$Y_{ij} = \beta_{00} + \mu_{0j} + e_{ij} \quad (7)$$

En tal ecuación (7) se tiene de manera resumida la información de ambos niveles, donde cada una de las partes que lo componen fue descrita previamente. Es importante destacar que en el modelo de la expresión 7 y en el que se va a presentar a continuación, se asumen que la variable respuesta tiene una distribución normal.

De La Cruz (2008) y Acevedo (2008) mencionan que se puede tener un modelo de dos niveles que incluya más información si se toman en cuenta algunas variables predictoras. En el caso de que se tenga solamente una variable independiente el modelo de dos niveles sería escrito de la siguiente manera:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \mu_j + e_{ij} \quad (8)$$

En este caso el coeficiente β_1 está asociado a la variable independiente medida para la unidad i en el grupo j. El número de variables independientes que se puede incluir en el modelo no está definido ya que esto depende del objetivo de estudio; sin embargo, es importante recalcar como se mencionó antes, que estas pueden estar medidas a nivel individual o a nivel grupal, es decir, en los distintos niveles contemplados.

De La Cruz (2008) comenta que los coeficientes β_0 y β_1 son usualmente los parámetros fijos y el conjunto de varianzas y covarianzas son los parámetros aleatorios. Se debe resaltar que se pueden tener modelos de más de dos niveles, pero se mantiene la estructura anidada. Ejemplo: en un primer nivel se tienen las personas costarricenses, en el siguiente nivel (nivel 2) una estructura de distritos (los individuos agrupados según distrito de residencia), en el nivel 3 se tienen cantones los cuales agrupan a los distritos, final e hipotéticamente podría tenerse un cuarto nivel que sea las provincias donde se encuentran anidados los cantones. Cada nuevo nivel agrega complejidad al modelo y en algunas ocasiones una estructura más compleja no significa una ganancia en términos de varianza explicada.

Morgenstern (1995) desarrolla que no solamente existen estudios ecológicos y estudios individuales, si no que existe información agrupada e individual que se suele combinar. De igual manera, según Haneuse y Bartell (2011), cuando ambos niveles se combinan este tipo de diseño se denomina multinivel. Es importante subrayar que estos modelos también se suelen conocer como modelos jerárquicos.

De La Cruz (2008) define el coeficiente de correlación intraclase como un estimador de la proporción de varianza explicada en la población. La correlación intraclase es igual a la proporción estimada de la varianza del nivel grupo comparada con la varianza total estimada. El coeficiente de correlación intraclase se calcula mediante la siguiente ecuación:

$$\rho = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_e^2} \quad (9)$$

Garrido y Murillo (2014) comentan que este coeficiente servirá para evaluar el grado de similitud entre unidades de nivel individual (o menor nivel) que pertenecen al mismo grupo. Si la correlación es baja indica que la variabilidad total está altamente determinada por la diferencia que existe entre individuos, es decir, que la estructura jerárquica no está ayudando a analizar la variabilidad; caso contrario es si toma un valor alto, ya que indicaría que un alto porcentaje de la variabilidad total está siendo explicada por la estructura jerárquica que existe en los datos.

2.4 Modelos lineales generalizados mixtos

En los modelos lineales generalizados mixtos (GLMM, por sus siglas en inglés) la variable respuesta sigue una distribución distinta de la normal siempre y cuando la misma pertenezca a la familia exponencial. Por otra parte, se debe de tomar en cuenta que los valores de la variable respuesta están correlacionados y además se tienen los efectos (fijos y aleatorios). (West et al. (2006); Dobson (2008); Jiang & Nguyen (2021))

Tanto en los LMM y los GLMM se da cabida a los modelos multinivel o jerárquicos, donde se tiene información de un conjunto de individuos, además de una agrupación de estos (el segundo nivel que anida al primer nivel). Dobson (2008) comenta que las variables independientes pueden ser mediciones del nivel 1 (individuales) o del nivel 2 (grupos), sin embargo, la variable respuesta siempre corresponde a mediciones del nivel 1.

Al utilizar LMM y GLMM se solventa las significancias espurias de los coeficientes de regresión estimados. De igual manera al utilizar una estructura jerárquica se evita la falacia ecológica, interpretación de datos agrupados como individuales; así como la falacia atomística, interpretación agregada de datos individuales (Aparicio y Morera, 2007).

Monsalve (2013) comenta que debido a que los individuos que se estudian no son extraídos de una población homogénea, especialmente cuando los individuos comparten una determinada cercanía, los modelos jerárquicos son capaces de tratar esta heterogeneidad debido a los niveles con los que trabaja.

La variable respuesta de los modelos lineales generalizados mixtos condicionada a las variables independientes puede seguir cualquier distribución siempre y cuando pertenezca a la familia exponencial. Por lo tanto, de igual manera que con los modelos lineales generalizados se puede trabajar con la distribución Poisson, Binomial y Gamma, entre otras.

2.5 Análisis de datos espaciales

de Corso y Pinilla (2017) indican que realizar un análisis exploratorio de los datos mediante un gráfico, en este caso un mapa, es ideal, ya que de esta manera se puede observar la distribución espacial de las diferentes variables y su comportamiento. Por ejemplo, el análisis exploratorio espacial permite observar si se presentan agrupamientos o si el comportamiento de las variables parece ser aleatorio; de igual manera a partir de este análisis previo se pueden formular hipótesis.

El problema de investigación de este documento consiste en determinar si existe en Costa Rica una asociación entre la incidencia de cáncer y la exposición espacial a plaguicidas, para lo cual se tiene información individual de las personas (falleció o no de cáncer) e información a un nivel más alto (exposición de plaguicidas de manera agrupada). Para poder unir la información que se encuentra a dos niveles distintos, no se puede utilizar los modelos lineales de regresión ya que estos asumen que los datos son independientes, por lo que se debe de explorar otra manera de poder analizarlos.

Sánchez (2012) comenta que hay dos grandes efectos que estructuran las relaciones en el espacio. La primera de ellas es la heterogeneidad espacial que se presenta en las áreas geográficas de estudio; donde heterogeneidad espacial hace referencia a las diferencias en las distribuciones (media, varianza) en un subgrupo de los datos. Sánchez (2012) menciona que el segundo factor al que hace referencia es la autocorrelación espacial que tienen las unidades geográficamente cercanas en correspondencia con la Ley de Tobler (1970), la cual indica que *"Todas las cosas se parecen a otros objetos, pero se parecen más a los objetos más cercanos"* (Tobler 1970, pp. 236).

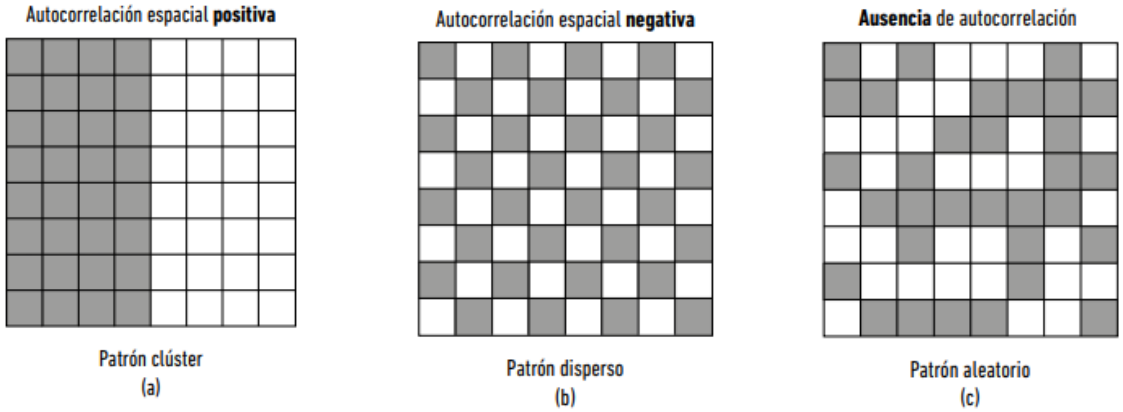
2.5.1 Coeficientes de correlación espacial

Para tener una cuantificación respecto a la correlación espacial se procederá a calcular el coeficiente de correlación I de Moran.

El coeficiente I de Moran es un coeficiente que varía entre -1 y 1; donde 0 implica la no existencia de un patrón definido, mientras valores cercanos a -1 indicarían autocorrelación

negativa, y 1, el máximo de autocorrelación positiva. Tal coeficiente puede considerarse como un coeficiente de correlación de Pearson ponderado por una matriz de pesos geográficos. La diferencia entre los dos coeficientes radica en que la correlación de Pearson analiza la relación entre dos variables, mientras que en el caso de los coeficientes de correlación espacial se analiza únicamente una variable y se analiza qué tanta relación tiene una unidad geográfica con las respectivas unidades cercanas. De manera gráfica, Siabato y Guzmán (2019) muestran cómo se observan tales resultados en la siguiente figura:

Figura 2.1. Patrones espaciales y su relación con la autocorrelación espacial.



Fuente: Siabato, W., y Guzmán, J. (2019). “La autocorrelación espacial y el desarrollo de la geografía cuantitativa. pp.5 ”

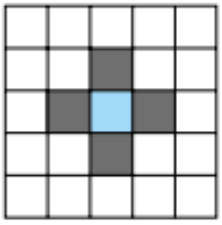
Sánchez, L. (2012) plantea la ecuación para el cálculo del coeficiente de correlación I de Moran:

$$I = \frac{\sum_{k=1}^{n}\sum_{l=1}^{n} W_{kl}(Y_k-\bar{Y})(Y_l-\bar{Y})}{\sum_{k=1}(Y_k-\bar{Y})^2} \quad (10)$$

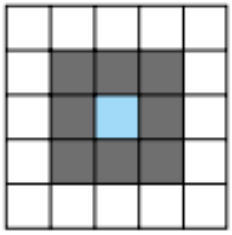
Donde Y_k representa el valor de la variable a estudiar en la unidad geográfica de análisis k y Y_l, la observación para la unidad l. La $\bar{Y}$ representa el promedio de tal variable y W_{kl} representa la matriz de pesos geográficos, la cual determina que observaciones son consideradas vecinas entre sí.

La matriz W_{kl} tiene pesos espaciales que son función de alguna medida de contigüidad en la matriz de datos originales. Siabato y Guzmán (2019) comentan algunos de estos criterios de vecindad, de los cuales solo se tomarán en cuenta Torre y Reina, ya que son los que consideran a los vecinos con los que comparten algún límite. Como se puede notar en la figura 4 la diferencia entre Torre y Reina consiste en que la Torre considera únicamente los vecinos con los que comparte una frontera, mientras que Reina considera a estos mismos vecinos y de igual manera los que se encuentran diagonalmente cercanos.

Figura 2.2. Criterios de vecindad.

Torre: 

4 vecinos

Reina: 

8 vecinos

Fuente: Siabato, W., y Guzmán, J. (2019). “La autocorrelación espacial y el desarrollo de la geografía cuantitativa. pp.8”

Finalmente, la matriz que contiene los pesos a utilizar para calcular tal coeficiente de correlación es una matriz cuadrada con una diagonal de ceros, ya que un elemento *i* no puede ser considerado vecino de sí mismo. Fuera de la diagonal se tendrán valores de 1 cuando se defina mediante alguno de los criterios de vecindad (torre o reina) que son elementos vecinos y 0 cuando este no sea el caso.

Con el coeficiente de correlación *I* de Moran se puede poner a prueba la hipótesis para determinar si este coeficiente es significativamente distinto de 0. Las hipótesis por plantear son las siguientes:

H_0: El atributo analizado está distribuido en forma aleatoria entre las unidades del área de estudio, no existe una correlación espacial.

H_1: El atributo analizado no está distribuido en forma aleatoria entre las unidades del área de estudio, existe una correlación espacial.

El coeficiente de Moran junto con el C de Geary y el G de Getis y Ord son los coeficientes globales más utilizados; sin embargo, debe de considerarse que estos indicadores no son los más adecuados si existe heterogeneidad espacial; es allí donde los coeficientes de correlación locales toman relevancia ya que trabajan a partir de subzonas (Siabato & Guzmán, 2019).

2.5.2 Modelos espaciales

Teniendo en cuenta lo expuesto la sección 2.5 referente a los modelos lineales mixtos generalizados, se procede ahora a presentar algunos modelos espaciales. Inicialmente se considera lo planteado por Xu (2014) quien indica un primer modelo estándar multinivel asumiendo que la variable respuesta es Bernoulli. Para un sujeto i en un vecindario j, usando la función logit como enlace para linealizar la relación entre la respuesta y las variables independientes es el siguiente:

$$logit\left[E\left(Y_{ij}\right)\right] = logit\left(p_{ij}\right) = \alpha_0 + X_{ij}\beta + Z_j\gamma + \mu_j \quad (16)$$

Donde:

Y_{ij}: representa la observación binaria para el individuo i en el grupo j.

p_{ij}: probabilidad de éxito, dado que la variable se asume sigue una distribución Binomial.

α_0: es la intersección.

$X_{ij}\beta$: son los coeficientes de regresión para las variables a nivel individual.

$Z_j\gamma$: son los coeficientes de regresión para las variables a nivel de vecinos (segundo nivel).

μ_j: es el parámetro que captura la correlación dentro de los vecindarios, donde se asume se distribuye normalmente con promedio 0 y varianza σ_u^2.

El modelo presentado en la ecuación 16 es útil para poder adquirir conocimiento respecto al modelado de datos espaciales. Xu (2014) indica un modelo puramente espacial, el cual se presenta a continuación:

$$logit\left[E\left(Y_{ij}\right)\right] = logit\left(p_{ij}\right) = \alpha_0 + X_{ij}\beta + Z_j\gamma + s_j \quad (17)$$

En este caso s_j es el parámetro que considera la correlación entre vecindarios.

Una manera de especificar s_j es mediante los modelos condicionalmente auto regresivos (CAR). Esta es una de las maneras de capturar la correlación espacial; sin embargo, como lo indican Hoef et al. (2018) no es la única manera. Entre otros se deben de tomar en cuenta los modelos autorregresivos simultáneos (SAR, por sus siglas en inglés) y los modelos intrínsecos autorregresivos (IAR) (caso especial de un modelo CAR), siendo los CAR y los SAR los de uso más común.

Hoef et al. (2018) indican 6 principales objetivos por los cuales el uso de modelos autorregresivos es óptimo:

"1. Selección de modelos, utilizando las matrices de varianzas y covarianzas de todos los modelos a utilizar ya sean SAR o CAR y a partir de las estimaciones de los modelos utilizar criterios tales como los de Akaike o el criterio de deviancia para realizar la selección del modelo que mejor se adecua a los datos.

2. Regresión espacial, para poder analizar y entender la relación entre la variable respuesta y predictoras.

3. Estimación de autocorrelación, entender la fuerza de la autocorrelación puede revelar la conectividad e interrelación de sistemas ecológicos.

4. Estimación de otros parámetros de conectividad, comprensión de los efectos de las covariables directos sobre la autocorrelación.

5. Predicción espacial, realizar predicciones sobre ubicaciones no muestreadas.

6. Suavizamiento espacial, uso de valores de lugares cercanos para proporcionar mejores estimaciones del fenómeno de interés." (pp. 37-38)

Los modelos geoespeciales presentan un cercamiento para analizar datos espaciales, estos son especialmente útiles cuando la información que se tiene está georeferenciada; por ejemplo, por medio de coordenadas GPS y por lo tanto la autocorrelación espacial se mide por la distancia entre los puntos. Sin embargo, no en todos los casos se tiene este tipo de

información. Por lo contrario, se tiene información de áreas tales como provincias, cantones y distritos entre otros, y es en estas ocasiones que los modelos espaciales autorregresivos toman ventaja (Hoef et al., 2018).

Suponga que se tiene un modelo especificado de la siguiente manera:

$$Y = \beta X + Z + \varepsilon \quad (18)$$

Donde, Y corresponde a la variable respuesta, β corresponde a los coeficientes de regresión asociados a las X variables independientes, Z corresponde al error espacial aleatorio con distribución $N(0, \Sigma)$ y ε corresponde al error aleatorio con promedio 0 y varianza $\sigma^2 I$, donde I es la matriz identidad.

Z: que corresponde al error aleatorio espacial puede tener una varianza (Σ) que corresponde a SAR o CAR. Hoef et al. (2018) indica la forma que tiene la varianza en ambos casos de la siguiente manera:

Modelo SAR:

$$\Sigma = \sigma_z^2 [(I - B)\left(I - B^{`}\right)]^{-1} \quad (19)$$

Modelo CAR:

$$\Sigma = \sigma_z^2 (\mathrm{I} - \mathrm{C})^{-1} \mathrm{M} \quad (20)$$

I corresponde a la matriz identidad, donde en la diagonal tiene unos y fuera de la diagonal 0, por lo tanto I=M. B=$\{b_{ij}\}$, C=$\{c_{ij}\}$, donde el valor que toman es 1 si los elementos son vecinos, 0 si no son vecinos, y en los casos específicos de c_{ii} y b_{ii} son 0 ya que no se puede ser vecino de sí mismo. M= $\{m_{ij}\}$, donde todos los valores fuera de la diagonal son 0 y m_{ii} es proporcional a la varianza condicional de Z_i dada por todos los vecinos. B=pW y C=pW, p controla la fuerza de la dependencia y W corresponde a una matriz de pesos.

En cuanto a los modelos especiales autorregresivos, es importante destacar que los CAR y los SAR comparten similitudes y a su vez diferencias. Hoef et al. (2018) indican que los modelos CAR permiten especificar las correlaciones parciales o covarianzas en lugar de la autocorrelación directamente, mientras que los modelos SAR en vez de utilizar las

correlaciones parciales utiliza la raíz cuadrada de la matriz de precisión. En ambos casos el valor de p no puede ser seleccionado de manera tal que (I-C) y (I-B) sean inversas y tengan un autovalor positivo para tener de esta manera una distribución espacial apropiada.

s_j se puede especificar por medio de los efectos aleatorios que se asume tienen una distribución normal con promedio 0 y varianza $\sigma_s^2 H(\varphi)$. Xu (2014) indica que σ_s^2 denota la varianza de los efectos aleatorios, conocida como el umbral parcial en estadística espacial, mientras que $H(\varphi)$ representa una matriz de correlaciones que indica como a medida que aumenta la distancia entre dos puntos de análisis (vecindarios) así va a disminuir la correlación entre ellos.

Xu, H. (2014) comenta que $H(\varphi)$ se representa mediante la siguiente función:

$$H(\varphi)_{ij} = p(d_{ij}, \varphi) \quad (21)$$

En este caso p() representa a una función la cual asume que la relación que existe entre dos localidades depende únicamente de la distancia que hay entre ellas, la cual se mide mediante d_{ij}. φ controla la tasa de disminución de la correlación espacial a medida que aumenta la distancia entre dos ubicaciones. Xu (2014) citando a Banerjee et al. (2004) expone que la distancia a la cual dos unidades se encuentran y que genera que la correlación caiga por debajo del 5% se considera como una correlación no significativa. Finalmente, Xu (2014) apunta que la función de correlación p puede ser una función exponencial o normal, sin embargo, la función normal presentó problemas de convergencia.

2.5.3 Métodos de estimación bayesiana.

En el presente apartado se expondrán métodos computacionales de estimación bayesiana teniendo en cuenta que hasta hace algunas décadas el uso de este tipo de análisis se ha podido implementar. (Press, 2003)

Métodos computacionales bayesianos: Cadenas de Markov vía Montecarlo (MCMC).

Una vez se haya definido la función de probabilidad de la distribución previa, se debe de recurrir a la estimación para lo cual se hará referencia a los métodos de estimación

bayesiana. Tales métodos involucran la creación de una cadena de Markov, la cual Press (2003) define como "*cadenas que tienen la propiedad de que la densidad condicional de* θ^j *condicionado a toda la historia precedente de la cadena depende únicamente del valor previo* θ^{j-1}*, lo cual es conocido como la densidad de transición.*" (pp. 120).

La densidad de transición debe de converger a la distribución posterior desde cualquier punto que se elija, la selección del punto puede ser desde el momento inicial. En este proceso normalmente se deja una cantidad de iteraciones iniciales que serán descartadas esto para facilitar el proceso de convergencia (Press, 2003).

Para la construcción correcta las cadenas de Markov se pueden utilizar el algoritmo de Metropolis Hastings, introducido por autores con estos mismos apellidos o el algoritmo de Gibbs, el cual es un caso del muestreo de Metropolis Hastings.

Una vez construida la cadena de Markov haciendo uso de alguno de los algoritmos antes mencionados se debe de verificar la convergencia, es decir, que el comportamiento de la cadena está cerca de ser estacionario por lo que se deben de evaluar ciertos diagnósticos de convergencia.

Algunos de los diagnósticos que pueden utilizarse para determinar que el proceso de estimación ha alcanzado la convergencia al utilizar la metodología de MCMC son:

- Gráfico de traza

Lunn et al. (2012) indican que este gráfico ilustra una línea continua que representa las diferentes realizaciones contra el número de iteraciones del muestreo de Gibs o Metropolis Hastings. En este tipo de gráfica se espera que después de un determinado número de iteraciones (quemado) se observe un comportamiento estable (estacionariedad) para determinar que se ha alcanzado convergencia.

- Gelman y Rubin

Lunn et al. (2012) indican que en este caso se comienza con cadenas que tienen valores iniciales simulados que tienen sobredispersión; en este caso la convergencia se evalúa mediante la comparación de la variabilidad dentro y entre cadenas. Se determina que se

alcanzó convergencia si $\hat{R} < 1.05$. De igual manera que el caso de la gráfica de traza, este criterio de convergencia se puede analizar gráficamente buscando que después de un número de iteraciones, $\hat{R}$ tienda al valor de 1.

Métodos computacionales bayesianos: Aproximación Anidada Integrada de Laplace (INLA).

INLA (Aproximación Anidada Integrada de Laplace, INLA por sus siglas en inglés) ofrece otra forma de estimación bayesiana de los coeficientes de regresión (Martino & Riebler, 2020). Es un nuevo enfoque de la inferencia estadística para los campos aleatorios de cadenas latentes Gaussianas de Markov. Estos mismos autores comentan que la ventaja que presenta esta estimación respecto a MCMC es una computación más rápida.

Martino y Riebler (2020) indican que Modelos Guassianos Latentes (LGM) son el tipo de modelos que se pueden estimar utilizando INLA. Rue et al. (2016) citando a Rue et al. (2009) indican que los LGM representan una abstracción muy útil que incluye una gran clase de modelos estadísticos, ya que, la inferencia estadística se puede unificar para la clase entera. Para realizar las respectivas inferencias se debe de especificar el modelo usando una especificación jerárquica de 3 etapas del modelo. Para ello Rue et al. (pp 3, 2016) comentan que las observaciones "X" se asumen condicionalmente independientes dados un campo aleatorio Guassiano "X" y los hiperparámetros:

$$Y \mid X, \theta_1 \sim \prod_{i \in I} \pi(Y_i \mid X_i, \theta_1) \quad (22)$$

Donde Y representa las observaciones, las cuales pueden ser consideradas como condicionalmente independientes, dado un campo latente Gaussiano X y los hiperparámetros θ_1.

La versatilidad de la clase del modelo recae en la especificación del campo latente gaussiano

$$X \mid \theta_2 \sim N(\mu(\theta_2), Q^{-1}(\theta_2)) \quad (23)$$

Donde se incluyen todos los términos aleatorios en el modelo estadístico describiendo de esta manera la estructura de los datos. Los hiperparámetros controlan el efecto latente

Guassiano y la verosimilitud de los datos, teniendo finalmente que la distribución posterior tiene la siguiente forma:

$$\pi(x, \theta \mid Y) \propto \pi(X \mid \theta) \prod_{i \in I} \pi(Y_i \mid X_i, \theta) \quad (24)$$

Rue et al. (2016) indican que los modelos lineales generalizados de manera aditiva pueden representarse mediante la siguiente ecuación:

$$\eta_i = \mu + \sum_j \beta_j z_{ij} + \sum_k f_k, j_k(i) \quad (25)$$

Donde:

a) μ representa el promedio general
b) Z representa las covariables fijas
c) β efectos lineales de las covariables

La diferencia con los modelos lineales generalizados está en agregar los términos f_k los cuales se usan para representar los procesos Gaussianos específicos; entre los cuales se encuentran k modelos de series de tiempo autoregresivos, modelos de suavizamiento, modelos para medir los errores de medición, modelos de efectos aleatorios con diferentes tipos de correlación, modelos espaciales entre otros.

Para comprender las razones por las cuales la estimación mediante INLA es más eficiente respecto al método MCMC, es necesario definir lo que es un campo aleatorio Gaussiano de Markov (GMRF por sus siglas en inglés). Rue et al. (2016) indican que X es una variable GMRF Gaussiana con propiedades condicionales adicionales, lo que significa que X_i y X_j son condicionalmente independentes dados los restantes elementos X_{ij} para unas cuantas {i, j}.

Rue et al. (pp. 5, 2016) indican que se necesita obtener varias veces la forma de la distribución conjunta del campo latente y esto dependerá de los hiperparámetros θ; por lo tanto, para evitar el costo computacional de las operaciones matriciales, la distribución conjunta puede considerarse como GMRF, cuya matriz es sencilla de calcular y esta es una

de las razones claves por las cuales la estimación utilizando INLA es tan eficiente, así como la estructura dispersa de la matriz de precisión.

Martino, Rue y Chopin (2009) indican que la estimación mediante INLA se puede utilizar en otros casos cuando la variable respuesta siga una distribución distinta a la normal, siempre y cuando la variable respuesta tenga una distribución que pertenezca a la familia exponencial y que se pueda ligar a una serie de variables independientes a través de una función de enlace.

2.5.4 Modelos jerárquicos bayesianos

En este caso al igual que en el caso frecuentista, se cuenta con diferentes niveles, donde la variable respuesta está medida en el nivel inferior; mientras que las variables independientes se pueden medir a ese mismo nivel, pero al menos hay una de las variables que genera un conglomerado de las unidades estadísticas. A diferencia del caso frecuentista, la estimación de los coeficientes de los modelos de regresión se realiza a partir de la Aproximación Anidada Integrada de Laplace y definiendo las distribuciones previas para los mismos.

2.5.5 Modelos jerárquicos espaciales bayesianos

Lunn et al. (2012) estudian la manera bayesiana de trabajar con datos jerárquicos, donde indican que la información previa tiene una estructura multinivel. Estos mismos autores indican que los parámetros poblacionales son efectos aleatorios ya que son asignados por una distribución de probabilidad previa. Es importante destacar que, tanto desde una perspectiva frecuentista como bayesiana, la variable respuesta puede seguir cualquier otro tipo de distribución distinta a la normal.

Seguidamente si se toma en cuenta que los datos tienen una estructura espacial, Jin et al. (2005) comentan que cuando se tiene información para regiones geográficas, tales como condados, tractos censales, entre otros, los modelos que de manera más frecuente se utilizan son condicionales autorregresivos. De igual manera estos mismos autores indican que las distribuciones CAR son utilizadas como distribuciones de los efectos aleatorios en la estructura del promedio de los modelos jerárquicos.

Xu (2014) expone que los modelos estándar multinivel y los espaciales puros pueden combinarse en un modelo hibrido cuya ecuación se presenta a continuación:

$$logit\left[E\left(y_{ij}\right)\right] = logit\left(p_{ij}\right) = \alpha_0 + X_{ij}\beta + Z_j\gamma + \mu_j + s_j \quad (26)$$

Teniendo que

α_0: es la intersección.

$X_{ij}\beta$: son los coeficientes de regresión para las variables a nivel individual.

$Z_j\gamma$: son los coeficientes de regresión para las variables a nivel de vecinos (segundo nivel).

μ_j: es el parámetro que captura la correlación dentro de los vecindarios, donde se asume se distribuye normalmente con promedio 0 y varianza σ_u^2.

s_j: es el parámetro que considera la correlación espacial entre vecindarios.

Lunn et al. (2012) señalan que una manera de capturar la correlación espacial entre vecindarios es mediante s_j, haciendo uso de modelos condicionalmente auto regresivos (CAR) con un conjunto de distribuciones condicionales univariadas. Lunn et al. (2012) citando a Besag et al. (1991) comentan que una de las formulaciones más usada es la siguiente:

$$S_i|S_{\setminus i} \sim Normal\left(\sum_{j\neq i}\frac{w_{kj}S_j}{w_{k+}}, \frac{w^2}{w_{k+}}\right) \quad (27)$$

Donde w_{kj} son pesos utilizados para expresar dependencia espacial entre las locaciones j y k, se tiene además que $w_{kj} = w_{jk}$, $w_{kk} = 0$ y $w_{k+} = \sum_j w_{jk}$. Una opción simple y ampliamente usada es que $w_{kj} = 1$ cuando las áreas sean adyacentes, es decir, comportan alguna frontera y $w_{kj} = 0$ cuando no es así. La dependencia espacial debe ser determinada mediante alguno de los criterios de vecindad; puede ser Torre o Reina.

2.6 Modelos Espaciales y Jerárquicos en el Estudio de la Relación entre Plaguicidas y Cáncer

Una vez definidos los conceptos importantes de las técnicas estadísticas correspondientes para realizar el análisis de datos, se retomará la discusión sobre la relación existente entre los plaguicidas y el cáncer. En este contexto, es fundamental considerar que el estudio de dicha relación se enmarca en la epidemiología. Fajardo (2017) define la epidemiología como la rama de la medicina que estudia la distribución de las enfermedades en la

población y sus determinantes. Por lo tanto, es crucial entender las técnicas estadísticas para abordar adecuadamente la investigación sobre el impacto de los plaguicidas en la letalidad del cáncer de mama y próstata.

La incorporación de modelos jerárquicos y espaciales es esencial para comprender mejor estas relaciones. Monsalve (2013) comenta que la variabilidad espacial es algo intrínseco en los datos, además de que los modelos jerárquicos con un enfoque bayesiano permiten combinar la información espacial de manera efectiva. Lunn et al. (2012) destacan que los modelos espaciales son ampliamente utilizados en estudios ambientales y ecológicos, y mencionan que la manera más flexible de modelar dependencia espacial es a través de una distribución de efectos aleatorios en un modelo jerárquico. Estos autores también señalan que el tipo de modelo a utilizar depende del tipo de información espacial disponible, ya sean coordenadas espaciales o unidades geográficas como distritos.

Morgenstern (1995) comenta que los estudios ecológicos son muy utilizados en la epidemiología cuando se tiene información de grupos de personas en lugar de datos individuales. Las medidas ecológicas, como las características físicas del lugar donde los grupos viven o trabajan, son cruciales para estos estudios. En este caso específico, el interés de este estudio es utilizar este tipo de medida ecológica ambiental, dado que se contará con información por distritos o cantones respecto a la exposición a plaguicidas de las personas que residen en ellos.

Hoeting (2009) y Morgenstern (1995) advierten que ignorar la autocorrelación espacial puede llevar a conclusiones erróneas respecto a la significancia de las variables analizadas, debido a que los errores estándar pueden estar subestimados. Monsalve (2013) añade que, en tales casos, la significancia estadística puede estar sobreestimada. Por lo tanto, los modelos de regresión que consideran la estructura espacial son cruciales para abordar adecuadamente estas relaciones y evitar supuestos incorrectos de independencia de las observaciones.

CAPÍTULO 3: Estudio empírico

3.1 Metodología

El objetivo de la presente investigación es identificar un método para el análisis de datos de cáncer en Costa Rica que cuentan con una estructura jerárquica, así como la inclusión de un componente espacial que permita capturar la autocorrelación espacial de los mismos. Para el cumplimiento de tal objetivo se realizó un estudio empírico, así como un estudio de simulación el cuál será detallado en el siguiente capítulo.

3.1.1 Datos

Los datos e información que contempla el estudio se obtuvieron del Instituto de Investigaciones en Salud con autorización de la exdirectora PhD. Carolina Santamaría de dicha Unidad Académica se presentan el cuadro 3.1.1.

Cuadro 3.1.1 Datos por utilizar según tipo de la fuente de la información, institución y período de referencia.

Datos	Fuente	Institución	Período
Indicadores de uso de plaguicidas	Encuesta	Instituto Regional de Estudios en Sustancias Tóxicas de la Universidad Nacional (IRET/UNA)	1999-2000
Censo Nacional Agropecuario	Censo	Instituto Nacional de Estadística Censos, INEC	1984
Registro Nacional de Tumores	Registro	Ministerio de Salud	2011-2015
Proyecciones Oficiales de Población	Registro	Instituto Nacional de Estadística Censos, INEC	2002-2013
Límites geográficos de Costa Rica	Registro	Sistema Nacional de Información Territorial	-
Censo de variedades de caña de azúcar de Costa Rica	Censo	LIGA AGRÍCOLA INDUSTRIAL DE LA CAÑA DE AZÚCAR	2000
Defunciones	Registro	Tribunal Supremo de Elecciones	Actualizado hasta mayo 2019
Clasificación de componentes tóxicos	Registro	Agencia de Protección Ambiental de Estados Unidos (EPA), Agencia Internacional para la Investigación del Cáncer (IARC), la Red Internacional de Acción contra los Plaguicidas (PAN).	-

Como se puede observar del cuadro 3.1.1, la información asociada a los casos de cáncer corresponder al período del 2011 al 2015, sin embargo, la información de uso de plaguicidas refiriere a un período de entre 30 a 15 años de antigüedad al período antes mencionado. Esta diferencia entre los períodos se debe a como anteriormente fue mencionado, al período de latencia, es decir, entre el momento de la exposición a los plaguicidas y al momento en que aparece la enfermedad.

Cavalier et al. (2023) citando al IARC indican que basado en experiencia con humanos el período entre la primera exposición y la aparición del cáncer en algunas ocasiones superior a los 20 años y que períodos sustancialmente inferiores de 30 años no pueden evidenciar falta de evidencia de carcinogenicidad. Es por esto que en la presente investigación se hipotetiza que al utilizar este período de latencia de 15 a 30 de años de exposición se puede evidenciar la relación entre la exposición a plaguicidas y la mortalidad por estos casos específicos de cáncer.

3.1.2 Casos de cáncer

En este caso específico se plantea analizar los casos de las personas que fueron diagnosticadas con cáncer de mama y cáncer de próstata.

El archivo de datos relacionado a tumores fue brindado por el Registro Nacional de Tumores y contiene información de las personas que han sido diagnosticadas con cáncer desde el año 1980 hasta el año 2015, los años utilizados para esta este trabajo corresponden al período del 2011 al 2015. Además, el archivo incluye el nombre, número de cédula, fecha de nacimiento, fecha de diagnóstico del cáncer, código del cáncer con el que fue diagnosticada la persona, así como la provincia, cantón y distrito donde reside la persona. Se tiene información de personas extranjeras diagnosticadas con cáncer, sin datos respecto a provincia, cantón, distrito, por lo que son considerados como datos perdidos; específicamente 900 personas entre el 2011 y 2015. Se destaca que hay personas sin número de cédula, pero con la información referente a la unidad administrativa de residencia por lo que se incluyen en el análisis.

Por otra parte, de algunas personas solamente se tiene información de la provincia, o de la provincia y cantón de residencia, por lo que para no perder tales observaciones se realiza una asignación de estas personas utilizando como referencia a las personas que tienen todos

los datos necesarios. A modo de ejemplo, se presentará la forma de asignar el código de cantón en caso de que no esté presente: suponiendo que se tiene la información incompleta (solamente se conoce la provincia) de 100 personas que residen en Cartago, provincia que está compuesta por 8 cantones, con la información disponible se obtiene una distribución de frecuencias, es decir, qué proporción de personas diagnosticadas reside en cada cantón; la información hipotética se presenta el cuadro 3.1.2.1

Cuadrado 3.1.2.1. Distribución porcentual de la población del cantón de Cartago según distrito.

Distrito	Proporción
Cartago	0.08
Paraíso	0.15
La Unión	0.06
Jiménez	0.10
Turrialba	0.35
Alvarado	0.10
Oreamuno	0.12
El Guarco	0.04

La asignación de las 100 personas sin cantón sería la siguiente utilizando la información del cuadro 3.1.2.2:

Cuadro 3.1.2.2. Distribución de los casos diagnosticadas del cantón de Cartago según distrito.

Distrito	Casos
Cartago	8
Paraíso	15
La Unión	6
Jiménez	10
Turrialba	35
Alvarado	10
Oreamuno	12
El Guarco	4

Para las personas para las cuales no se posee información referente al distrito de residencia el procedimiento a seguir es el mismo que el descrito en los cuadros 3.1.2.1 y 3.1.2.2. Para cada cantón se obtiene la frecuencia relativa de casos de cáncer para cada distrito y las

observaciones incompletas (sin distrito) son asignadas de manera proporcional a cada uno de ellos.

Una vez que se tiene la información del código PCD (provincia, cantón, distrito) se procede a calcular las tasas de cáncer para cada distrito o cantón, en este caso se tiene un conteo de casos por lugar (ya sea distrito o cantón) por tipo de cáncer. Para este estudios los tipos de cáncer son cáncer de mama y próstata

Las tasas de morbilidad se calculan de la siguiente forma:

$$Tasa\ bruta\ por\ grupo\ cancerígeno = \frac{Casos\ de\ cáncer\ por\ grupo\ diagnósticados}{Población\ a\ mitad\ de\ período * Años} \quad (28)$$

La ecuación (28) será utilizada para calcular las tasas de cáncer para los cánceres de mama y próstata por grupo quinquenal de edad por distrito o cantón. La variable "Años" hace referencia a la cantidad de años transcurridos entre el inicio y fin del período.

La tasa bruta de la expresión 28 tiene la desventaja, según Arriaga (1996), que las tasas calculadas están afectadas por la estructura según edades de la población y no mide adecuadamente el fenómeno de interés. Debido a esta razón se procede a calcular las tasas estandarizadas; como menciona el mismo autor, el objetivo que se persigue al realizar tal cálculo es que la estructura por grupos de edad no afecte las tasas calculadas.

Arriaga (1996), de igual manera, indica que existen algunos problemas cuando se estandarizan las tasas, entre los cuales se tienen que si se presenta un cambio en la población que se utiliza para realizar la estandarización, se pueden producir cambios en el rango del nivel de mortalidad. Por otra parte, menciona que hacer comparaciones históricas de estas tasas estandarizadas no es una tarea sencilla.

Para realizar la estandarización de las tasas, se utilizó como población base la que se presenta en el cuadro 3.2.

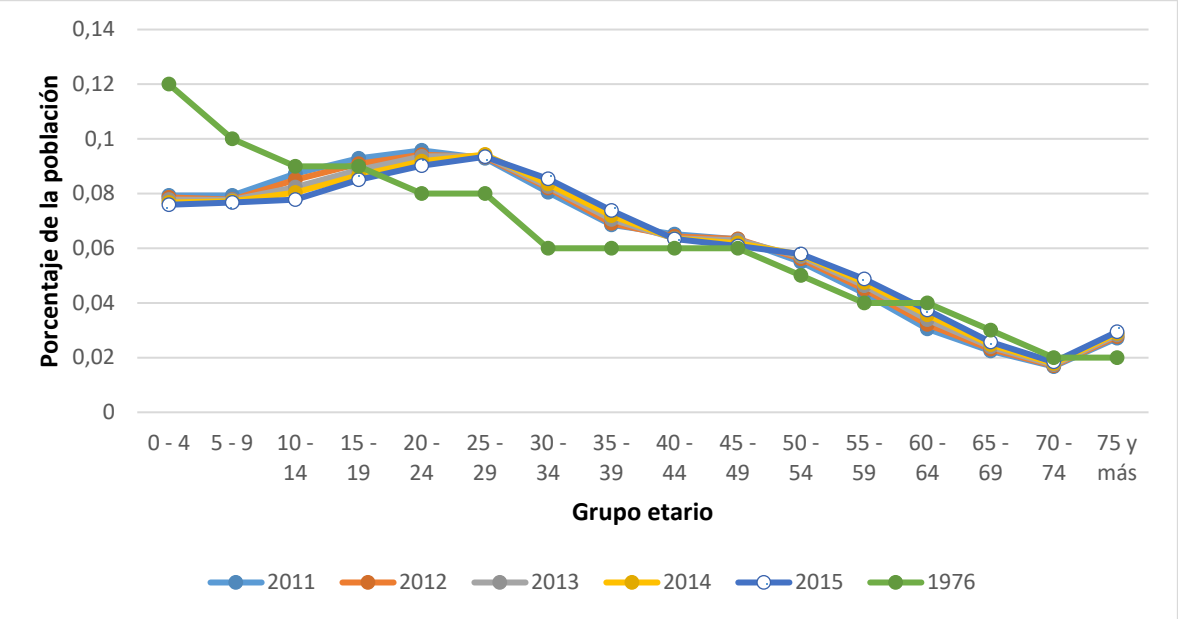
La población base que se muestra en el cuadro 3 corresponde a un proceso de estandarización de la población de Costa Rica de los períodos del 2011 al 2015. Los datos fueron obtenidos de las proyecciones de población realizadas por el INEC.

Inicialmente se procedió a graficar las curvas de población según grupo etario para cada uno de los años en el período del 2011 al 2015, para determinar si se tenían comportamientos similares o si por el contrario en algún período se tenía un comportamiento distinto, además se comparó con la población estándar propuesta por Waterhouse et al. en 1976.

Cuadro 3.2 Población estándar según grupo de edad.

Grupo de edad	Población estándar
0 a 4 años	7 764
5 a 9 años	7 770
10 a 14 años	8 251
15 a 19 años	8 880
20 a 24 años	9 307
25 a 29 años	9 354
30 a 34 años	8 283
35 a 39 años	7 079
40 a 44 años	6 397
45 a 49 años	6 245
50 a 54años	5 673
55 a 59 años	4 621
60 a 64 años	3 394
65 a 69 años	2 403
70 a 74 años	1 747
75 años y más	2 832
Total	100 000

Gráfico 3.1.1 Distribución porcentual de la población según grupo etario para Costa Rica (2011-2015) y población mundial (1976).



Fuente: INEC. (2018) Proyecciones Oficiales de Población.

Del gráfico 3.1.1 se pudo determinar que el comportamiento de la población para cada grupo etario en el período analizado era muy similar y además la composición poblacional era muy distinta a la población mundial estándar de 1976. Las principales diferencias se encuentran para los primeros grupos de edad, donde Costa Rica presenta bajos porcentajes de población, así como para el grupo de 20 a 39 años de edad, donde en Costa Rica este grupo hay un mayor porcentaje de población respecto a la población estándar.

Por lo tanto, se procedió a calcular un promedio de las poblaciones del período del 2011 al 2015 y se utilizó como la población estándar para el cálculo posterior de las tasas.

Las tasas de morbilidad por grupos quinquenales son calculadas a partir de la ecuación 24, seguidamente se multiplican por la población estándar, para de esta manera tener una cantidad de casos esperados. Para finalizar, el resultado de esta multiplicación se divide entre la suma total de la población estándar (100 000) y se amplía por 1 000.

Respecto a la estandarización de las tasas, Silva et al. (2003) indican que es una manera de la estadística clásica en la que se ha intentado corregir la variabilidad que se puede presentar en las mismas debido a las diferentes estructuras por edad presentes en los grupos a analizar, por ejemplo, al hacer comparaciones de diversos cantones o distritos.

3.1.3 Exposición a plaguicidas

En el análisis empírico se propone el uso de la variable Índice de Exposición a Plaguicidas (IEP), tal variable se tiene por cantones, lo cual indica que se podría considerar el análisis espacial de los datos. Es importante destacar que tener la información para una unidad geográfica más pequeña no es factible.

La información referente a exposición a plaguicidas presenta la peculiaridad de que no se puede obtener de manera individual para las personas incluidas en este estudio. Para poder realizar el cálculo de un índice de exposición a plaguicidas se necesita obtener la cantidad de áreas sembradas por cultivo, además de la cantidad de ingredientes activos que se usan por área para un determinado producto.

Se tiene dos fuentes de datos para obtener la información de plaguicidas. Una de ellas corresponde al Censo Nacional Agropecuario que se llevó a cabo en el año 1984, por el Instituto Nacional de Estadística y Censos; del cual se tiene la información para cada provincia cantón y distrito sobre la cantidad de hectáreas por cultivo que han sido tratadas con herbicidas u otro tipo de plaguicidas. A partir de estos datos se procede a calcular un total distrital de hectáreas. El total distrital surge de la suma de la cantidad de hectáreas tratadas con herbicidas u otras sustancias para cada cultivo, luego esos totales se suman para todos los cultivos que se siembran dentro de un distrito.

La otra fuente de información referente a plaguicidas son las encuestas elaboradas por el Instituto Regional de Estudios en Sustancias Tóxicas de la Universidad Nacional (IRET/UNA), donde se obtiene un listado de ingredientes activos (ai) según cultivo, así como dosis o cantidad de uso y su frecuencia por cantón.

Además, se utiliza información de las proyecciones oficiales de población que son realizadas por el Instituto Nacional de Estadística y Censos, del cual se obtiene información referente al tamaño de la población por año según cantón. Para el cultivo de piña los datos se obtuvieron del Sistema Nacional de Información Territorial el cual brinda información sobre el área del paisaje productivo de piña en Costa Rica en el año 2000. Finalmente, para la caña de azúcar la información se extrae del Censo de Variedades de Caña de Azúcar de Costa Rica 2000.

Haciendo uso de las tres fuentes antes mencionadas, se plantea la construcción de un Índice de Exposición a Plaguicidas, el mismo es una modificación del índice de exposición a plaguicidas propuesto por Wesseling et. al (1999). El índice original se conforma de la siguiente manera:

$$IEP = \frac{\sum_{i=1}^{m} h_l * n_l * a_l}{Población\ total\ en\ cada\ cantón} \quad (29)$$

Donde:

l: cultivos agrícolas (1,2, ... ,m)
h_l: hectáreas tratadas con plaguicidas para cada cultivo i
n_l: promedio estimado anual de aplicaciones de plaguicida para cada cultivo i
a_l: factor de corrección por aplicación aérea de plaguicidas para cada cultivo i

De tal forma, el nuevo índice contiene las hectáreas de cada cultivo, la población proyectada por el INEC en cada cantón, la extensión en hectáreas del cantón y el peso anual promedio de plaguicida aplicado según ingrediente activo. Además, para cada ingrediente activo se hace una corrección según su factor de toxicidad. Los niveles de toxicidad son estimados a partir de Highly Hazardous Pesticides (HHP), mientras que el factor de toxicidad es un ranking de la siguiente forma:

0 → No hay riesgo

1 → Riesgo muy bajo

2 → Riesgo bajo

3 → Riesgo alto

4 → Riesgo muy alto

Se contempla la información sobre la toxicidad de cada ingrediente activo mediante la recopilación de las distintas entidades encargadas de la clasificación de componentes tóxicos, entre ellos: EPA (2021), la Agencia Internacional para la Investigación del Cáncer (OMS, 2021) y la Red Internacional de Acción contra los Plaguicidas (PAN por sus siglas en inglés, 2021). Además, se decide usar una cuarta clasificación para los ingredientes activos que están prohibidos en la actualidad debido a su toxicidad en humanos. A partir de las diversas fuentes se asigna un puntaje según como clasifican al ingrediente activo, posteriormente a partir de la suma de puntajes para cada ingrediente activo se clasifica en la escala de niveles de toxicidad.

Ahora bien, en cuanto a los cultivos, para este estudio se incluyeron seis: arroz, banano, caña de azúcar, café, piña y palma africana; debido a que fueron los principales cultivos sembrados en el territorio nacional. Finalmente, el indicador se compone de la siguiente forma:

$$IEP = \frac{\sum_{l=1}^{m} h_l * (c_l * F_l)}{Densidad\ de\ población\ en\ el\ cantón} \quad (30)$$

Donde:

l: cultivos agrícolas (1,2, ... ,m)
h_l: hectáreas tratadas con plaguicidas para cada cultivo i
c_l: peso anual promedio de aplicaciones realizadas para el cultivo i
F_l: corrección por toxicidad para el ingrediente activo j

Para el cálculo del IEP se utiliza el factor de corrección de toxicidad para el ingrediente activo utilizado considerando que no todos los ingredientes son igualmente peligrosos. Para el cálculo del índice se obtiene para cada cantón una cuantificación de la exposición a los plaguicidas corregido por toxicidad, es decir, dándole más peso a plaguicidas más tóxicos, finalmente se divide esta exposición entre la densidad poblacional del cantón para

determinar en cuales cantones existe mayor exposición debido a la cantidad de personas que residen en el mismo.

Tal índice de exposición a plaguicidas se estandariza para que varíe entre 0 a 1, donde 0 es el cantón con la menor exposición a plaguicidas y el valor de 1 corresponde al cantón que tiene una mayor exposición.

3.1.4 Análisis de datos: Modelos de regresión

El análisis incluye una variable dependiente que sigue una distribución binomial, específicamente se evalúa la muerte o no de las personas a partir de un conjunto de variables independientes; la cuales se describen en esta sección.

El manejo adecuado de este conjunto de datos requiere utilizar un modelo de regresión lineal mixto espacial, debido a la estructura jerárquica de los datos además de la asociación espacial que se espera ocurra entre los datos.

La letalidad de las personas diagnosticadas con cáncer de próstata y mama se analizará mediante la estimación de un modelo para hombres y otro para mujeres debido a que la esperanza de vida de las mujeres es usualmente mayor a la de los hombres. (Rogers, R et al., 2010)

Las principales variables para analizar corresponden a:

Dependiente:

- Muerte de las personas diagnosticados con cáncer de mama y próstata (sí, no).

La selección de las personas diagnosticadas con cáncer de mama y próstata responde a lo mencionado por Santamaria (2009, pp. 9) y Xu et al. (2010) que indican que algunos pesticidas pueden mimetizar hormonas sexuales como estrógeno, así como la testosterona. Como parte de los resultados obtenidos por Xu et al. (2010) se encontró que los pesticidas organoclorados estaban positivamente asociados con la prevalencia del padecimiento de cáncer de próstata. Por otra parte, Mink et al. (2008) así como Pardo et al. (2020) indican que, en el caso de los hombres expuestos directamente por su trabajo a la exposición a plaguicidas, su riesgo de padecer cáncer de próstata incrementa.

Independientes:

- Índice de exposición a plaguicidas (cantonal, distrital).

El índice de exposición a plaguicidas ha sido seleccionado como una de las variables independientes en esta investigación para poder determinar cuál es su relación con la mortalidad de las personas diagnosticadas con cáncer según la relación entre dicha variable y la incidencia de enfermedades de cáncer.

- Edad.

Fernandes et al. (2023) indican que en el caso de las mujeres diagnosticadas con cáncer la edad es un factor de pronóstico importante, ya que a mayor edad las mujeres tenían una menor probabilidad de sobrevivencia.

La variable edad en el estudio de Xu et al. (2010) estuvo significativamente asociada con la presencia de pesticidas organoclorados, personas en el mayor grupo de edad tenían concentraciones más altas de pesticidas organoclorados que las personas en el menor grupo de edad.

- Tasa de incidencia de cáncer según grupo.

La inclusión de las tasas de morbilidad de cáncer, donde se consideran el cáncer de mama y de próstata es debido al riesgo de muerte que estas enfermedades generan luego del diagnóstico. En el caso de los dos cánceres seleccionados para esta investigación En Cheng et al. (2022) comentan que existe un exceso de riesgo de mortalidad sostenido, pero bajo después del diagnóstico de cáncer.

Asociado a la selección de tasas de incidencia para el análisis de la letalidad de cáncer de mama y próstata Camus & Band (1994), analizaron la relación entre mortalidad e incidencia de cáncer, cuantificando que el exceso de mortalidad de Montreal era atribuible al exceso de la letalidad, por lo tanto, en esta investigación se desea analizar que existe entre la incidencia de la enfermedad y la letalidad de la misma.

Se estandarizó la variable de tasas de cáncer para que de igual manera que el IEP varie de 0 a 1. La estandarización se realizó con el objetivo de que el rango de las variables no afectara la estimación de los coeficientes, de igual manera para que los coeficientes de regresión puedan ser comparables. (Kutner et al., 2004)

A partir de las variables antes mencionadas se ajustan los diferentes modelos de regresión bayesianos. Seguidamente se debe de evaluar los diversos criterios de convergencia, así

como la pertinencia de agregar otro nivel, es decir, que ganancia en varianza explicada se tiene al incluir la estructura jerárquica.

3.1.5 Especificación de las distribuciones a priori

Parte muy importante de los resultados obtenidos al realizar análisis de datos desde una perspectiva bayesiana, depende de la selección de las distribuciones a priori que se realice. La selección de esta puede surgir como resultado de un criterio subjetivo de la persona investigadora, partiendo desde su conocimiento.

De igual manera se suele hacer referencia a la selección de la distribución a priori clasificándolas como priori informativa o no informativa. En el segundo caso se tiene menos información respecto al fenómeno de interés que en el primer caso, donde se tiene información respecto al parámetro que se desea estimar.

Es importante recalcar que aun cuando se selecciona una priori denominada informativa, se debe tener precaución, ya que puede llegar a influir los resultados; además se podrían generar estimaciones de coeficientes que no tengan sentido con los datos analizados.

En este caso, la distribución a priori de la intersección de los diferentes modelos de regresión se determinó como una distribución normal con promedio 0 y precisión de 0.0001 (es decir, varianza 10000), con la finalidad de que sea una distribución previa no informativa. De igual manera para la distribución del coeficiente asociado a la variabilidad espacial su distribución es loggamma con un parámetro de forma de 0.5 y de tasa de 0.0005, debido a que como menciona Lunn et al. (2012) se recomienda usar previas débilmente informativas ya que normalmente se cuenta con poca información.

3.1.6 Modelo de regresión a estimar.

Finalmente se definirá cual es el modelo de regresión de manera general (tanto para hombres como para mujeres) que se procederá a estimar en las siguientes secciones. El mismo es presentado a continuación:

$$logit\left[E\left(y_{ij}\right)\right] = \beta_{0j} + Edad_{ij}\beta_{1j} + Tasa_j\, B_{2j} + IEP_j\, B_{3j} + s_j \quad (31)$$

$$\beta_{1j} = \beta_{11}$$

$$\beta_{2j} = \beta_{22}$$

$$\beta_{3j} = \beta_{33}$$

$$\beta_{0j} = \beta_{00} + u_{oj};\ u_{oj} \sim N(0, \tau_{00})$$

Donde:

ij: representan a los individuos i dentro del conglomerado j.

j: representa al conglomerado.

y_{ij}: corresponde a la variable respuesta donde se tienen los valores 1: persona fallecida, 0: persona viva.

β_{0j}: es el intercepto. En el caso de incluir el conglomerado se utilizará como efecto aleatorio, de lo contrario será utilizado como un efecto fijo.

β_1: representa el coeficiente de regresión asociado a la variable edad.

β_2: representa el coeficiente de regresión asociado a la variable de tasa de cáncer especifica de mama o próstata. Efecto fijo.

β_3: representa el coeficiente de regresión asociado a la variable del índice de exposición a plaguicidas.

s_j: es el parámetro que considera la correlación entre vecindarios.

u_{oj}: efecto específico del distrito en el intercepto.

Es importante destacar que se plantea el mismo modelo para hombres y mujeres, además se tiene el supuesto de que la migración interna en el país es baja y no afecta en las estimación de la exposición a plaguicidas y por ende letalidad.

3.2 Resultados del estudio empírico

3.2.1 Análisis exploratorio de datos espaciales

Se hará un análisis previo de la información relacionada a las tasas de cáncer y el índice de exposición a plaguicidas a partir de una representación gráfica con los mapas. Ante la presencia de valores extremos se debe analizar el comportamiento de los datos a partir de histogramas y gráficos de cajas, en especial la información relativa a exposición.

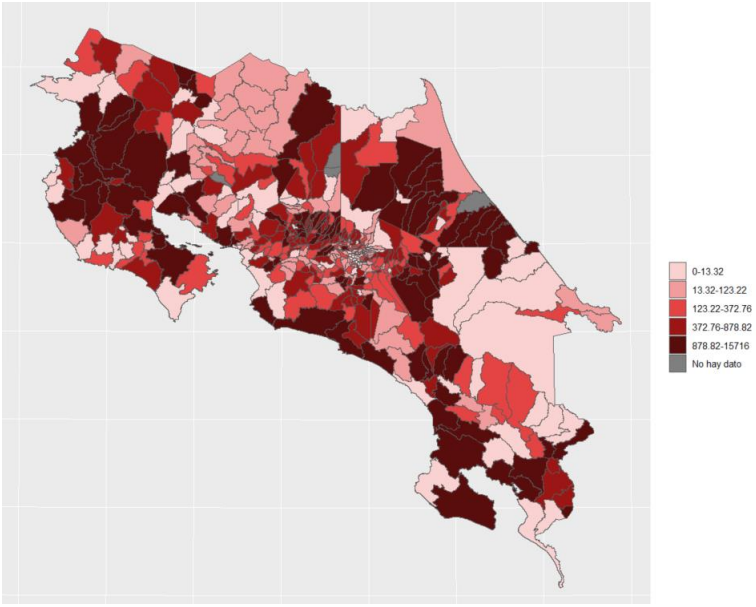
La información generada a partir de las dos fuentes de información que se tienen para plaguicidas se va a mostrar en mapas, ya que como menciona Silva, L et al. (2003), la

visualización de los datos en mapas brinda más información de los patrones que se pueden mostrar geográficamente respecto a si se muestran de manera tabular.

En relación con la primera fuente de información del Censo Agropecuario de 1984 sobre los plaguicidas se va a trabajar a nivel distrital. La distribución de la cantidad de hectáreas cultivadas y tratadas con plaguicidas se muestra en el mapa de la Figura 3.2.1.

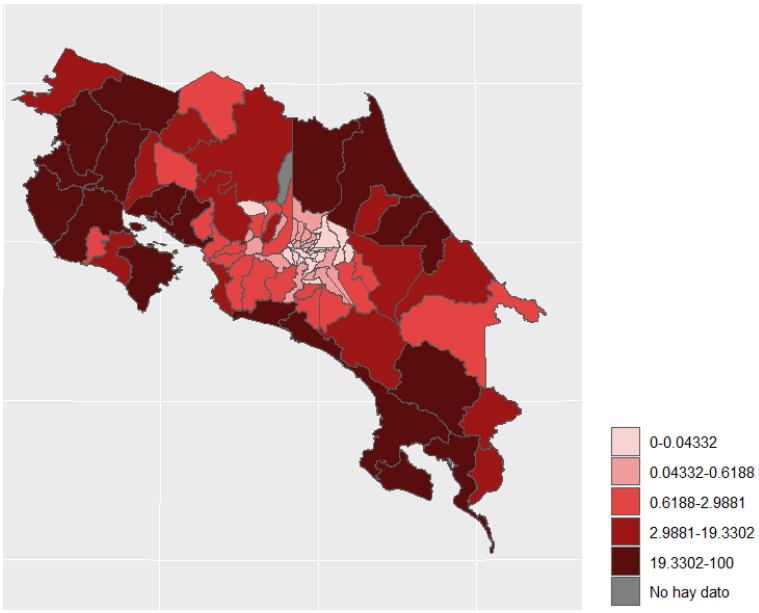
De la figura 3.2.1 se puede observar que la provincia de Limón tiene una distribución bastante homogénea respecto a la cantidad de hectáreas tratadas con plaguicidas. En el caso de las otras provincias costeras (Guanacaste y Puntarenas) se puede observar que hay áreas con poca siembra como los distritos de Garita, Santa Elena, Agua Buena, La Cuesta entre otros. Sin embargo, hay otros distritos donde se tiene una gran cantidad de hectáreas tratadas con estos productos como lo son Santa Cruz, Belén De Nosarita, Corredores, Parrita entre otros. Finalmente, el panorama de las 4 provincias centrales es complicado de observar debido a que los distritos son más pequeños, sin embargo, se puede observar que en algunos lugares tales como el distrito de Aguas Zarcas, Venecia en Alajuela y Puerto Viejo en Heredia donde hay una alta concentración de plaguicidas.

Figura 3.2.1. Distribución por distritos de la cantidad de hectáreas tratadas con plaguicidas. 1984



Seguidamente se muestra la distribución de exposición a plaguicidas a nivel cantonal medida por el Índice de Exposición a Plaguicidas en la figura 3.2.2.

Figura 3.2.2. Distribución por cantones del índice de exposición a plaguicidas estandarizado. 2000.



En el mapa de la Figura 3.2.2 se observa que la provincia de Limón tiene cantones con una alta exposición a plaguicidas, al igual que algunos cantones de Guanacaste, en la parte central del país se puede observar que hay una menor concentración de cantones expuestos a plaguicidas.

Es importante comentar que para ambos mapas la escala de los gráficos se construyó utilizando los quintiles de las variables a analizar, es decir, hectáreas que fueron tratadas con herbicidas u otros plaguicidas y según los quintiles del Índice de Exposición a Plaguicidas estandarizado respectivamente. Destacando que los valores de la figura 3.2.1 no están estandarizados. Además es importante resaltar que existen unidades geográficas para las cuales no había información disponible sobre la exposición a plaguicidas.

3.2.2 Verificación del supuesto de autocorrelación espacial.

En este caso se realizó la prueba de hipótesis de autocorrelación espacial, asociadas al coeficiente de correlación I de Moran, para las tasas de cáncer desde el período de 1980 al

2015 calculadas por grupo de cáncer, de manera distrital y en décadas. En todos los casos estas hipótesis se pusieron a prueba con un 5% de significancia, en este caso se tuvo suficiente evidencia estadística para rechazar la hipótesis nula que indica que las tasas de cáncer antes mencionadas se distribuyen de manera aleatoria. De igual manera, al realizar la prueba de hipótesis de autocorrelación espacial para el Índice de Exposición a Plaguicidas, se rechaza la hipótesis nula de no autocorrelación, además cuenta con una magnitud positiva y moderada (0.40).

3.2.3 Resultados descriptivos de las variables.

Se presentan los resultados descriptivos de las variables incluidas en los diferentes modelos de regresión a estimar. Inicialmente se presenta los resultados de los casos de cáncer de mama y próstata de 2011 al 2015. En este conjunto de datos se presenta la información de 10612 personas. La distribución de estas se puede observar en el cuadro 3.2.1.

Cuadro 3.2.1. Distribución de las personas diagnosticadas con cáncer según condición de fallecimiento por sexo, 2011-2015

Muerte	Hombres		Mujeres	
	Conteo	Porcentaje	Conteo	Porcentaje
Sí	1236	25.24	1110	19.42
No	3660	74.74	4606	80.58
Total	4896	100	5716	100

Seguidamente se muestran en el cuadro 3.2.2 medidas de posición y variabilidad de las variables independientes incluidas en el modelo de regresión.

Cuadro 3.2.2. Medidas de posición y variabilidad de la edad, tasa de morbilidad e índice de exposición a plaguicidas por sexo, 2011-2015

Medidas	Hombres			Mujeres		
	Edad	Tasa	IEP	Edad	Tasa	IEP
Promedio	68.07	0.08	0.07	57.49	0.07	0.06
Mediana	69	0.06	0.00	57	0.04	0.00
Desviación estándar	9.71	0.14	0.17	13.26	0.15	0.15
Mínimo	22	0.00	0.00	16	0.00	0.00
Máximo	103	1.00	1.00	101	1.00	1.00

Gráfico 3.2.3.1. Histograma de la tasa de cáncer de mama. 2011-2015.

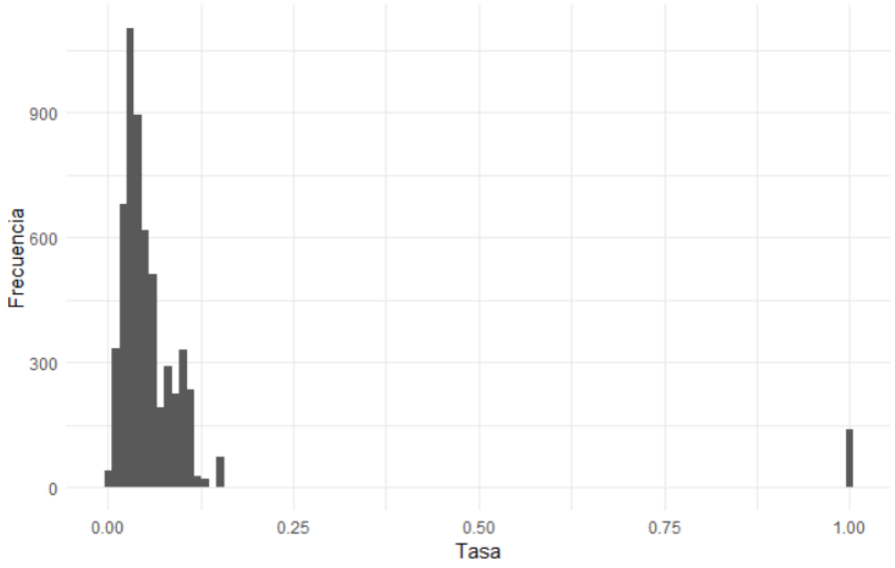
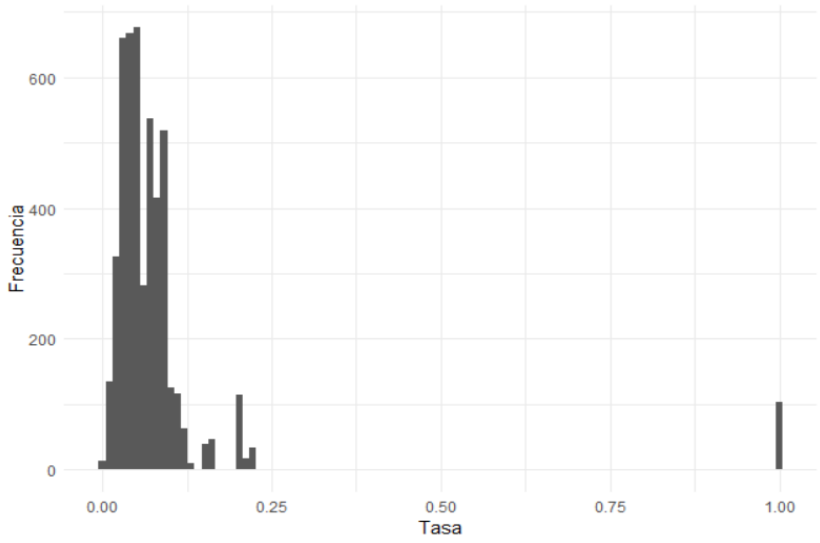


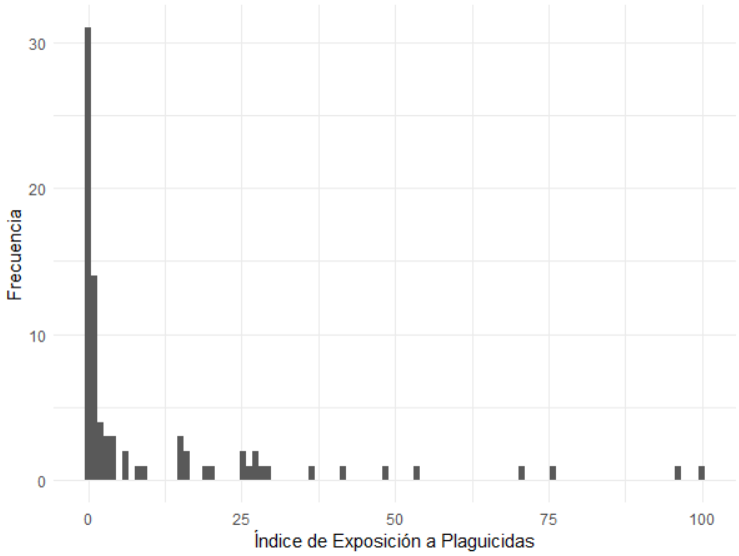
Gráfico 3.2.3.2. Histograma de la tasa de cáncer de próstata. 2011-2015.



Al observar el gráfico 3.2.3.1 y 3.2.3.2 se puede determinar que la variable de tasa de cáncer de mama y próstata presentan un comportamiento bastante asimétrico. La mayoría de los cantones tienen tasas estandarizadas cercanas a valores de entre 0 y 0.001.

Seguidamente se presenta en el gráfico 3.2.3.3, el comportamiento de la variable índice de exposición a plaguicidas (el cuál fue obtenido mediante la ecuación 27).

Gráfico 3.2.3.3 Histograma del índice de exposición a plaguicidas. 1991-2000.



Del gráfico 3.2.3.3 se tiene que se presenta un comportamiento bastante asimétrico, donde la mayoría de los cantones tiene asociado un bajo índice de exposición de plaguicidas. Seguidamente en el gráfico 3.2.3.4 y 3.2.3.5 se presenta el histograma para analizar la forma que posee la variable de la edad según sexo.

Gráfico 3.2.3.4 Histograma de la edad de las personas diagnosticadas con un cáncer de próstata. 2011-2015.

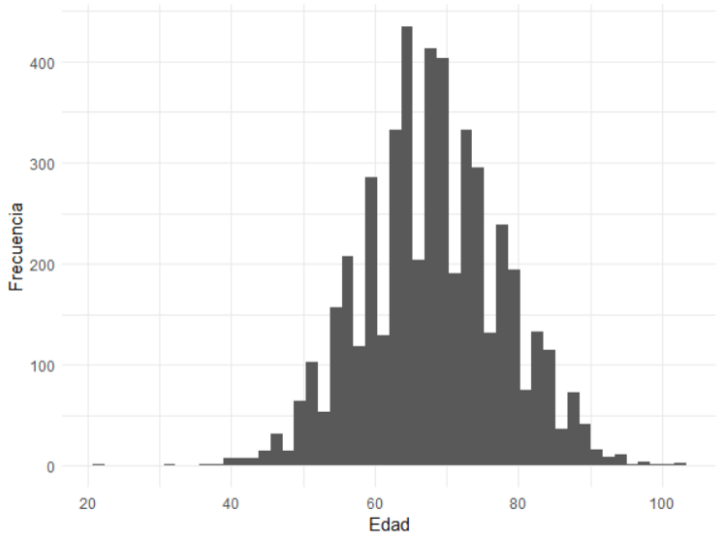
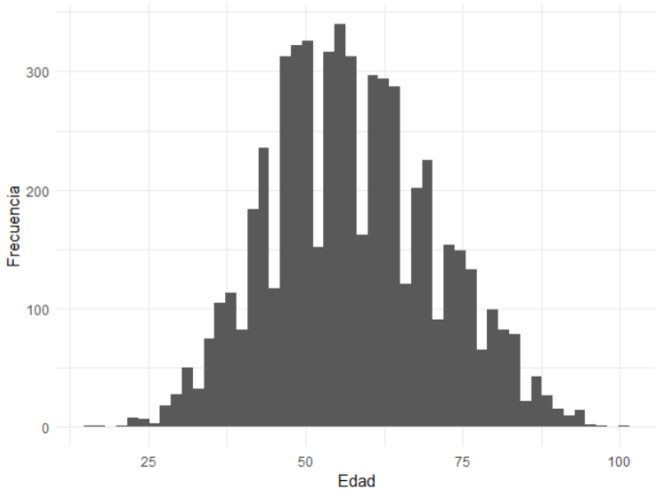


Gráfico 3.2.3.5. Histograma de la edad de las personas diagnosticadas con un cáncer de mama. 2011-2015



Del gráfico 3.2.3.5 se observa un comportamiento bastante simétrico de la variable edad, mientras que en el gráfico 3.2.3.4 que la edad presenta una leve asimetría negativa dado que la mayoría de las personas que se analizaron en este caso tienen edades entre los 60 y 80 años (55% en el caso de los hombres y 51.3% en el caso de las mujeres). Sin embargo, en el caso de las mujeres un 7.7% de ellas tienen edades de los 20 a los 40, destacando que existen 55 (1%) mujeres de muy baja edad (menores de los 20 años) que han sido diagnosticados con cáncer de mama. La distribución de frecuencias de la edad a la que las personas fueron diagnosticadas con cáncer para ambos sexos se encuentra en el anexo 5.

En la figura 3.2.3.1 y 3.2.3.2 se presenta el comportamiento espacial de las tasas de cáncer de mama y próstata del período del 2011 al 2015, donde se puede observar que se presenta una concentración de valores altos en lugares de Puntarenas y Guanacaste, más no tan claro en la provincia de Limón.

En los anexos 3 y 4 se puede encontrar un acercamiento del área central del país tanto para las tasas de cáncer de próstata como para las tasas de cáncer de mama.

Figura 3.2.3.1. Distribución por distritos de las tasas de cáncer de mama. 2011-2015.

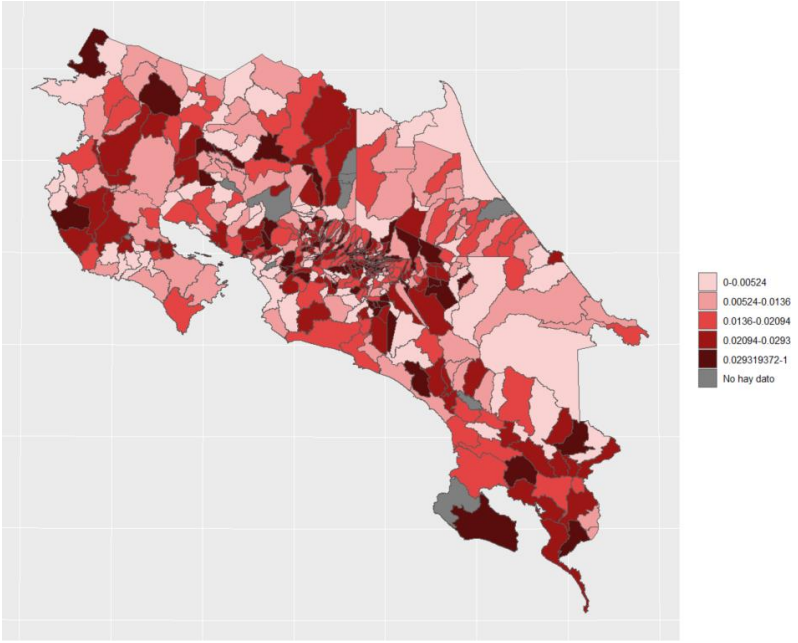
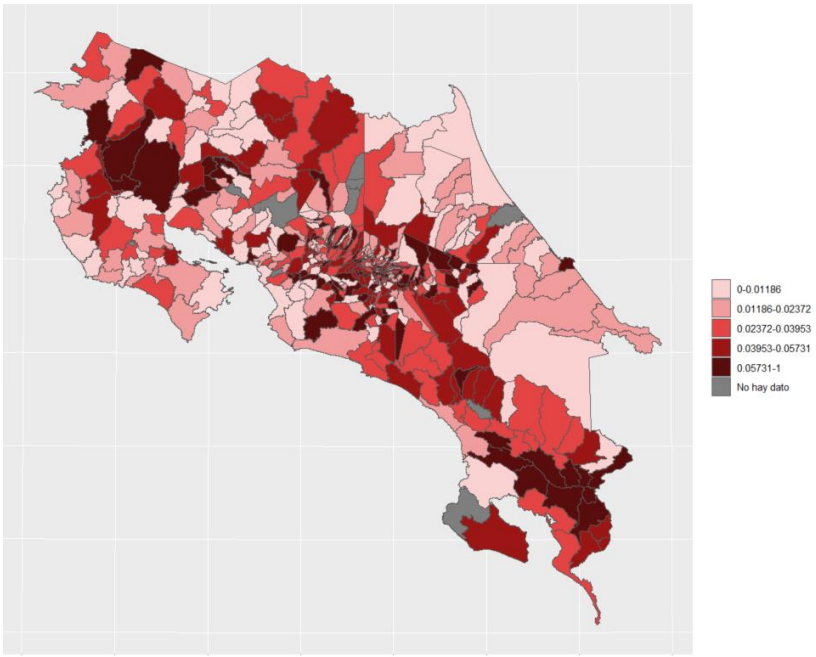


Figura 3.2.3.2. Distribución por distritos de las tasas de cáncer de próstata. 2011-2015.



3.2.4. Resultados de los modelos estimados

Se estimaron diferentes modelos de regresión jerárquicos: modelo de un solo nivel, modelo jerárquico de dos niveles, modelo jerárquico de dos niveles espacial (variando el criterio de vecindad entre Torre y Reina), tanto para hombres como para mujeres. Los resultados de la estimación se muestran en los siguientes cuadros.

A continuación, se describen los resultados de los modelos estimados para los hombres, destacando que estos mismos modelos se estimaron también para las mujeres. No obstante, debido a que el comportamiento observado en las mujeres es muy similar al descrito para los hombres, con la salvedad de que la magnitud de los coeficientes es diferente, los resultados se presentan en la sección de anexos (ver Anexo 6).

El cuadro 3.3.1 muestra que, en términos generales, al agregar la estructura jerárquica tanto el criterio de información de Akaike Watanabe y el criterio de información de devianza se mantienen constantes en todos los casos. Los cambios se presentan en la *log* verosimilitud marginal donde esta decrece cuando se incluyen los componentes espaciales. El comportamiento en los indicadores WAIC podría indicar un peor ajuste de los modelos espaciales respecto a los más simples, sin embargo, como es mencionado por Oaks et al. (2019) agregar un parámetro a un modelo puede disminuir la verosimilitud marginal, lo cual generaría que se prefiera un modelo más simple.

Un resultado muy importante en este caso es que ambos modelos espaciales tienen valores de precisión muy elevados para tal componente, lo cual indica que se tiene una muy baja variabilidad (debido a que la precisión es el equivalente a la inversa de la varianza), sin embargo, es preferible el modelo con el criterio de vecindad de torre ya que tiene una precisión más baja. Al comparar el modelo jerárquico espacial que utiliza como criterio de vecindad “reina” contra el criterio “torre” presentan un comportamiento similar, es decir, que independientemente del criterio utilizado se presentó básicamente el mismo desempeño.

Al analizar los resultados de los coeficientes estimados se tiene que en general se presentó el comportamiento esperado, donde por cada aumento de un año los hombres son

aproximadamente un 10% más propensas de morir (OR=1.099). Por otro lado, a mayor tasa de cáncer de próstata, menor riesgo de morir (OR=0.69). Finalmente, por cada aumento de una unidad en el índice de exposición a plaguicidas se tiene que los hombres diagnosticados con cáncer de próstata son aproximadamente un 52% más propensas de fallecer.

Es importante destacar que la interpretación de los diferentes coeficientes de regresión únicamente se presentará para el primer caso mas no para los siguientes debido a la similitud que presentan todos los resultados en términos de coeficientes estimados.

Seguidamente, se procedió a realizar el mismo análisis con la salvedad de que en esta ocasión se realizaron agrupamientos de los distritos por los cuales está conformado el país, donde esta agrupación se realizó dentro de cada provincia y es distinto al agrupamiento generado por los cantones. Para observar el agrupamiento que se realizó de las áreas, el cuadro resumen se muestra en el anexo 1.

Al evaluar los resultados del cuadro 3.3.2 se observa que de igual manera que en el cuadro 3.3.1 se presentan diferencias en el rendimiento del modelo básico (que no toma en cuenta la estructura jerárquica) respecto al modelo jerárquico y los modelos espaciales. Esto se determinó a partir de las medidas de bondad de ajuste analizadas, ya que, se presenta una disminución en los valores analizados como el DIC y WAIC así como en la log verosimilitud marginal. Los indicadores determinan que no parece existir una ganancia al utilizar la estructura jerárquica espacial respecto a la estructura jerárquica o dicho de otra manera y teniendo en cuenta a Oaks (2019) no hay una gran desmejora al tener un modelo más complejo.

Seguidamente y para el cumplimiento del objetivo de creación de potenciales conglomerados, se realizó un análisis de conglomerados, para generar una nueva agrupación de las unidades de estudio. Para la creación de los conglomerados, se analizó el comportamiento de los distritos según la exposición a plaguicidas, las tasas de cáncer y la clasificación del distrito según el grado de urbanización. Tal clasificación es obtenida del INEC del Manual de Clasificación Geográfica con Fines Estadísticos de Costa Rica (2016),

donde se clasifican las unidades geográficas en: urbano, predominantemente urbano, rural y predominante rural.

Debido a que las variables utilizadas para la generación de conglomerados no corresponden al mismo nivel de medición, se utilizó la distancia de Gower para agrupar los distritos. (Gower, 1971). Una vez generados los conglomerados dentro de cada provincia basado en la similitud de estos, se procedió a verificar la validez geográfica de los mismos, es decir, que geográficamente la unión de los distritos fuera válida. Para ello se generaron mapas con los resultados de los agrupamientos y se realizaron uniones de distritos que hayan resultado del análisis de conglomerados y que compartieran frontera geográficas. Para observar el agrupamiento que se realizó de las áreas, el cuadro resumen se muestra en el anexo 2.

Cuadro 3.3.1 Resultados de los modelos de regresión bayesianos estimados para la unidad geográfica de distritos entre la letalidad por cáncer de próstata y la edad, tasa de cáncer e Índice de Exposición a Plaguicidas (N=4896).

Modelo	Criterio de vecindad	Coeficientes				Criterios de bondad de ajuste			Precisión	
		Intercepto	Edad	Tasa de cáncer	Índice de Exposición a Plaguicidas	Log verosimilitud marginal	DIC	WAIC	Estructura jerárquica	Estructura espacial
De un nivel	-	-7.757*	0.095*	-0.371*	0.415*	-2457.33	4880.45	4880.62		
Jerárquico no espacial	-	-7.757*	0.095*	-0.371*	0.414*	-2457.47	4880.02	4880.51	19905.65	
Jerárquico espacial	Reina	-7.758*	0.095*	-0.369*	0.417*	-2641.27	4880.06	4880.52	19426.11	20534.65
Jerárquico espacial	Torre	-7.758*	0.095*	-0.370*	0.417*	-2650.27	4880.21	4880.51	25206.20	23146.49

*: utilizado para indicar que el intervalo de credibilidad calculado al 90% no contiene al 0.

Cuadro 3.3.2 Resultados de los modelos de regresión bayesianos estimados para la nueva unidad geográfica entre la letalidad por cáncer de próstata y la edad, tasa de cáncer e Índice de Exposición a Plaguicidas (N=4896).

Modelo	Criterio de vecindad	Coeficientes				Criterios de bondad de ajuste			Precisión	
		Intercepto	Edad	Tasa de cáncer	Índice de Exposición a Plaguicidas	Log verosimilitud marginal	DIC	WAIC	Estructura jerárquica	Estructura espacial
De un nivel	-	-7.750*	0.095*	-0.476*	0.428*	-2458.84	4879.66	4879.83		
Jerárquico no espacial	-	-7.750*	0.095*	-0.456*	0.428*	-2456.87	4879.37	4879.79	2100.83	
Jerárquico espacial	Reina	-7.751*	0.095*	-0.456*	0.429*	-2478.92	4879.41	4879.79	20716.79	18815.07
Jerárquico espacial	Torre	-7.751*	0.095*	-0.456*	0.429*	-2480.38	4879.49	4879.78	26462.75	24852.01

*: utilizado para indicar que el intervalo de credibilidad calculado al 90% no contiene al 0.

Del cuadro 3.3.3 se puede observar que, en términos generales al utilizar la unidad geográfica resultante de la agrupación de conglomerados, la estimación de los coeficientes de regresión es igual en todos los casos a excepción del caso de tasa de cáncer. Al analizar los intervalos de credibilidad al 90% se puede observar que el valor de 0 está incluido para esta variable por lo que se destaca que este coeficiente no tiene relevancia práctica.

En términos de precisión, se observa que el modelo con criterio de reina se captura de mejor manera el componente espacial respecto al modelo de torre ya que presenta una log verosimilitud marginal inferior y las estimaciones para este componente son más precisas. Para el caso del componente jerárquico, tiene un mejor desempeño el modelo no espacial y seguidamente el modelo con criterio de vecindad de torre indicado a partir de la precisión de la estructura jerárquica.

Los resultados de la estimación de los diversos modelos estimados para las unidades de geográficas anteriormente descritas se muestran en los cuadros 3.3.4 a 3.3.6.

Cuadro 3.3.3. Resultados de los modelos de regresión bayesianos estimados para la unidad geográfica resultante de la agrupación de conglomerados entre la letalidad por cáncer de próstata y la edad, tasa de cáncer e Índice de Exposición a Plaguicidas (N=4896).

Modelo	Criterio de vecindad	Coeficientes				Criterios de bondad de ajuste			Precisión	
		Intercepto	Edad	Tasa de cáncer	Índice de Exposición a Plaguicidas	Log verosimilitud marginal	DIC	WAIC	Estructura jerárquica	Estructura espacial
De un nivel	-	-7.807*	0.095*	0.218	0.336*	-2458.15	4884.65	4884.71		
Jerárquico no espacial	-	-7.808*	0.095*	0.233	0.336*	-2458.3	4884.58	4884.73	20300.91	
Jerárquico espacial	Reina	-7.808*	0.095*	0.234	0.336*	-2473.05	4884.6	4884.76	17688.7	21429.53
Jerárquico espacial	Torre	-7.808*	0.095*	0.234	0.336*	-2473.46	4883.6	4884.76	18268.9	18227.55

*: utilizado para indicar que el intervalo de credibilidad calculado al 90% no contiene al 0

En el cuadro 3.3.4 se muestran los resultados del agrupamiento de los individuos por distritos, modelo que utiliza distribuciones previas no informativas (mencionadas anteriormente) para la estimación de la precisión de los coeficientes de la estructura jerárquica y de la estructura espacial.

En el cuadro 3.3.5 se presentan los resultados del modelo estimados para la nueva unidad geográfica (cuadro 3.3.2), incluyendo distribuciones previas para las partes aleatorias del modelo.

Es importante mencionar que, según las medidas de bondad de ajuste en el primer análisis realizado (cuadro 3.3.4) utilizando como unidad de agrupamiento el distrito, parece indicar que utilizar en el modelo con el componente jerárquico, tiene un mejor desempeño. El comportamiento se repite al utilizar el nuevo agrupamiento propuesto, sin embargo, las diferencias entre las log verosimilitudes marginales de los modelos son muy pequeñas. Se debe de resaltar que cuando se utilizan los distritos existen 39 distritos que solamente tienen la información de una persona, mientras que al utilizar el nuevo agrupamiento únicamente quedan dos grupos que tienen la información de una persona, dando de esta manera estimaciones más confiables de la varianza.

En ambos casos (cuadro 3.3.4 y 3.3.5) es preferible utilizar entre los modelos espaciales, el modelo que tiene una distribución previa únicamente para el componente espacial y que utiliza el criterio de reina ya que tiene una precisión en el componente espacial más bajo que en los otros casos.

Al realizar las comparaciones de la nueva agrupación con respecto a la agrupación de distritos, se prefiere la segunda debido a que los comportamientos entre modelos son más similares en términos de bondad de ajuste.

Cuadro 3.3.4. Resultados de los modelos de regresión bayesianos estimados para la unidad geográfica de distritos utilizando distribuciones previas entre la letalidad por cáncer de próstata y la edad, tasa de cáncer e Índice de Exposición a Plaguicidas (N=4896).

Modelo	Criterio de vecindad	Distribución previa		Coeficientes				Criterios de bondad de ajuste			Precisión	
		Jerárquica	Espacial	Intercepto	Edad	Tasa de cáncer	Índice de Exposición a Plaguicidas	Log verosimilitud marginal	DIC	WAIC	Estructura jerárquica	Estructura espacial
Jerárquico espacial	Reina	Normal(0,0.001)	Loggamma(0.5, 0.0005)	-7.785*	0.096*	-0.357	0.424*	-2643.69	4880.28	4880.58	2766.63	467.59
Jerárquico espacial	Torre	Normal(0,0.001)	Loggamma(0.5, 0.0005)	-7.771*	0.095	-0.359	0.434	-2654.51	4879.91	4880.8	26930.4	1189.66
Jerárquico no espacial	-	Normal(0,0.001)	-	-7.758*	0.095*	-0.371	0.413*	-2461.35	4878.49	4880.61	2.89E+19	
Jerárquico espacial	Reina	-	Loggamma(0.5, 0.0005)	-7.769*	0.095*	-0.358	0.439*	-2641.19	4880.14	4880.76	18831.75	944.35
Jerárquico espacial	Torre	-	Loggamma(0.5, 0.0005)	-7.769*	0.095*	-0.359	0.439*	-2650.51	4880.22	4880.78	18243.69	1040.17

*: utilizado para indicar que el intervalo de credibilidad calculado al 90% no contiene al 0.

Cuadro 3.3.5. Resultados de los modelos de regresión bayesianos estimados para la para la nueva unidad geográfica utilizando distribuciones previas entre la letalidad por cáncer de próstata y la edad, tasa de cáncer e Índice de Exposición a Plaguicidas (N=4896).

Modelo	Criterio de vecindad	Distribución previa		Coeficientes				Criterios de bondad de ajuste			Precisión	
		Jerárquica	Espacial	Intercepto	Edad	Tasa de cáncer	Índice de Exposición a Plaguicidas	Log verosimilitud marginal	DIC	WAIC	Estructura jerárquica	Estructura espacial
Jerárquico espacial	Reina	Normal(0,0.001)	Loggamma(0.5, 0.0005)	-7.757*	0.095*	-0.452*	0.435*	-2482.62	4879.79	4880.3	1.68E+18	1.33E+03
Jerárquico espacial	Torre	Normal(0,0.001)	Loggamma(0.5, 0.0005)	-7.758*	0.095*	-0.452*	0.436*	-2484.07	4879.67	4880.4	6.48E+10	1.43E+03
Jerárquico no espacial	-	Normal(0,0.001)	-	-7.750*	0.095*	-0.456*	0.428*	-2460.46	4879.51	4879.81	3.97E+45	
Jerárquico espacial	Reina	-	Loggamma(0.5, 0.0005)	-7.757*	0.095*	-0.451*	0.438*	-2478.89	4879.81	4880.38	20786.43	1108.02
Jerárquico espacial	Torre	-	Loggamma(0.5, 0.0005)	-7.757*	0.095*	-0.451*	0.438*	-2480.22	4879.92	4880.38	19304.11	1169.56

*: utilizado para indicar que el intervalo de credibilidad calculado al 90% no contiene al 0.

En el cuadro 3.3.6 se presentan los resultados de los modelos estimados haciendo uso de distribuciones previas para los componentes jerárquico y jerárquico espaciales.

En términos de estimación de coeficientes el comportamiento en todos los casos es muy similar. Por otra parte, los modelos que cuentan con una distribución previa tanto para el componente jerárquico como para el espacial tienen una precisión muy alta para el componente jerárquico. Mientras que para el caso de los modelos con distribución únicamente para el componente espacial tienen precisiones más bajas.

Para los dos primeros escenarios evaluados los modelos con criterio de torre presentan una mejor precisión para el componente espacial respecto a los modelos con criterio de reina.

Al comparar las estimaciones del modelo utilizando las agrupaciones resultantes de los conglomerados con respecto a los otros dos casos, en términos de estimación de los coeficientes se tienen comportamientos bastante similares y medidas de bondad de ajuste levemente mejores. Los dos agrupamientos realizados de igual manera presentan un mejor comportamiento en términos de precisión y bondad de ajuste que el agrupamiento original de los distritos indicando de esta manera que la creación de unidades geográficas más grandes funciona de mejor manera.

Al comparar los resultados de los modelos estimados para la unidad geográfica de distritos y para la nueva unidad geográfica generada, el coeficiente de regresión asociado a la tasa de exposición a plaguicidas presenta un valor negativo y además su intervalo de credibilidad no incluye el valor de 0. Este resultado es consistente con los resultados obtenidos al utilizar una distribución previa para la estimación de tales coeficientes.

Al analizar el modelo de regresión cuya unidad espacial corresponde a los conglomerados generados, se tiene que tal coeficiente de regresión es positivo, sin embargo, su intervalo de credibilidad si contiene el valor de 0.

El coeficiente de regresión negativo para la variable tasas de cáncer asociado a los modelos de regresión con la unidad original de distritos y la nueva unidad geográfica, es un resultado inesperado. Se espera que, a mayor prevalencia del tipo de cáncer bajo estudio, mayor sea la letalidad por el mismo.

Esta dirección inversa en estos modelos puede deberse a que las agrupaciones de distritos y la nueva geográfica de distritos dentro de provincias no son las ideales para capturar esas relaciones, como si se obtuvo en el modelo cuya unidad geográfica es de conglomerados.

Se esperaría que al incluir determinantes sociales de la salud al modelo cuya unidad geográfica resulta del uso de conglomerados, esta relación mantenga un coeficiente positivo pero que además tenga un peso importante, es decir, que su intervalo de credibilidad no incluya el 0.

Cuadro 3.3.6. Resultados de los modelos de regresión bayesianos estimados para la unidad geográfica resultante de la agrupación de conglomerados utilizando distribuciones previas entre la letalidad por cáncer de próstata y la edad, tasa de cáncer e Índice de Exposición a Plaguicidas (N=4896).

Modelo	Criterio de vecindad	Distribución previa		Coeficientes				Criterios de bondad de ajuste			Precisión	
		Jerárquica	Espacial	Intercepto	Edad	Tasa de cáncer	Índice de Exposición a Plaguicidas	Log verosimilitud marginal	DIC	WAIC	Estructura jerárquica	Estructura espacial
Jerárquico espacial	Reina	Normal(0,0.0001)	Loggamma(0.5, 0.0005)	-7.815*	0.095*	0.285	0.336*	-2477.27	4885.50	4885.09	30508.20	1836.76
Jerárquico espacial	Torre	Normal(0,0.0001)	Loggamma(0.5, 0.0005)	-7.813*	0.095*	0.253	0.337*	-2478.02	4884.97	4885.27	3.38E+12	973
Jerárquico no espacial	-	Normal(0,0.0001)	-	-7.808*	0.095*	0.228	0.336*	-2460.82	4884.62	4884.71	8.51E+118	
Jerárquico espacial	Reina	-	Loggamma(0.5, 0.0005)	-7.813*	0.095*	0.253	0.337*	-2473.46	4884.92	4885.26	20532.51	1279.12
Jerárquico espacial	Torre	-	Loggamma(0.5, 0.0005)	-7.813*	0.095*	0.248	0.337*	-2473.88	4884.98	4885.25	19092.8	1057.13

*: utilizado para indicar que el intervalo de credibilidad calculado al 90% no contiene al 0.

CAPÍTULO 4: Estudio de simulación

Para el desarrollo del estudio de simulación se hará uso de los resultados obtenidos del estudio empírico de letalidad de los hombres, tanto de los resultados descriptivos de las variables bajo estudio, así como de los diversos modelos estimados; no se incluyen los resultados para el caso de las mujeres debido a la similitud con los resultados para los hombres. Con base a los resultados ofrecidos por el estudio empírico se propone el estudio de simulación donde se pueda indagar como diferentes agrupamientos geográficos de las unidades de estudio, así como criterios de vecindad pueden afectar las estimaciones de los coeficientes.

A continuación, se presenta una descripción de cómo fueron simuladas las variables tanto dependiente como independientes para realizar el estudio, así como los criterios que permitan realizar evaluaciones del desempeño de los estimadores.

4.1 Metodología

Como se mencionó en una sección previa el principal interés de esta investigación se centra en el efecto que pueda tener el uso de plaguicidas en la salud de las personas. Por lo que se procederá a analizar a detalle este coeficiente de regresión, así como su error estándar; para ello se hará uso entre otros, del modelo descrito en la sección 3.1.6. Así como se analizará la predicción correcta de las personas que han fallecido a partir de los diversos modelos.

Con el fin de poder realizar tal evaluación se utilizará un diseño de estudio de simulación para generar conjuntos de datos.

4.1.1 Diseño del estudio de simulación

En cada uno de los conjuntos simulados la variable respuesta va a ser definida en función de las variables independientes pseudoaleatorias las cuales son edad, tasa de morbilidad de cáncer de próstata y la respectiva medición de exposición a plaguicidas, así como un error pseudoaleatorio que genere variabilidad entre las unidades de estudio.

Se utilizará el error cuadrático medio, la raíz del error cuadrático medio y el error absoluto medio para evaluar el desempeño de los estimadores de interés en cada uno de los

escenarios planteados, esto con el fin de poder determinar en qué caso se recupera de mejor manera los parámetros asociados.

Para realizar la generación de los datos, así como de los diferentes escenarios que se procederán a mencionar, se utilizó el software R versión 4.2.

Tabla 4.1. Resumen y descripción de las variables y factores utilizados en el estudio de simulación.

Fatores	Detalles
Edad	Se generó como una variable aleatoria con una distribución normal, con su respectivo promedio de 68.07108 y desviación estándar de 9.705882.
Variable independiente: exposición a plaguicidas	Se establecieron tres escenarios distintos para esta variable tomando en cuenta los resultados obtenidos de los modelos del estudio empírico: • un valor bajo: 0.2 • un valor intermedio: 0.35 • un valor alto: 0.5
Unidades geográficas (mismas agrupaciones utilizadas en el estudio empírico)	• Distritos • Agrupaciones que se realizan de manera manual de las unidades de distritos. • Agrupaciones latentes, construidas a partir de conglomerados; es importante destacar que en este caso se evaluó la validez de tales agrupaciones en función que sean grupos que sean cercanos espacialmente o compartan alguna frontera.
Variable dependiente: Muerte por cáncer de próstata	A partir de las variables antes mencionadas y los coeficientes estimados del estudio empírico se calcula la probabilidad de muerte para cada individuo. Posterior al cálculo de la probabilidad de muerte se comparará con un valor aleatorio

	generado a partir de la distribución uniforme. Si la probabilidad calculada es mayor al valor aleatorio se declara a la persona como fallecida; caso contrario se identificará a la persona como no fallecida

4.1.1.1. Generación de variables.

El tamaño de muestra, o bien, la cantidad de datos a simular para este estudio está relacionado con la cantidad de personas diagnosticadas con cáncer de próstata en los diferentes períodos, únicamente se analizarán los modelos para los casos de cáncer de hombres debido a la similitud de resultados entre hombres y mujeres. Para el período de 2011 al 2015 se tiene un total de 10 612 observaciones, donde 4896 son hombres. Por lo que se generaran un archivo con tal tamaño. Es importante destacar que estas observaciones se van a asignar proporcionalmente a los distritos.

Para determinar la distribución a utilizar para simular las variables independientes se analizó el comportamiento de estas a partir de histogramas. Además de la información descriptiva obtenida de los gráficos, se hizo uso del paquete *rriskDistributions* del software RStudio, mediante el cual se comparan medidas de bondad de ajuste tales como la log verosimilitud, el Criterio de Información de Akaike (AIC, por sus siglas en inglés) y el Criterio de Información Bayesiano (BIC, por sus siglas en inglés). Con la información obtenida se determinó cuál de las distribuciones tiene un mejor ajuste a los datos. Se destaca que se seleccionaron las distribuciones de probabilidad que tuvieran valores más bajos en los criterios mencionados, ya que esto refleja un mejor ajuste.

La generación de la variable edad se hará utilizando la distribución normal, con su respectivo promedio y varianza. Seguidamente se generarán las tasas de morbilidad para cada uno de los distritos haciendo uso de igual manera de la distribución normal con su respectivo promedio y varianza, tales tasas se asignarán a los individuos que inicialmente se asignaron a los distritos.

4.1.1.2 Índice de Exposición a Plaguicidas

Para la generación de la respectiva cuantificación del Índice de Exposición a Plaguicidas se utilizará el Índice de Exposición a Plaguicidas a nivel cantonal.

Así mismo, para la generación de los datos relacionados a la exposición a plaguicidas se propone hacer uso de una distribución normal con un promedio positivo y un término que considere la autocorrelación espacial de los datos.

4.1.1.3 Variable respuesta

Con los datos simulados de todas las variables independientes se procederá a simular la variable respuesta. Inicialmente se calculará una probabilidad de muerte o no muerte, teniendo en cuenta que la variable es dicotómica y que por ende se ajustará a una distribución binomial. Obtenida la probabilidad se comparará con un valor aleatorio generado a partir de la distribución uniforme. En caso de que la probabilidad estimada sea mayor al valor aleatorio se colocará el valor de 1, que identificará a la persona como fallecida; caso contrario se colocará el valor de 0, que identificará a la persona como no fallecida.

En el caso de los hombres, para calcular la probabilidad de muerte para cada individuo se hará uso de los coeficientes estimados del estudio empírico, mediante la siguiente ecuación:

$$\pi_{Muerte} = \frac{e^{-7.758+0.095*Edad+-0.369*Tasa\ de\ cáncer+\beta_p*Índice\ de\ Exposición\ a\ plaguicidas}}{1+e^{-7.758+0.095*Edad+-0.369*Tasa\ de\ cáncer+\beta_p*Índice\ de\ Exposición\ a\ plaguicidas}} \quad (32)$$

En este caso se propone utilizar tres diferentes valores de coeficientes para la variable de exposición a plaguicidas, un valor bajo definido como 0.2, uno intermedio que toma el valor de 0.35 y finalmente uno alto 0.5, estos valores se escogieron con base a los resultados del estudio empírico. Esto con el objetivo de poder determinar en qué combinación de casos de simulación permiten mejor su recuperación.

4.1.1.4 Criterios para variar en los modelos a estimar

Una vez se hayan simulado las variables independientes, así como la variable respuesta (ecuación 32), se procederá a estimar los modelos bajo las condiciones de unión de

unidades geográficas, así como los criterios de vecindad, con el objetivo de determinar en qué caso el coeficiente asociado a la exposición a plaguicidas es mejor recuperado.

4.1.1.5 Unión de diferentes unidades geográficas.

Se utilizaron diferentes agrupamientos de las unidades geográficas, diferentes a los agrupamientos sociodemográficos con los que cuenta el país, es decir, provincia, cantón distrito.

Entre las unidades geográficas a considerar se encuentran:

- Distritos
- Agrupaciones que se realizan de manera manual de las unidades de distritos.
- Agrupaciones latentes, que se construirán a partir de clusters o conglomerados; es importante destacar que en este caso se evaluó la validez de tales agrupaciones en función que sean grupos que espacialmente sean cercanos o compartan alguna frontera.

En este caso se quiere verificar si se tiene un mejor rendimiento con alguna de las agrupaciones que se puedan generar que sea distinta a las sociodemográficas con las que cuenta el país. Esto ya que como menciona Morris (1993) cuando se hace uso de las divisiones geográficas existentes se pueden tener resultados de valores estimados inestables si se tienen localidades con poca población o si se tienen áreas innecesariamente grandes.

En el estudio de enfermedades raras mediante el uso de tasas se tiene que las mismas van a ser inestables y esto va a generar que se tenga una pérdida de detalle en los gráficos a utilizar. Debido a esta problemática que se presenta se considera que una alternativa para su solución es la agregación de unidades geográficas más pequeñas en unidades geográficas más grandes. (Morris, 1993)

Para realizar la unión de los distritos se escogería un punto al azar en el mapa y a partir del mismo se procede a realizar la agrupación de los distritos cercanos al mismo, posterior a esta unión se procede a analizar mediante alguno de los criterios de vecindad cuales son las nuevas unidades geográficas que se consideran vecinas.

4.1.1.6 Criterios de contigüidad de los vecinos.

Sumado a la unión de diversos distritos, se desea conocer si la selección del criterio de contigüidad afecta las estimaciones que se generan. Para ello se utilizarán los siguientes criterios:

- Torre, donde para considerar a una unidad geográfica contigua como vecina debe de compartir un borde geográfico. Es importante tener en cuenta que este tipo de criterio de vecindad no considera a las unidades geográficas que comparten un borde con sus vecinos.
- Reina, este criterio toma en cuenta los vecinos que considera el criterio de torre, además de los "vecinos de los vecinos", es decir, que considera las unidades geográficas que si bien no comparten un límite geográfico están cerca. Por lo tanto, sería de esperar que tenga un mejor desempeño respecto al criterio de torre ya que toma en cuenta más información.

La importancia en la selección del método de vecindad y la posterior construcción de la matriz de vecinos antes mencionada radica en como lo mencionan Siabato y Gúzman (2019) en que *"Una elección inapropiada de la matriz puede conllevar resultados imprecisos, o poco relacionados con el fenómeno analizado, que pueden generar decisiones incorrectas y la implementación de políticas con impactos adversos."* (pp. 5)

Debido a esto es que surge la inquietud de conocer en el contexto específico de cáncer y plaguicidas como la relación entre ellos puede variar dependiendo de los criterios de contigüidad utilizados, o por el contrario si la relación es consistente aun cuando se usan criterios distintos.

4.1.1.7 Escenarios planteados

Tomando en consideración todas consideraciones especificadas en las secciones unión de diferentes unidades geográficas (4.1.5.1) y criterios de contigüidad de los vecinos (4.1.5.2) se plantean los diversos escenarios planteados para la estimación de los modelos de regresión:

- Agrupamiento original de las unidades geográficas, coeficiente de regresión de plaguicidas bajo.
- Agrupamiento original de las unidades geográficas, coeficiente de regresión de plaguicidas medio.
- Agrupamiento original de las unidades geográficas, coeficiente de regresión de plaguicidas alto.
- Agrupamiento de las unidades geográficas por provincia, coeficiente de regresión de plaguicidas bajo.
- Agrupamiento de las unidades geográficas por provincia, coeficiente de regresión de plaguicidas medio.
- Agrupamiento de las unidades geográficas haciendo uso de conglomerados, coeficiente de regresión de plaguicidas alto.
- Agrupamiento de las unidades geográficas haciendo uso de conglomerados, coeficiente de regresión de plaguicidas bajo.
- Agrupamiento de las unidades geográficas haciendo uso de conglomerados, coeficiente de regresión de plaguicidas medio.
- Agrupamiento de las unidades geográficas haciendo uso de conglomerados, coeficiente de regresión de plaguicidas alto.

4.1.2. Análisis de los datos

Una vez conformados los nuevos conglomerados, así como la simulación de las variables independientes y dependientes se realiza la estimación del modelo de regresión jerárquico bayesiano con el componente espacial. Para llevar a cabo el análisis de datos estadístico bayesiano espacial se utilizará en el método de INLA, así como su paquete programado en R.

A partir de los datos simulados, en cada uno de los escenarios propuestos se procederá a estimar los siguientes modelos:

- Modelo lineal generalizado, sin estructura jerárquica ni espacial.
- Modelo lineal mixto generalizado, con estructura jerárquica, sin estructura espacial

- Modelo lineal mixto generalizado, con estructura jerárquica y estructura espacial bajo el criterio de vecindad de torre.
- Modelo lineal mixto generalizado, con estructura jerárquica y estructura espacial bajo el criterio de vecindad de reina.

4.1.3 Evaluación del desempeño de los estimadores.

Para evaluar el desempeño del modelo para la recuperación del respectivo coeficiente estimado para la exposición a plaguicidas se utilizarán medidas comunes para cuantificar la exactitud.

Bruno y Moore (2005) definen la exactitud como la distancia total entre los valores estimados (u observados) y los verdaderos valores. Existen varios métodos para poder cuantificar tales distancias y algunas de ellas combinan el sesgo y la precisión. Respecto a estos dos conceptos son definidos por Bruno y Moore (2005) de la siguiente manera:

- Sesgo: La diferencia entre el promedio poblacional y los verdaderos o aceptados valores de referencia.
- Precisión: es una medida estadística de varianza del error de medición, variación de la muestra o del proceso de estimación.

Entre las principales medidas de exactitud definidas por Bruno y Moore (2005) son las siguientes:

- Error cuadrático medio: es el promedio de las diferencias al cuadrado, esta medida indica qué tan cerca está el estimador del verdadero valor.

$$ECM = \frac{1}{n} * \sum_{j-1}^{n} (E_j - A)^2 \quad (33)$$

- Raíz del error cuadrático medio: la única diferencia con la medida de ECM es que está es la raíz cuadrada de tal valor.

$$RECM = \sqrt{\frac{1}{n} * \sum_{j-1}^{n} (E_j - A)^2} \quad (34)$$

- Error absoluto medio: las medidas de las expresiones 33 y 34 pueden verse afectadas por la presencia de valores extremos. El error absoluto medio intenta solventar este problema.

$$MAE = \frac{1}{n} * \sum_{j-1}^{n} \left| E_j - A \right| \quad (35)$$

Donde:

n: La cantidad de simulaciones a realizar.

A: Representa el verdadero valor, en este caso el verdadero valor del parámetro estructural correspondiente a la cuantificación de la exposición a plaguicidas.

E_j: Es el valor observado del coeficiente para cada simulación realizada.

4.2 Resultados

Para determinar la cantidad de repeticiones necesarias para evaluar los estudios de simulación se analizaron distintas cantidades y se evaluó la desviación estándar del coeficiente de exposición a plaguicidas para el modelo con la agrupación original de las unidades geográficas. Los resultados se presentan a continuación:

Cuadro 4.2.1 Error estándar del coeficiente de regresión de exposición a plaguicidas según el modelo estimado por la cantidad de repeticiones que se realizó la simulación, para la unidad geográfica original.

Modelo	Cantidad de repeticiones					
	5	10	25	50	75	100
Bajo, $\beta = 0.2$						
Simple	0.6640	0.6259	0.6938	0.7093	0.6044	0.7549
Jerárquico	0.5786	0.7941	0.8982	0.7418	0.6824	0.7724
Torre	0.5793	0.7962	0.8982	0.7429	0.6833	0.7737
Reina	0.5792	0.7955	0.8961	0.7429	0.6831	0.7739
Medio, $\beta = 0.35$						
Simple	0.5851	0.5586	0.6425	0.6803	0.6423	0.7448
Jerárquico	0.8316	0.6234	0.9123	0.7130	0.6919	0.7641
Torre	0.8336	0.6342	0.9108	0.7132	0.6933	0.7650
Reina	0.8323	0.6338	0.9126	0.7138	0.6932	0.7652
Alto, $\beta = 0.50$						
Simple	0.5857	0.6651	0.6075	0.7105	0.6781	0.7206
Jerárquico	0.8892	0.7028	0.8609	0.6610	0.7545	0.7542
Torre	0.8954	0.7055	0.8643	0.6643	0.7595	0.7573
Reina	0.8905	0.7035	0.8629	0.6619	0.7587	0.7538

Modelo	Cantidad de repeticiones					
	125	150	175	200	225	250
Bajo, $\beta = 0.2$						
Simple	0.6387	0.7007607	0.6969014	0.6461643	0.6314301	0.6960
Jerárquico	0.6620	0.7513573	0.7378462	0.6698749	0.7321519	0.7286
Torre	0.6636	0.7518957	0.7390575	0.6709015	0.7334819	0.7300
Reina	0.6633	0.7523686	0.7395804	0.670977	0.733499	0.7301
Medio, $\beta = 0.35$						
Simple	0.6419	0.6664058	0.6883236	0.6395859	0.6430992	0.6888
Jerárquico	0.7024	0.6787186	0.7700641	0.7173126	0.7094559	0.7330
Torre	0.7032	0.6796144	0.7716843	0.7187456	0.7106478	0.7343
Reina	0.7031	0.679362	0.771569	0.7187941	0.710522	0.7344
Alto, $\beta = 0.50$						
Simple	0.6568	0.6694698	0.69037	0.6619891	0.6453243	0.6884
Jerárquico	0.6867	0.7115801	0.6929984	0.747565	0.7149835	0.7279
Torre	0.6892	0.7132951	0.695842	0.7521215	0.7197978	0.7319
Reina	0.6881	0.7119869	0.6943662	0.7488644	0.7172754	0.7291

Se determino a partir de los resultados del cuadro 4.2.1 que la cantidad optima de repeticiones para cada escenario propuesto es de 125.

A partir de la cantidad de repeticiones seleccionada, se procedió a la simulación de los siguientes escenarios propuestos en la sección 4.1.1.7.

4.2.1 Resultados de la simulación con unidades geográficas originales.

Se procedió a estimar distintos modelos de regresión para determinar en cuál de los casos se logra una mejor recuperación del coeficiente de regresión asociado a exposición a plaguicidas. Los resultados del estudio de simulación de dicho coeficiente se presentan en el siguiente cuadro.

Cuadro 4.2.2 Medidas de posición y variabilidad del coeficiente de regresión de exposición a plaguicidas según el modelo estimado, para la unidad geográfica de distritos.

Modelo	Promedio	Cuartil 1	Mediana	Cuartil 3	Máximo	Mínimo	Desviación estándar
Bajo, $\beta = 0.2$							
Simple	0.159	-0.292	0.097	0.6845	2.167	-2.047	0.6960
Jerárquico	0.1958	-0.2125	0.2105	0.67875	1.928	-1.873	0.7286
Torre	0.1967	-0.2032	0.213	0.68025	1.936	-1.874	0.7300
Reina	0.1966	-0.2117	0.213	0.6795	1.932	-1.876	0.7301
Medio, $\beta = 0.35$							
Simple	0.2659	-0.217	0.238	0.7582	2.059	-1.867	0.6888
Jerárquico	0.3331	-0.13	0.3295	0.774	2.348	-1.627	0.7330
Torre	0.3336	-0.1282	0.3305	0.7765	2.359	-1.629	0.7343
Reina	0.3337	-0.1282	0.3305	0.7782	2.353	-1.63	0.7344
Alto, $\beta = 0.50$							
Simple	0.3597	-0.117	0.3135	0.883	2.288	-17.766	0.6884
Jerárquico	0.5359	0.04374	0.5045	1.04	2.477	-1.217	0.7279
Torre	0.5386	0.0455	0.506	1.0425	2.488	-1.233	0.7319
Reina	0.5365	0.045	0.505	1.0441	2.48	-1.222	0.7291

Figura 4.2.1. Coeficiente de regresión según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica de distritos.

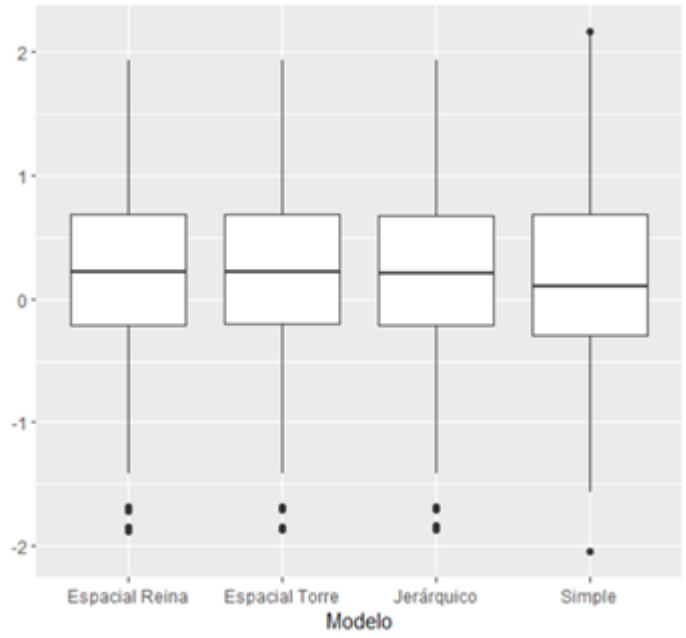


Figura 4.2.2 Coeficiente de regresión según modelo estimado para el caso del coeficiente de regresión medio, unidad geográfica de distritos.

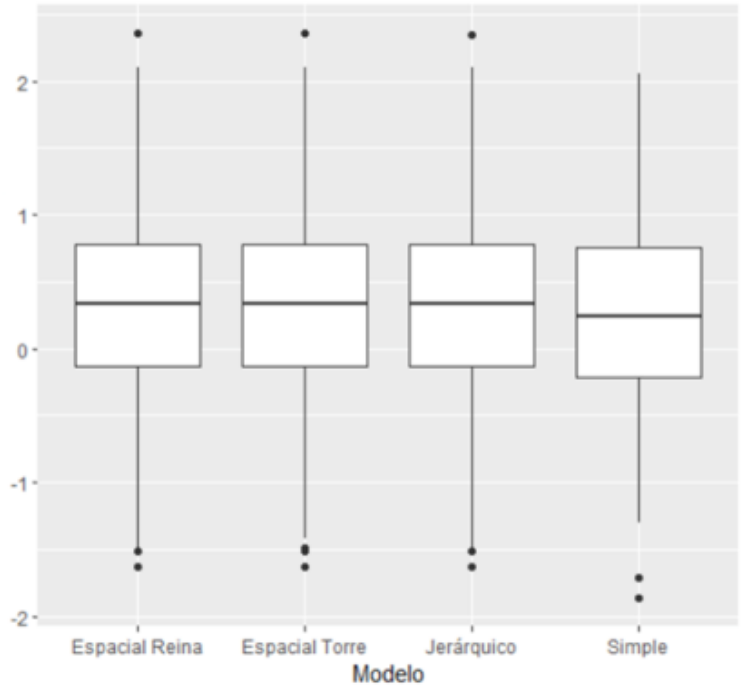
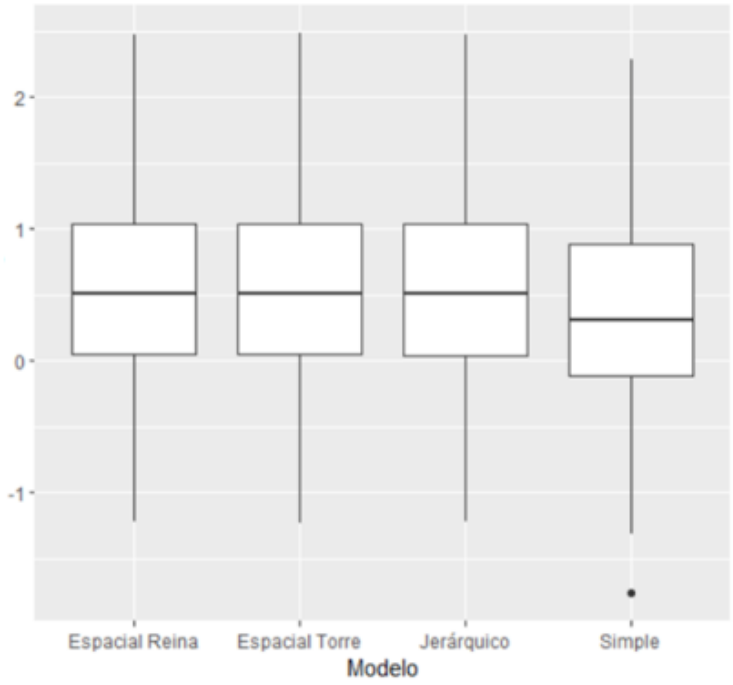


Figura 4.2.3. Coeficiente de regresión según modelo estimado para el caso del coeficiente de regresión alto, unidad geográfica de distritos.



Al observar los promedios se puede determinar que cuando se estimó el modelo simple, este subestima el coeficiente de regresión asociado a la exposición a plaguicidas. En el caso del modelo jerárquico y los modelos jerárquicos espaciales se acercan más al coeficiente utilizado para realizar la simulación. De igual manera sucede cuando se analiza la mediana.

Por otro lado, al analizar los cuartiles, se puede determinar que el recorrido intercuartílico es mayor para el modelo simple respecto a los otros modelos, esto indicar una mayor variabilidad en los valores que toma el coeficiente de regresión en la totalidad de las simulaciones evaluadas.

Al examinar el recorrido en el caso de la simulación con el coeficiente de regresión bajo y alto, se tiene que para el modelo simple se posee la mayor amplitud y los recorridos más pequeños corresponden para los modelos jerárquicos y jerárquicos espaciales. Sin embargo,

para el coeficiente de regresión medio la menor amplitud se presenta para el coeficiente de regresión medio.

Es importante destacar la presencia del valor extremo para el modelo simple cuando la simulación se realizó con un coeficiente de regresión alto.

Al analizar la desviación estándar de los 125 coeficientes de regresión estimador para los diversos modelos para los distintos valores de los coeficientes (bajo, medio y alto), se tiene que el modelo simpe es el que de manera general presentó menor variabilidad respecto a los otros modelos estimados.

Cuadro 4.2.3 Medidas de bondad de ajuste promedio según el modelo estimado, para la unidad geográfica de distritos.

Modelo	LML	WAIC	DIC
Bajo, $\beta = 0.2$			
Simple	-2911.7	5788.17	5788.19
Jerárquico	-2254.2	4158.67	4191.29
Torre	-2448.3	4158.6	4191.18
Reina	-2439.1	4158.62	4191.22
Medio, $\beta = 0.35$			
Simple	-2914.8	5794.36	5794.38
Jerárquico	-2257.2	4163.9	4196.24
Torre	-2451.4	4163.84	4196.14
Reina	-2442.2	4163.85	4196.15
Alto, $\beta = 0.50$			
Simple	-2920.9	5806.48	5806.5
Jerárquico	-2260.7	4169.33	4202.22
Torre	-2449.4	4169.18	4201.94
Reina	-2445.6	4169.26	4202.13

Figura 4.2.4. Log verosimilitud marginal según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica de distritos.

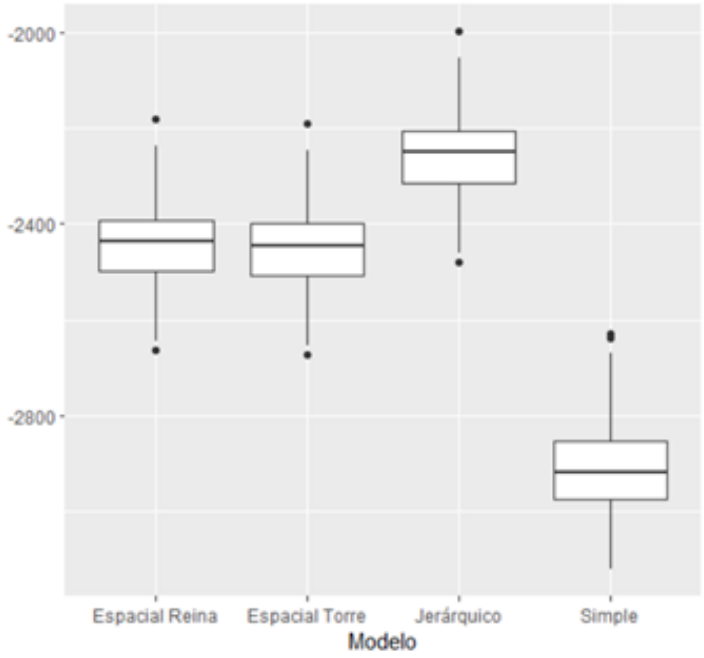


Figura 4.2.5 Criterio de información de Akaike Watanabe según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica de distritos.

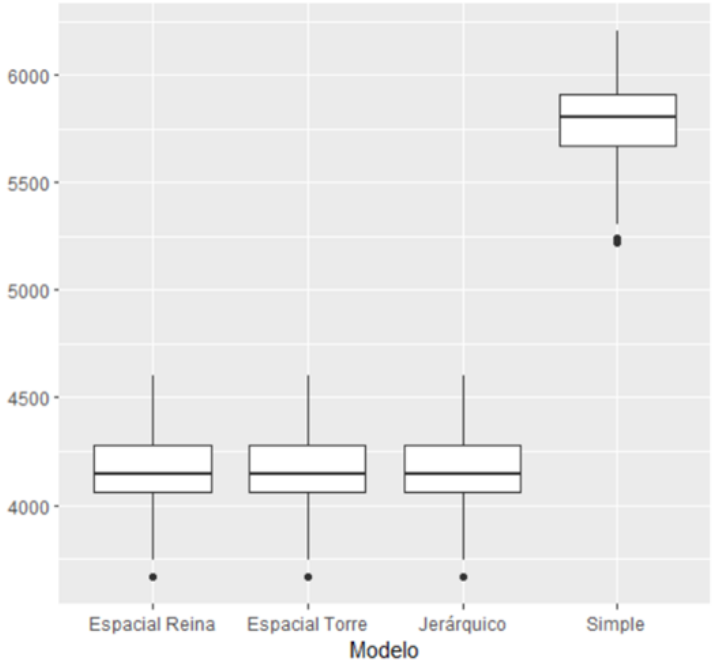


Figura 4.2.6. Criterio de información de deviancia según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica de distritos.

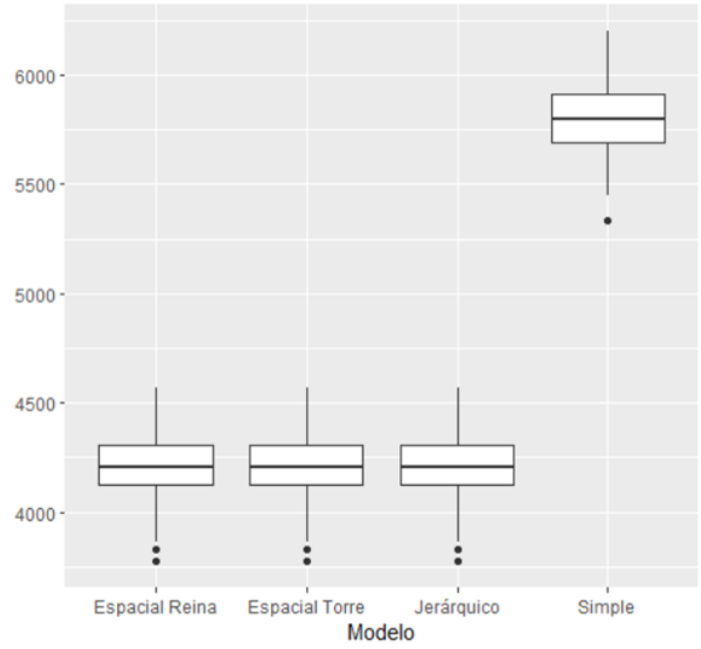


Figura 4.2.7. Log verosimilitud marginal según modelo estimado para el caso del coeficiente de regresión medio, unidad geográfica de distritos.

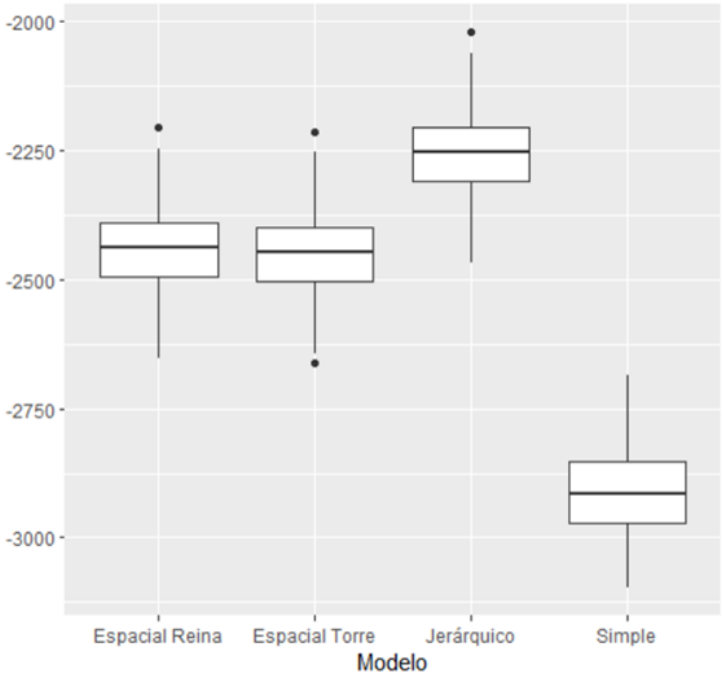


Figura 4.2.8. Criterio de información de Akaike Watanabe según modelo estimado para el caso del coeficiente de regresión medio, unidad geográfica de distritos.

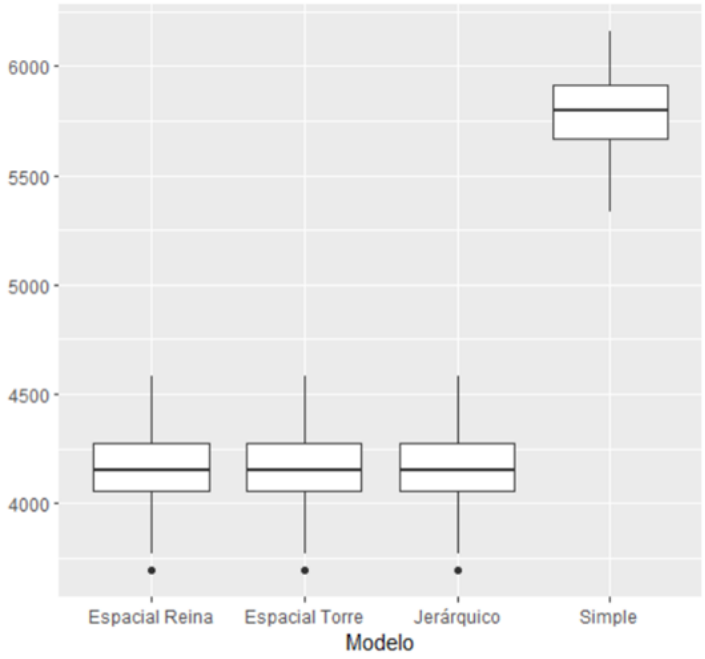


Figura 4.2.9. Criterio de información de deviancia según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica de distritos.

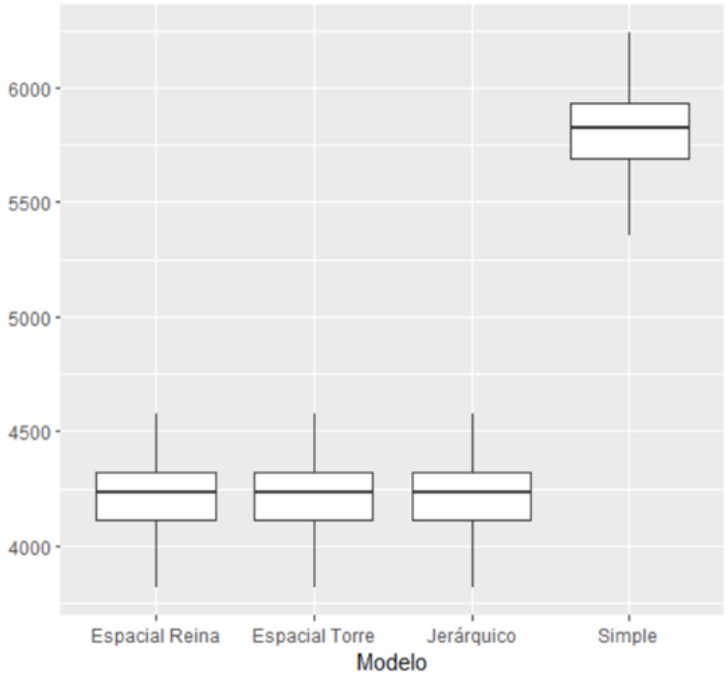


Figura 4.2.10. Log verosimilitud marginal según modelo estimado para el caso del coeficiente de regresión alto, unidad geográfica de distritos.

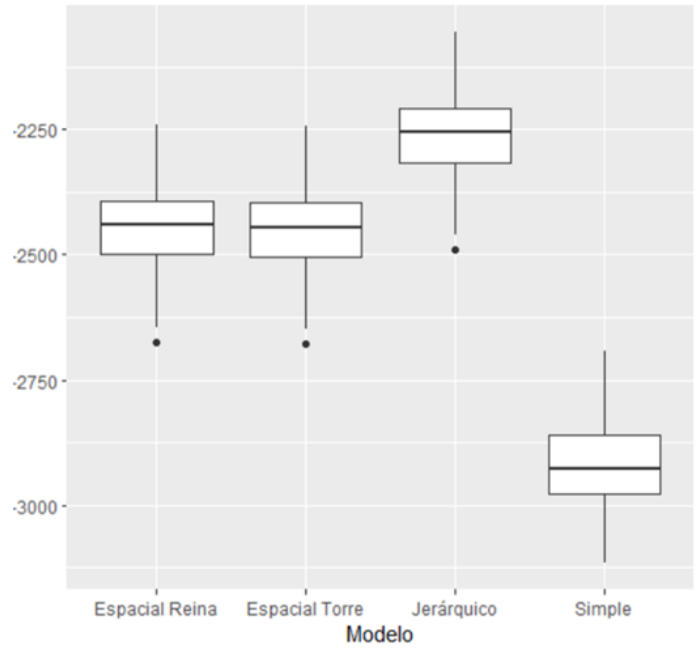


Figura 4.2.11. Criterio de información de Akaike Watanabe según modelo estimado para el caso del coeficiente de regresión alto, unidad geográfica de distritos.

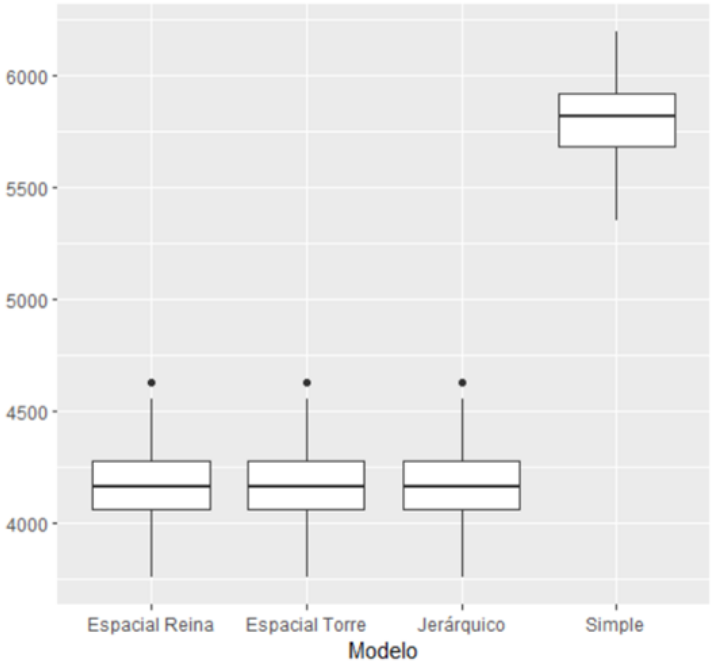
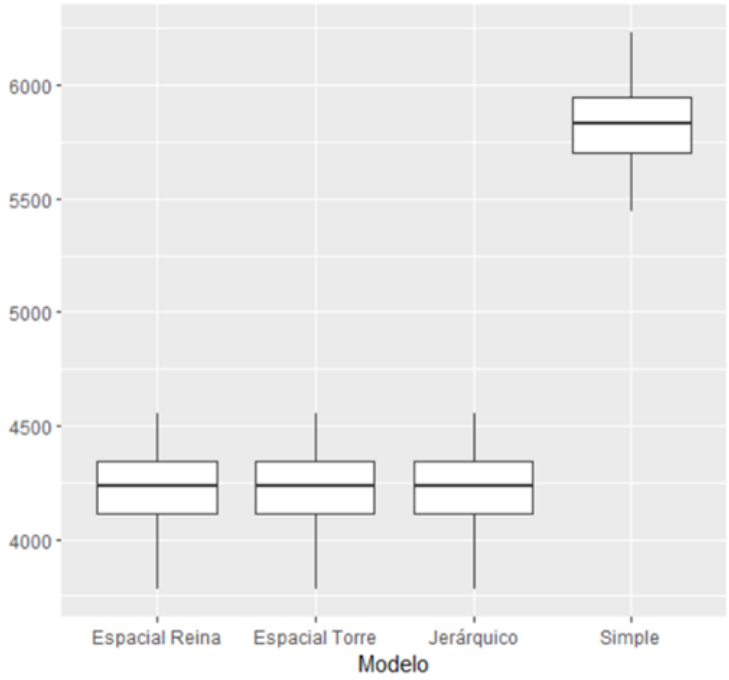


Figura 4.2.12. Criterio de información de deviancia según modelo estimado para el caso del coeficiente de regresión alto, unidad geográfica de distritos.



Al analizar las medidas de bondad de ajuste se puede determinar que en todos los casos la LML toma valores más alto para el modelo simple y se da una drástica disminución de este criterio si se compara con los demás modelos evaluados, esto para el coeficiente bajo, medio y alto. De manera general el modelo con el menor valor para la LML es el modelo jerárquico, Además el comportamiento de los modelos espaciales es muy similar entre sí, siendo que el modelo jerárquico utilizando el criterio de reina presenta un mejor ajuste si se compara con el criterio de torre.

Un comportamiento similar se presenta al analizar el WAIC y el DIC, existe una gran diferencia en estas medidas si se compara el modelo simple con los demás modelos. Sin embargo, la magnitud de estos indicadores para el modelo jerárquico y el de los modelos espaciales es básicamente la misma.

Cuadro 4.2.4 Medidas de bondad de ajuste según el modelo estimado, para la unidad geográfica de distritos.

Modelo	ECM	RMS	MAE
	Bajo, β=0.2		
Simple	0.0247	0.1572	0.0021
Jerárquico	0.0269	0.1643	0.0002
Torre	0.0271	0.1646	0.0001
Reina	0.0271	0.1646	0.0001
	Medio, β=0.35		
Simple	0.0244	0.1565	0.0042
Jerárquico	0.0273	0.1653	0.0008
Torre	0.0274	0.1656	0.0008
Reina	0.0274	0.1656	0.0008
	Alto, β=0.50		
Simple	0.0251	0.0022	0.0071
Jerárquico	0.027	0.0023	0.0018
Torre	0.0273	0.0019	0.0019
Reina	0.0271	0.0023	0.0018

El cuadro 4.2.4 muestra que el modelo simple tiene un mejor rendimiento respecto a los otros modelos si se compara el ECM o el RMS, sin embargo, se puede observar en el gráfico de los coeficientes de regresión estimados (Figuras 4.2.1, 4.2.2, 4.2.3) la presencia de valores extremos, por lo que estas dos medidas de bondad de ajuste pueden estar levemente sesgadas. Al analizar al MAE Los modelos jerárquicos y jerárquicos espaciales tienen un mejor desempeño respecto al simple.

El MAE tiene la ventaja de no verse afectada por la presencia de valores extremos, por lo que podría indicarse que estos tres modelos tienen un mejor rendimiento que el simple cuando se utiliza la unidad geográfica original.

Al comparar el modelo espacial con criterio de reina, suele tener un mejor desempeño que el modelo espacial con criterio de torre. Esto puede deberse a que utiliza más información de los vecinos que el caso del modelo de torre.

Cuadro 4.2.5 Precisión para el componente jerárquico y jerárquico espacial según el modelo estimado, para la unidad geográfica de distritos.

Modelo	Precisión jerárquica	Precisión espacial
Bajo, β=0.2		
Jerárquico	0.2736	-
Torre	0.2726	27333.25
Reina	0.2725	23068.25
Medio, β=0.35		
Jerárquico	0.2717	-
Torre	0.2701	50174.83
Reina	0.2697	23289.18
Alto, β=0.50		
Jerárquico	0.2698	-
Torre	0.2642	12307.37
Reina	0.2675	25236.13

Al analizar la precisión del componente jerárquico se tiene que, para los tres escenarios del coeficiente de regresión de exposición a plaguicidas, se obtiene la misma precisión para el modelo jerárquico, espacial con criterio de vecindad de torre y reina. Seguidamente cuando el coeficiente de regresión es bajo y medio la precisión del modelo estimado utilizando el criterio espacial de torre es más alta que el modelo con criterio de precisión de reina, patrón que se invierte cuando el coeficiente de regresión es alto, ya que la precisión es más alta, es decir, hay menos variabilidad en las estimaciones.

4.2.2 Resultados de la simulación de los conglomerados

Cuadro 4.2.6 Medidas de posición y variabilidad del coeficiente de regresión de exposición a plaguicidas según el modelo estimado, para la unidad geográfica resultante del análisis de conglomerados.

Modelo	Promedio	Cuartil 1	Mediana	Cuartil 3	Máximo	Mínimo	Desviación estándar
Bajo, β=0.2							
Simple	0.1672	-1.229	-0.046	1.485	6.935	-4.695	2.0274
Jerárquico	0.3865	-0.452	0.35	1.278	3.998	-2.631	1.2747
Torre	0.388	-0.452	0.35	1.279	4	-2.633	1.2761
Reina	0.3871	-0.453	0.349	1.279	4.003	-2.634	1.2752
Medio, β=0.35							
Simple	0.2639	-1.148	-0.064	1.683	6.732	-4.519	2.0297
Jerárquico	0.5864	-0.262	0.457	1.514	4.229	-2.968	1.2504
Torre	0.5858	-0.288	0.457	1.514	4.228	-2.971	1.2503
Reina	0.5868	-0.268	0.457	1.514	4.227	-2.972	1.2509
Alto, β=0.50							
Simple	0.3607	-0.995	0.086	1.528	6.634	-4.636	2.0101
Jerárquico	0.7168	-0.164	0.678	1.695	3.701	-2.728	1.3027
Torre	0.7163	-0.156	0.689	1.69	3.698	-2.747	1.3053
Reina	0.712	-0.165	0.677	1.696	3.701	-2.73	1.3032

Figura 4.2.13. Coeficiente de regresión según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica resultante del análisis de conglomerados.

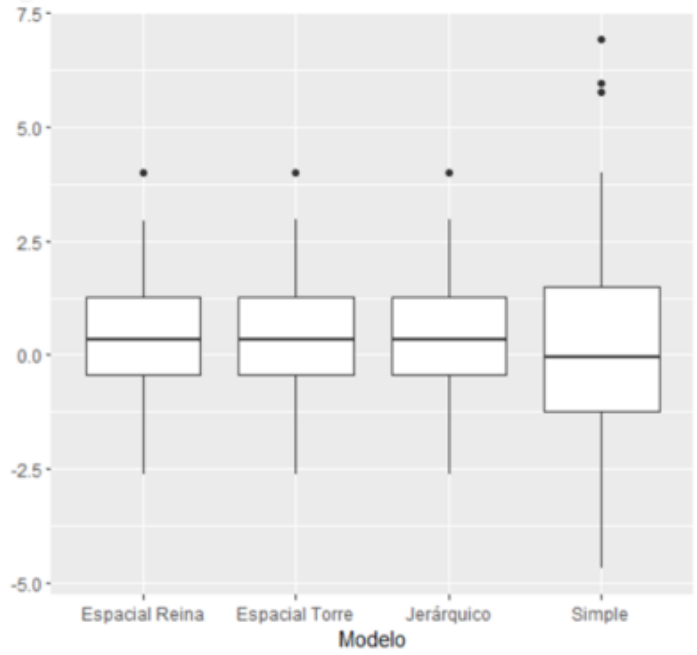


Figura 4.2.14. Coeficiente de regresión según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica resultante del análisis de conglomerados.

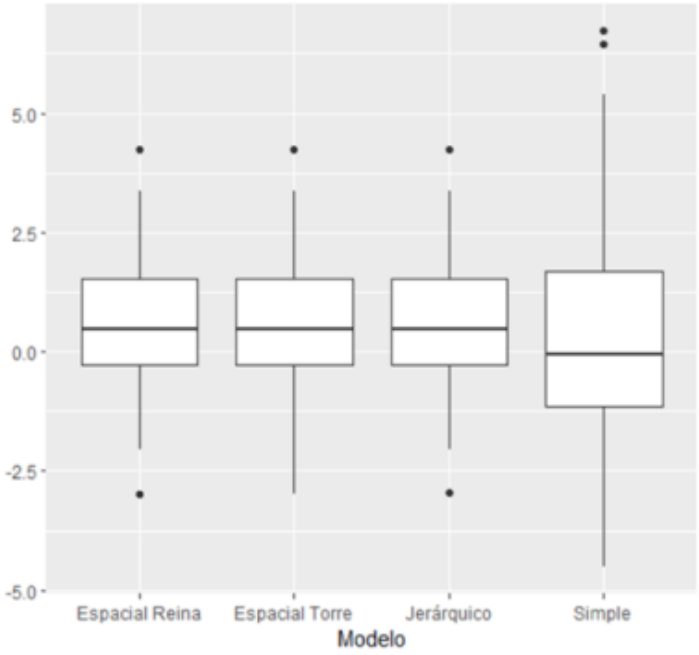
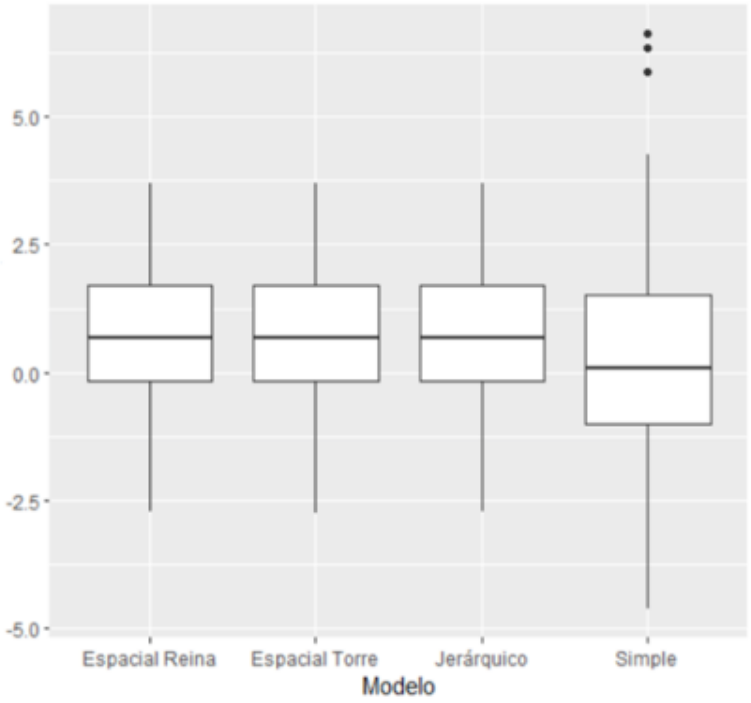


Figura 4.2.15. Coeficiente de regresión según modelo estimado para el caso del coeficiente de regresión alto, unidad geográfica resultante del análisis de conglomerados.



Al utilizar las unidades geográficas que se formaron a partir del análisis de conglomerados, se realizó la simulación de los datos para la estimación de los distintos modelos. En el caso del modelo simple se puede observar que el coeficiente de regresión promedio se subestima y, por el contrario, en los otros modelos el coeficiente asociado a los plaguicidas se sobreestima.

Analizando la mediana, se puede notar que es bastante similar al promedio en todos los modelos excepto para el modelo simple, donde la distribución de los coeficientes parece presentar una asimetría positiva.

Al analizar los valores máximos del coeficiente de regresión, así como el gráfico 4.2.13, 4.2.14, 4.2.15 se puede observar la presencia de valores extremos en todos los modelos estimados. Así como una mayor variabilidad (analizando el recorrido y el recorrido intercuartil) en el caso del modelo de regresión simple respecto a los demás casos.

Cuadro 4.2.7 Medidas de bondad de ajuste según el modelo estimado, para la unidad geográfica resultante del análisis de conglomerados.

Modelo	LML	WAIC	DIC
	Bajo, β=0.2		
Simple	-2735.2	5435.92	5435.83
Jerárquico	-2051.8	3908.66	3918.91
Torre	-2068.3	3908.62	3918.94
Reina	-2067.7	3908.63	3918.94
	Medio, β=0.35		
Simple	-2738.4	5442.22	5442.14
Jerárquico	-2054.1	3912.59	3922.7
Torre	-2070.6	3912.56	3922.72
Reina	-2069.9	3912.57	3922.73
	Alto, β=0.50		
Simple	-2743.2	5451.93	5451.86
Jerárquico	-2059.5	3924.24	3934.32
Torre	-2076.4	3924.2	3934.28
Reina	-2075.2	3924.23	3934.37

Figura 4.2.16. Log verosimilitud marginal según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica resultante del análisis de conglomerados.

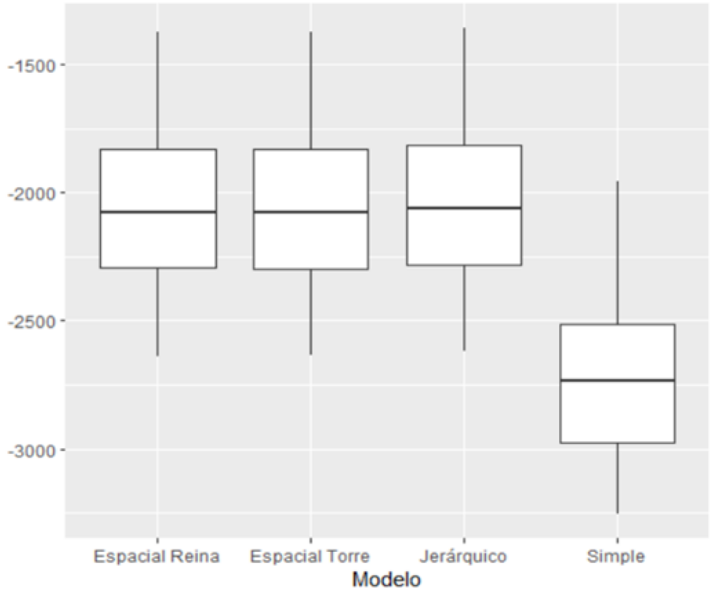


Figura 4.2.17. Criterio de información de Akaike Watanabe según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica resultante del análisis de conglomerados.

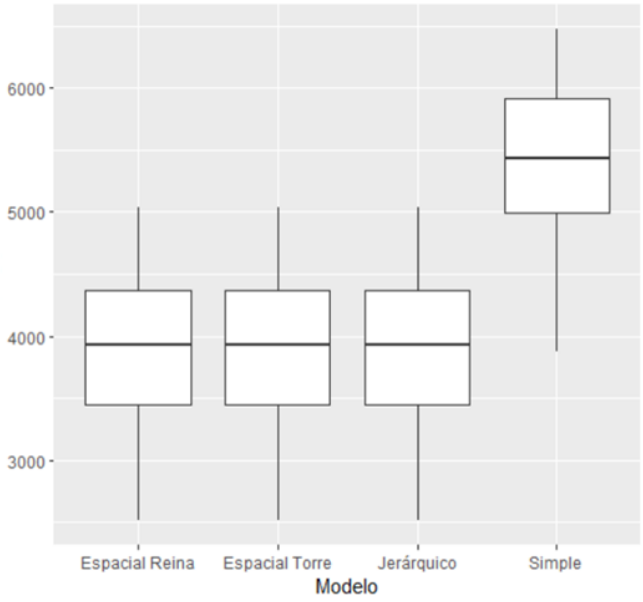


Figura 4.2.18. Criterio de información de deviancia según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica resultante del análisis de conglomerados.

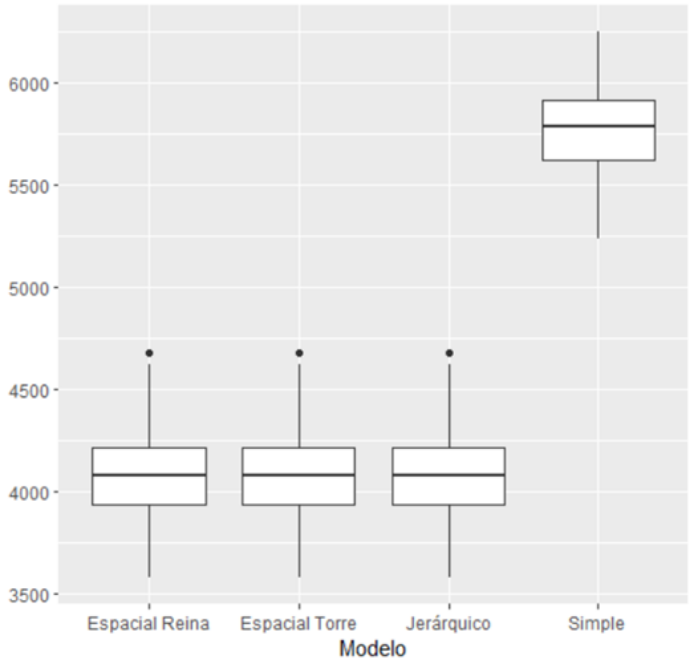


Figura 4.2.19. Log verosimilitud marginal según modelo estimado para el caso del coeficiente de regresión medio, unidad geográfica resultante del análisis de conglomerados.

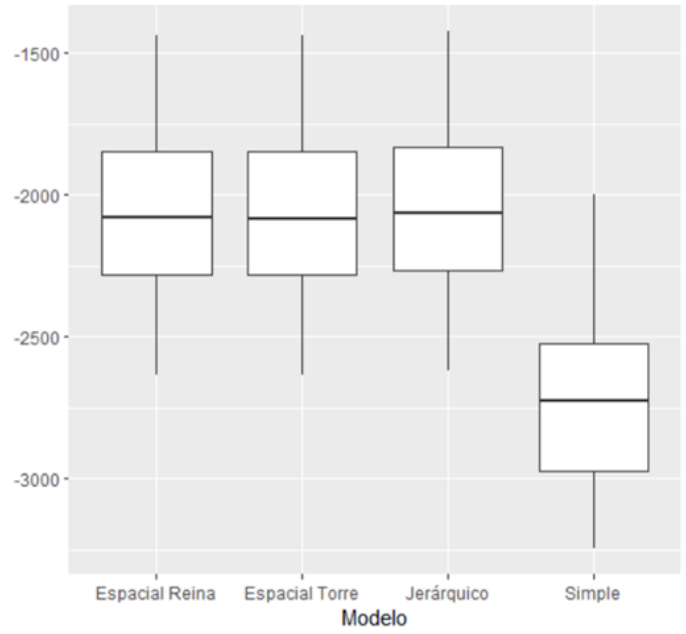


Figura 4.2.20. Criterio de información de Akaike Watanabe según modelo estimado para el caso del coeficiente de regresión medio, unidad geográfica resultante del análisis de conglomerados.

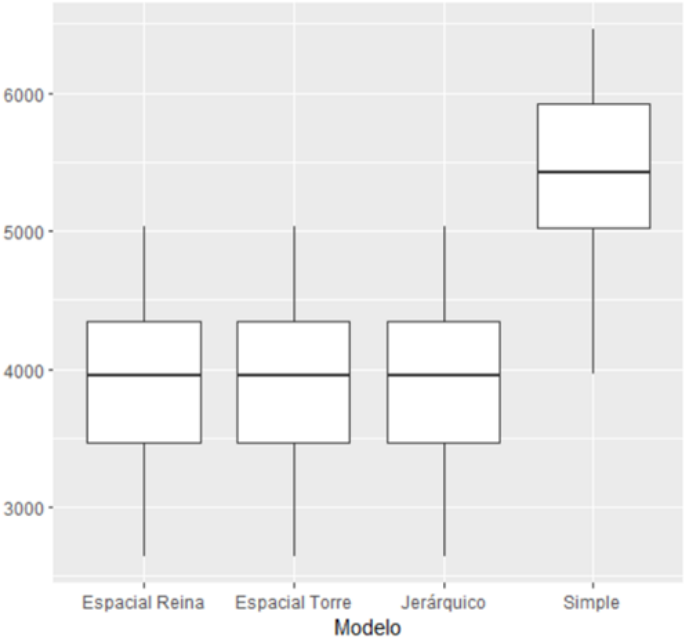


Figura 4.2.21. Criterio de información de deviancia según modelo estimado para el caso del coeficiente de regresión medio, unidad geográfica resultante del análisis de conglomerados.

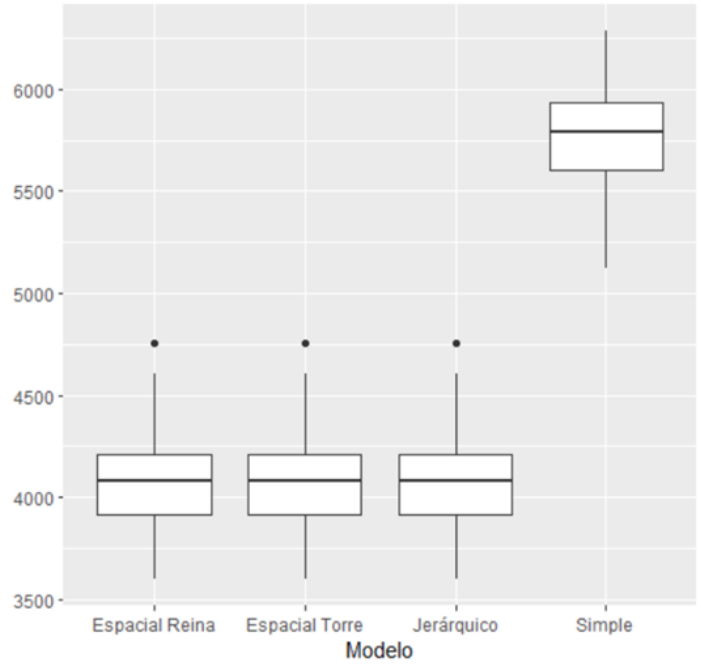


Figura 4.2.22. Log verosimilitud marginal según modelo estimado para el caso del coeficiente de regresión alto, unidad geográfica resultante del análisis de conglomerados.

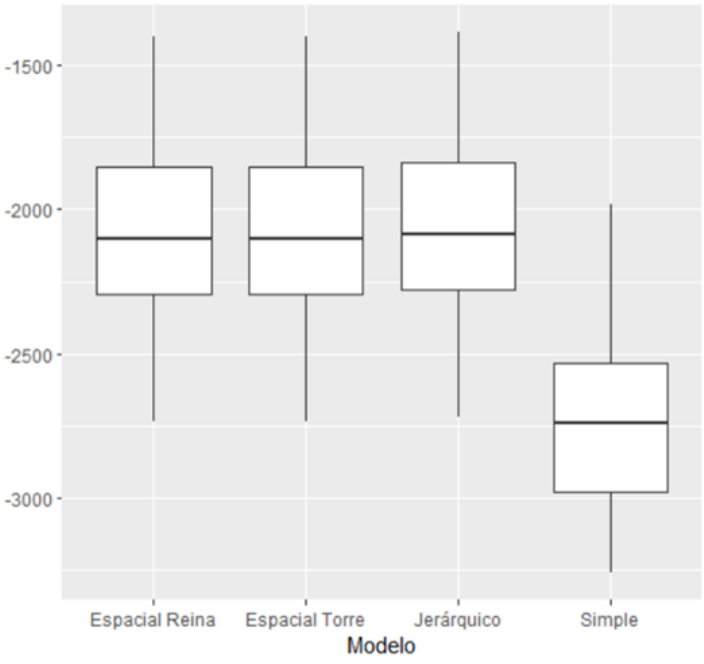


Figura 4.2.23. Criterio de información de Akaike Watanabe según modelo estimado para el caso del coeficiente de regresión alto, unidad geográfica resultante del análisis de conglomerados.

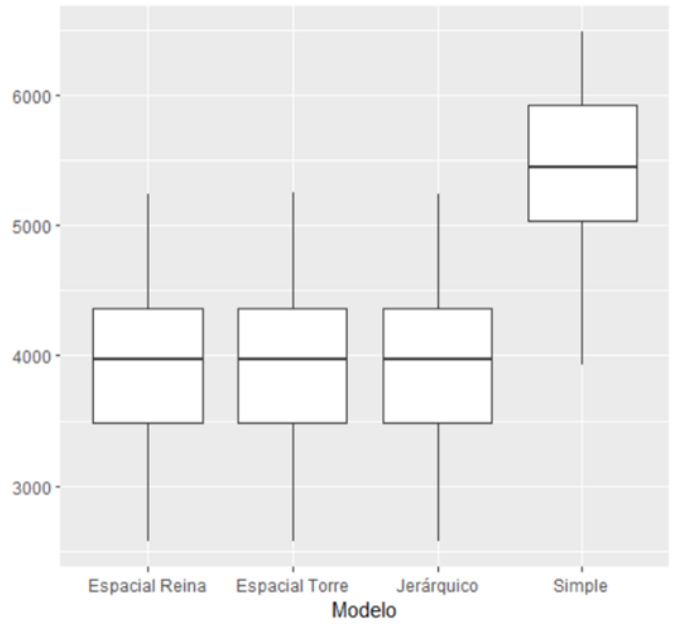
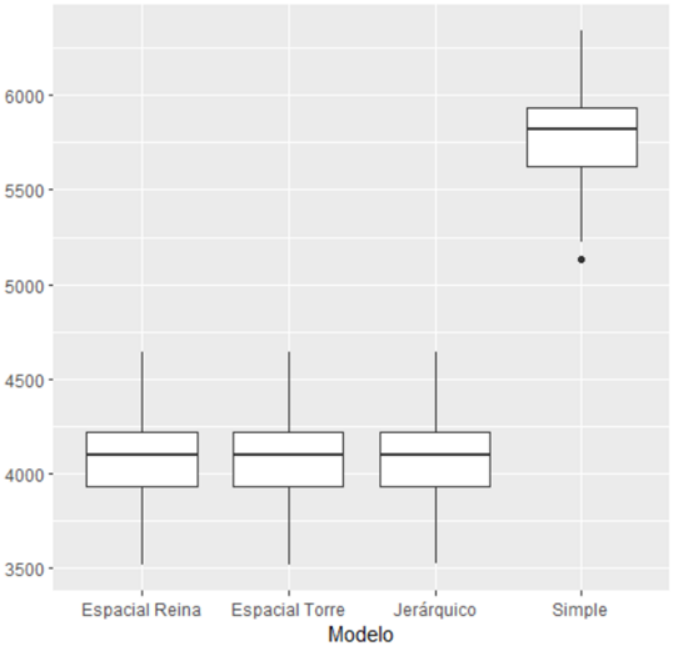


Figura 4.2.24. Criterio de información de deviancia según modelo estimado para el caso del coeficiente de regresión alto, unidad geográfica resultante del análisis de conglomerados.



De manera muy similar al caso evaluado cuando se utilizó la unidad geográfica original, existe una gran diferencia en las medidas de bondad de ajuste del modelo de regresión simple respecto a los demás modelos evaluados, donde el primero tiene el peor desempeño. Entre los modelos jerárquico y jerárquicos espaciales el comportamiento es bastante similar en términos del DIC y el WAIC, no tanto así para la LML.

Cuadro 4.2.8 Medidas de bondad de ajuste según el modelo estimado, para la unidad geográfica original.

Modelo	ECM	RMS	MAE
	Bajo, β=0.2		
Simple	0.1041	0.3226	0.0008
Jerárquico	0.042	0.205	0.0047
Torre	0.0421	0.2053	0.0047
Reina	0.042	0.2051	0.0047
	Medio, β=0.35		
Simple	0.1045	0.3233	0.0022
Jerárquico	0.041	0.2025	0.006
Torre	0.041	0.2025	0.006
Reina	0.041	0.2026	0.006
	Alto, β=0.50		
Simple	0.1028	0.0045	0.0035
Jerárquico	0.0441	0.003	0.0055
Torre	0.0443	0.0055	0.0055
Reina	0.0442	0.003	0.0055

Al analizar las medidas de bondad de desempeño del ECM y RMS se puede determinar que el modelo con peor rendimiento es el simple, mientras que los modelos jerárquico y jerárquico espacial tienen comportamientos bastante similares. Siendo que el modelo jerárquico espacial de reina en ocasiones tiene mejor rendimiento respecto al modelo cuyo criterio de vecindad es el de torre.

Cuadro 4.2.9 Precisión para el componente jerárquico y jerárquico espacial según el modelo estimado, para la unidad geográfica original.

Modelo	Precisión jerárquica	Precisión espacial
Bajo, β=0.2		
Jerárquico	0.2639	-
Torre	0.2745	22285.5
Reina	0.2789	22699.22
Medio, β=0.35		
Jerárquico	0.2651	-
Torre	0.2687	23072.74
Reina	0.2726	26236.21
Alto, β=0.50		
Jerárquico	0.2661	-
Torre	0.2680	1302.64
Reina	0.2731	119938.9

En este caso se puede observar que la precisión del componente jerárquico para los nueve escenarios evaluados es muy simular, es decir, para los tres tipos de modelos estimados bajo los tres tipos de coeficientes de regresión. Sin embargo, el modelo de reina tuvo mejores valores de precisión respecto al modelo de torre en los tres casos de los coeficientes de regresión. Indicando que para la unidad geográfica resultante de los conglomerados dentro de las provincias el modelo de reina genera una menor variabilidad en la estimación de la variabilidad del componente espacial.

4.5.3 Resultados simulación agrupación de distritos dentro de las provincias

Cuadro 4.2.10 Medidas de posición y variabilidad del coeficiente de regresión de exposición a plaguicidas según el modelo estimado, para la unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

Modelo	Promedio	Cuartil 1	Mediana	Cuartil 3	Máximo	Mínimo	Desviación estándar
Bajo, β=0.2							
Simple	-0.0319	-0.458	-0.061	0.567	1.933	-2.435	0.8596
Jerárquico	0.1231	-0.45	0.211	0.614	2.334	-2.134	0.8833
Torre	0.1236	-0.45	0.21	0.614	2.333	-2.134	0.8832
Reina	0.1237	-0.45	0.21	0.614	2.334	-2.135	0.8836
Medio, β=0.35							
Simple	0.0873	-0.376	0.098	0.691	2.011	-1.917	0.8614
Jerárquico	0.3235	-0.334	0.356	0.954	2.62	-1.817	0.9342
Torre	0.3235	-0.334	0.37	0.956	2.621	-1.817	0.9346
Reina	0.3234	-0.333	0.37	0.952	2.62	-1.815	0.9343
Alto, β=0.50							
Simple	0.1869	-0.194	0.164	0.823	2	-1.789	0.8386
Jerárquico	0.4644	-0.114	0.506	0.999	2.698	-1.41	0.8414
Torre	0.4692	-0.114	0.484	1.012	2.699	-1.408	0.8445
Reina	0.4664	-0.114	0.495	1.021	2.7	-1.41	0.8431

Figura 4.2.25. Coeficiente de regresión según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica resultado de la agrupación de distritos dentro de las provincias

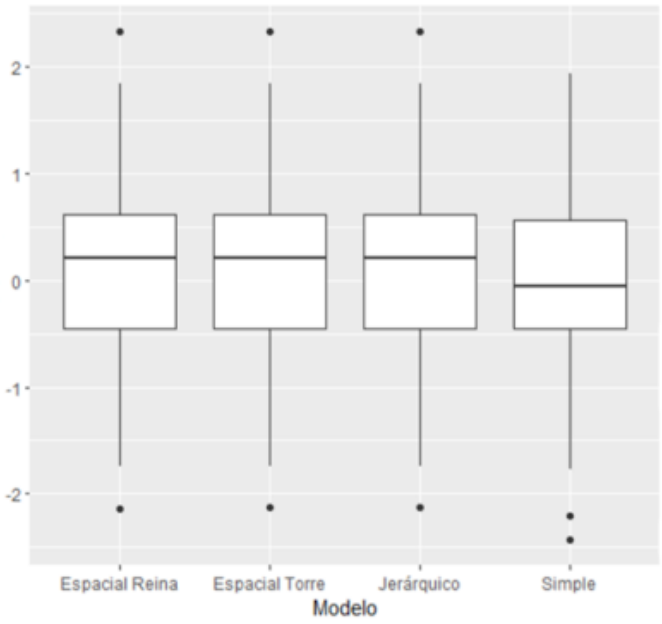


Figura 4.2.26. Coeficiente de regresión según modelo estimado para el caso del coeficiente de regresión medio, unidad geográfica resultado de la agrupación de distritos dentro de las provincias

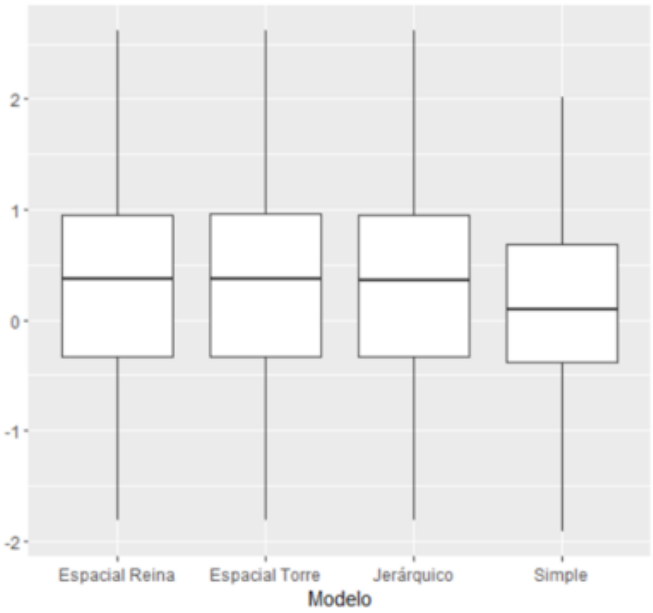
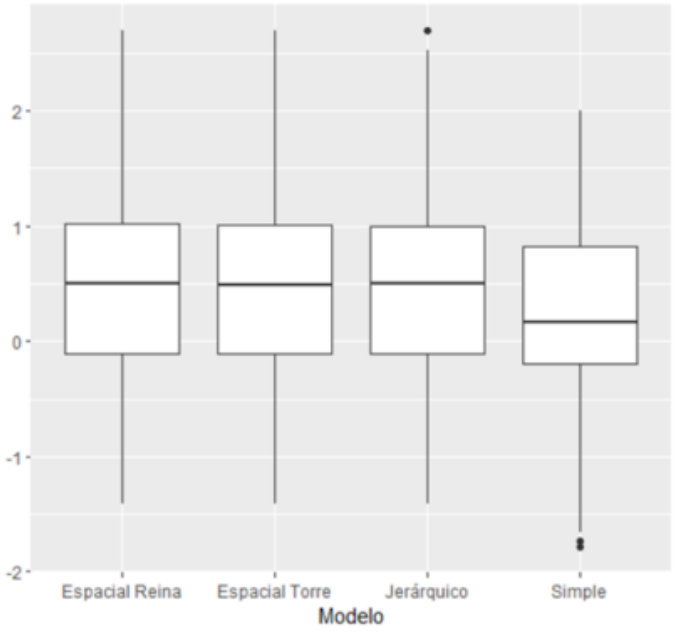


Figura 4.2.27. Coeficiente de regresión según modelo estimado para el caso del coeficiente de regresión alto, unidad geográfica resultado de la agrupación de distritos dentro de las provincias



A partir de los resultados de los distintos modelos de regresión estimados, se tiene que al evaluar el coeficiente promedio en todos los casos se subestima el coeficiente de regresión, sin embargo, el modelo simple es el que más se aleja del valor con el que se realizó la estimación. Por otra parte, al observar a la mediana del coeficiente, excepto para el modelo simple, el coeficiente estimado es muy similar al utilizado en la simulación.

A partir de los resultados obtenidos se tiene que, en general el modelo simple es el que presenta menor variabilidad si se analiza el recorrido, el recorrido intercuartílico, así como la desviación estándar. En el caso de la simulación del coeficiente de regresión bajo y alto, al evaluar los boxplot correspondientes, hay presencia de valores extremo, lo cual afecta al recorrido y la desviación estándar.

Cuadro 4.2.11 Medidas de bondad de ajuste según el modelo estimado, para la unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

Modelo	LML	WAIC	DIC
	Bajo, β=0.2		
Simple	-2911.9	5788.57	5788.59
Jerárquico	-2184	4071.65	4090.73
Torre	-2208.4	4071.62	4090.71
Reina	-2207	4071.59	4090.69
	Medio, β=0.35		
Simple	-2910.2	5785.08	5785.11
Jerárquico	-2187.7	4079.74	4098.35
Torre	-2211.8	4079.67	4098.31
Reina	-2210.9	4079.67	4098.32
	Alto, β=0.50		
Simple	-2917.6	5799.82	5799.85
Jerárquico	-2189.2	4082.56	4101.18
Torre	-2213.5	4082.46	4101.18
Reina	-2212.2	4082.5	4101.26

Figura 4.2.28. Log verosimilitud marginal según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

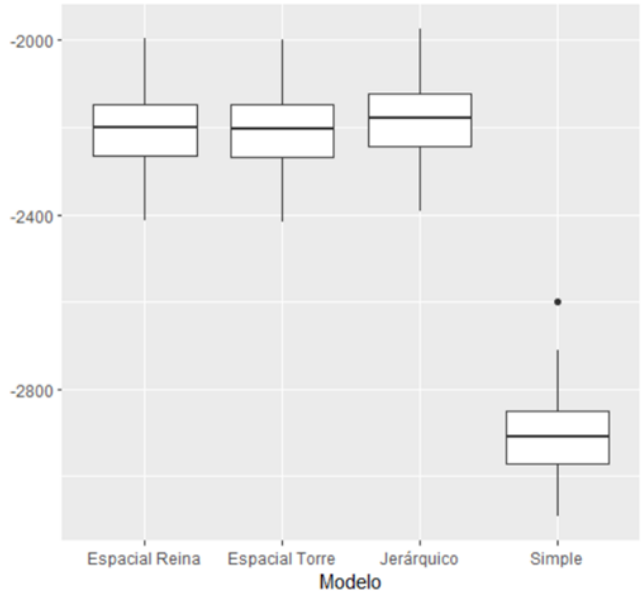


Figura 4.2.29. Criterio de información de Akaike Watanabe según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

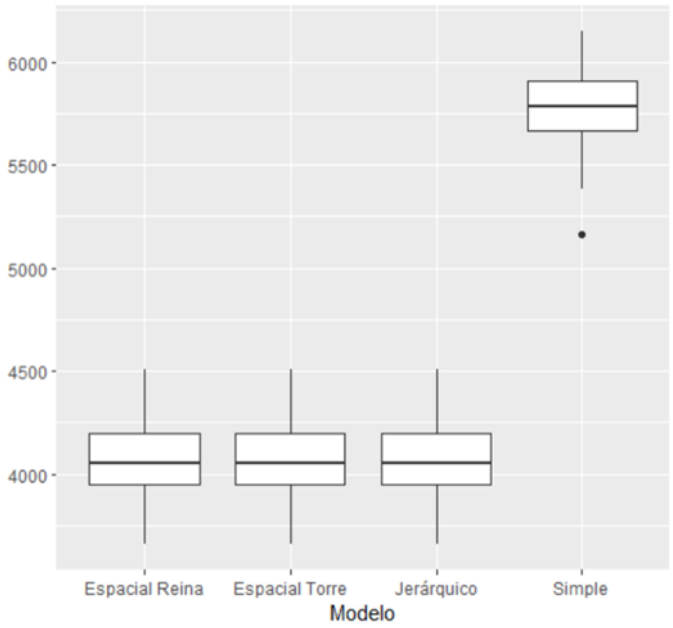


Figura 4.2.30. Criterio de información de devianza según modelo estimado para el caso del coeficiente de regresión bajo, unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

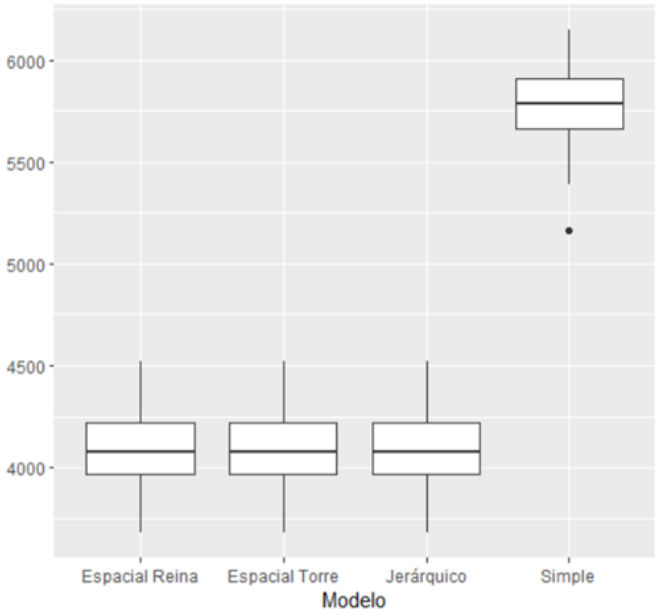


Figura 4.2.31. Log verosimilitud marginal según modelo estimado para el caso del coeficiente de regresión medio, unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

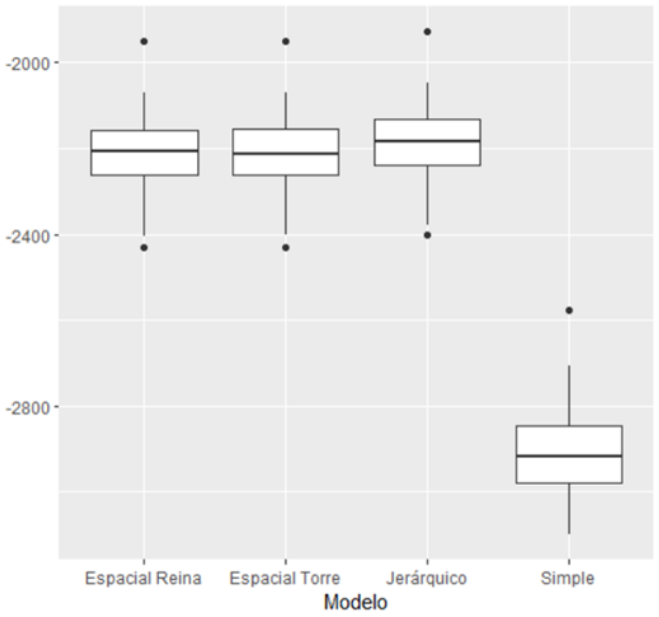


Figura 4.2.32. Criterio de información de Akaike Watanabe según modelo estimado para el caso del coeficiente de regresión medio, unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

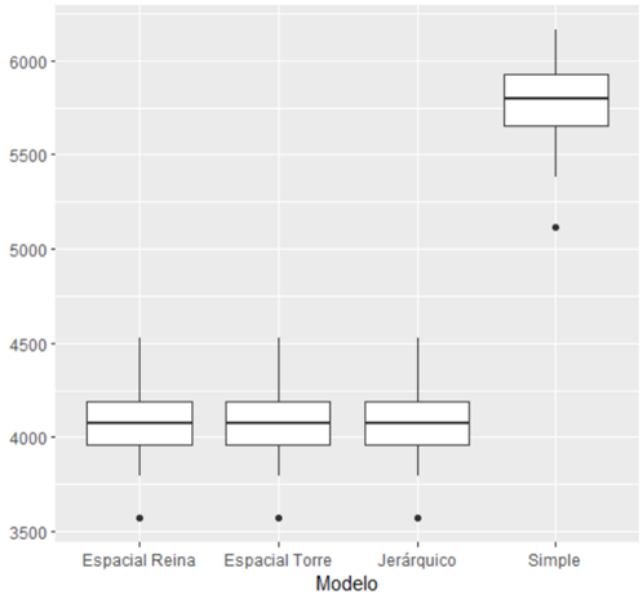


Figura 4.2.33. Criterio de información de devianza según modelo estimado para el caso del coeficiente de regresión medio, unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

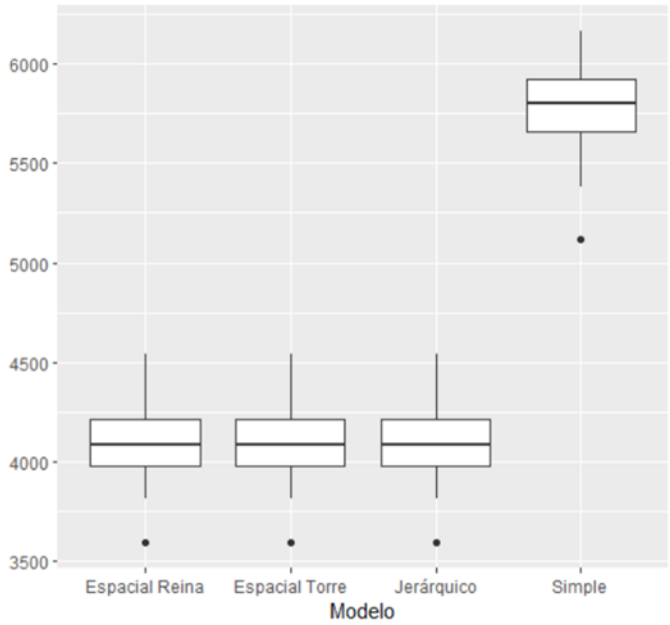


Figura 4.2.34. Log verosimilitud marginal según modelo estimado para el caso del coeficiente de regresión alto, unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

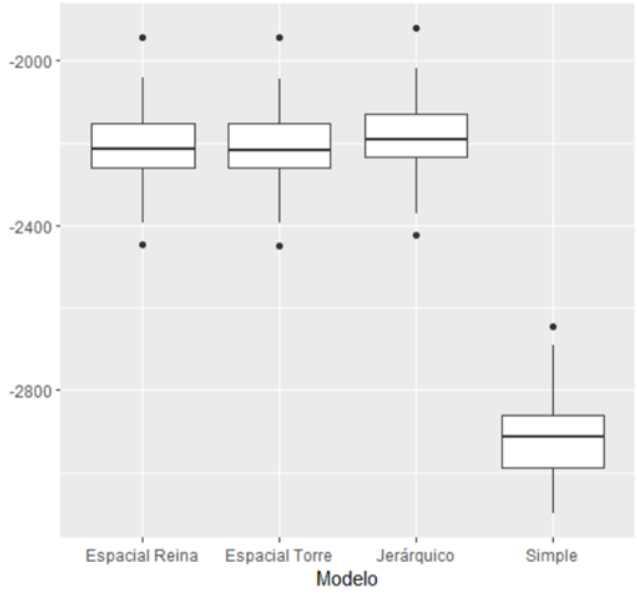


Figura 4.2.35. Criterio de información de Akaike Watanabe según modelo estimado para el caso del coeficiente de regresión alto, unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

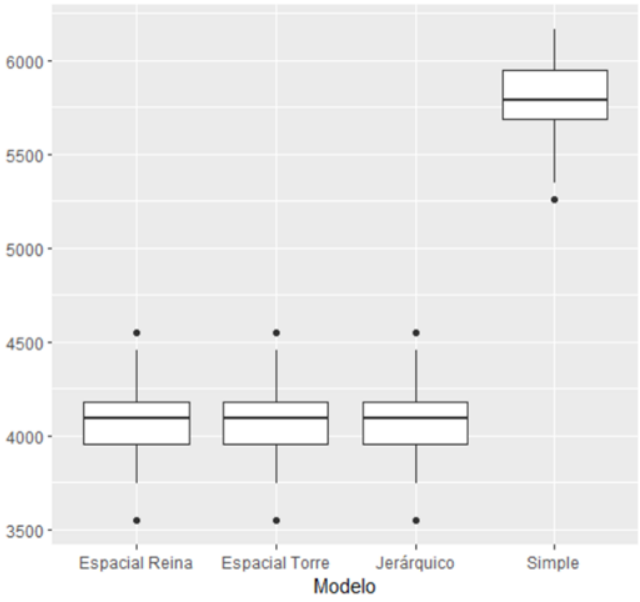
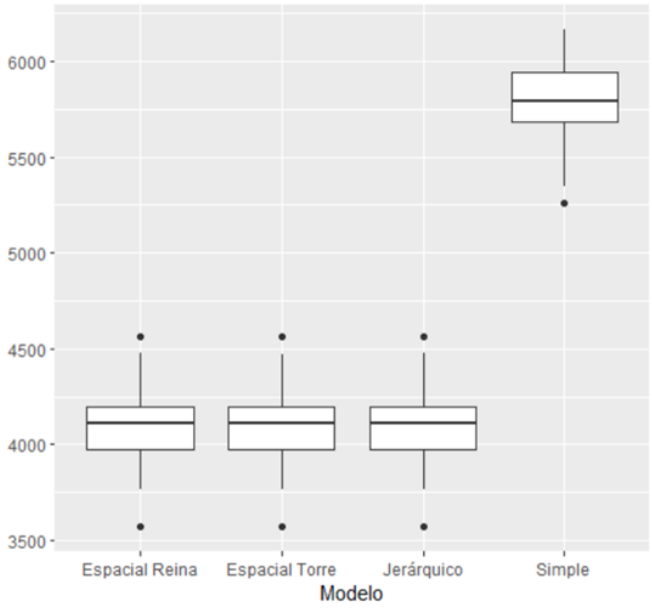


Figura 4.2.36. Criterio de información de devianza según modelo estimado para el caso del coeficiente de regresión alto, unidad geográfica resultado de la agrupación de distritos dentro de las provincias.



Para el coeficiente bajo, medio y alto se puede advertir como la LML, WAIC y DIC mejoran al tomar en cuenta la estructura jerárquica de los datos. En términos del LML se tiene que el modelo jerárquico tiene un mejor desempeño que sus alternativas espaciales, sin embargo, al comprar el WAIC y el DIC se puede determinar que los modelos se comportan de manera muy similar.

Cuadro 4.2.12 Medidas de bondad de ajuste según el modelo estimado, para la unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

Modelo	ECM	RMS	MAE
	Bajo, β=0.2		
Simple	0.02008	0.1417	0.0059
Jerárquico	0.01991	0.1411	0.0019
Torre	0.01991	0.141	0.0019
Reina	0.01992	0.1411	0.0019
	Medio, β=0.35		
Simple	0.0205	0.1433	0.0067
Jerárquico	0.0221	0.1487	0.00067
Torre	0.0221	0.1488	0.00067
Reina	0.0221	0.1487	0.00067
	Alto, β=0.50		
Simple	0.0203	0.002	0.0079
Jerárquico	0.0179	0.0019	0.0009
Torre	0.018	0.0007	0.0007
Reina	0.018	0.0019	0.0008

Respecto al error cuadrático medio, para el coeficiente de regresión bajo y alto indican que los modelos jerárquicos y jerárquicos espaciales recuperan de mejor manera el coeficiente de regresión utilizado en la simulación. En el caso del coeficiente medio, tal medida de bondad de ajuste indica una mejoría para el coeficiente simple.

Seguidamente la raíz del cuadrado medio del error, así como el error absoluto medio para los modelos jerárquico y espaciales indican una mejor recuperación respecto al modelo simple. En el caso del coeficiente de regresión alto, se tiene que la mejor recuperación se da con el modelo espacial con criterio de torre.

Cuando se analizó el error absoluto medio para el coeficiente de regresión medio, se obtuvo que los mejores modelos eran los jerárquico y espaciales.

Cuadro 4.2.13 Precisión para el componente jerárquico y jerárquico espacial según el modelo estimado, para la unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

Modelo	Precisión jerárquica	Precisión espacial
Bajo, β=0.2		
Jerárquico	0.2613	-
Torre	0.2635	22285.5
Reina	0.2634	22699.2
Medio, β=0.35		
Jerárquico	0.2679	-
Torre	197.0304	23072.7
Reina	200.4339	26236.2
Alto, β=0.50		
Jerárquico	0.263	-
Torre	171.6439	1302.635
Reina	182.3558	119938.9

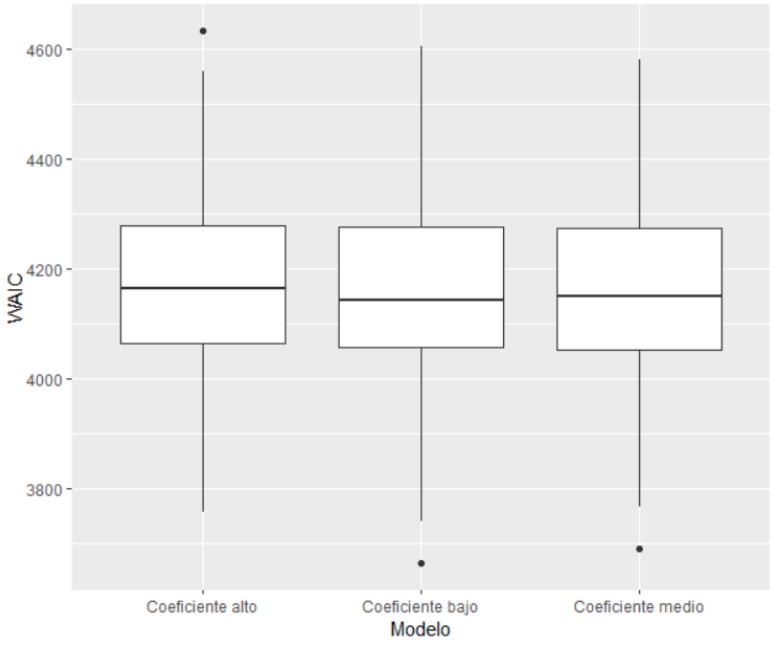
En el caso del modelo estimado cuando el coeficiente de regresión es bajo, los tres modelos tienen la misma precisión para el componente jerárquico, y al comparar los modelos con componentes espaciales la precisión del modelo reina es superior al caso de torre. En el caso del coeficiente medio y alto se tiene una mejoría en la precisión de la estimación del componente jerárquico y de igual manera la precisión del modelo espacial de reina es mejor que en el caso de torre.

Debido a la similitud en los comportamientos de los escenarios con los simulados utilizando distribuciones previas para los componentes jerárquico y jerárquico espaciales, los resultados se encuentran en anexos (ver anexos 6).

4.5.4 Comparación de resultados del modelo jerárquico espacial de reina para los distintos niveles de magnitud del coeficiente de exposición a plaguicidas y agrupamiento espacial.

En las siguientes figuras se compara al modelo jerárquico espacial con criterio de reina para los diferentes escenarios analizados en la simulación. Se presentará como varía el WAIC para los tres escenarios de magnitud del coeficiente de exposición a plaguicidas (bajo, medio y alto), esto para las tres agrupaciones propuestas.

Figura 4.2.37. Criterio de información de Akaike Watanabe según magnitud del coeficiente estimado, para la unidad geográfica original.



En el caso de la unidad geográfica original y de la agrupación de distritos dentro de provincias se puede notar la presencia de valores extremos en los valores de la log verosimilitud marginal. Se puede observar además que en el caso del coeficiente alto el valor del WAIC suele ser para todos los casos analizados levemente mayor al caso del coeficiente bajo y medio, teniendo que para el coeficiente bajo el valor del WAIC suele tener el menor valor.

Figura 4.2.38. Criterio de información de Akaike Watanabe n según magnitud del coeficiente estimado, para la unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

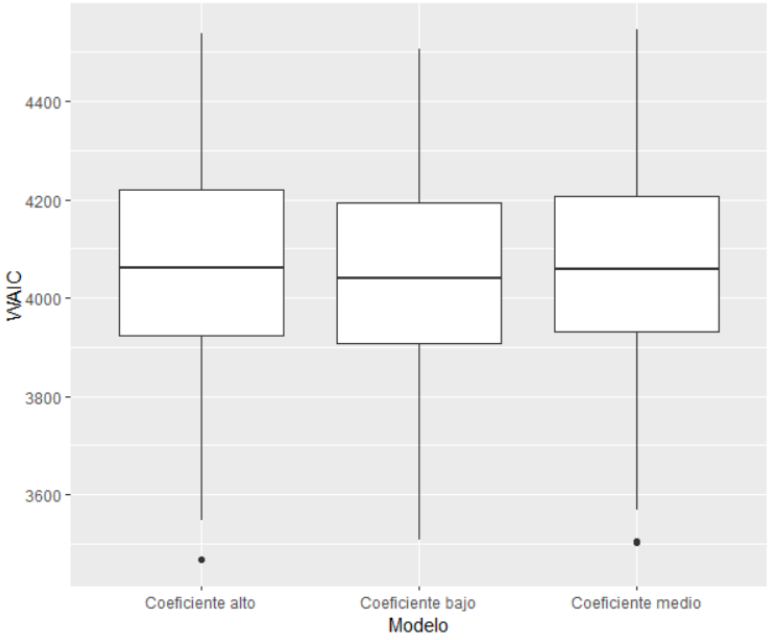
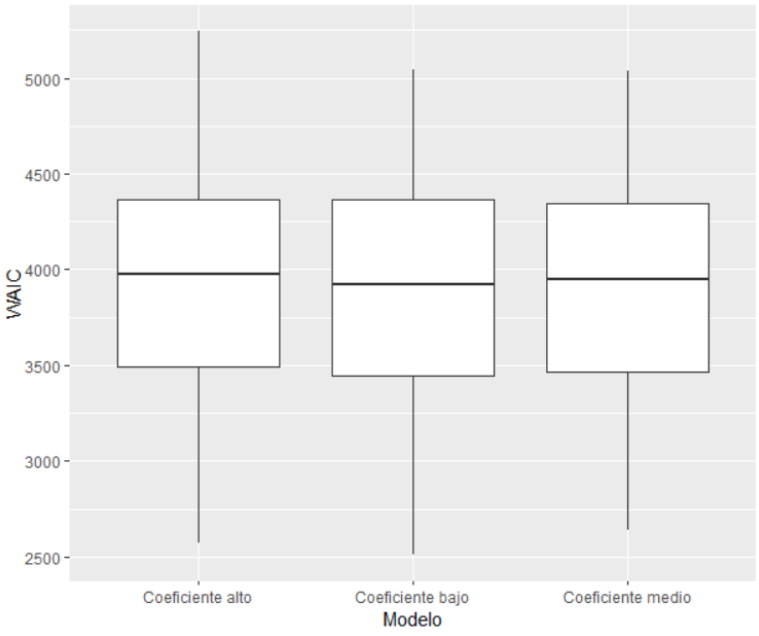
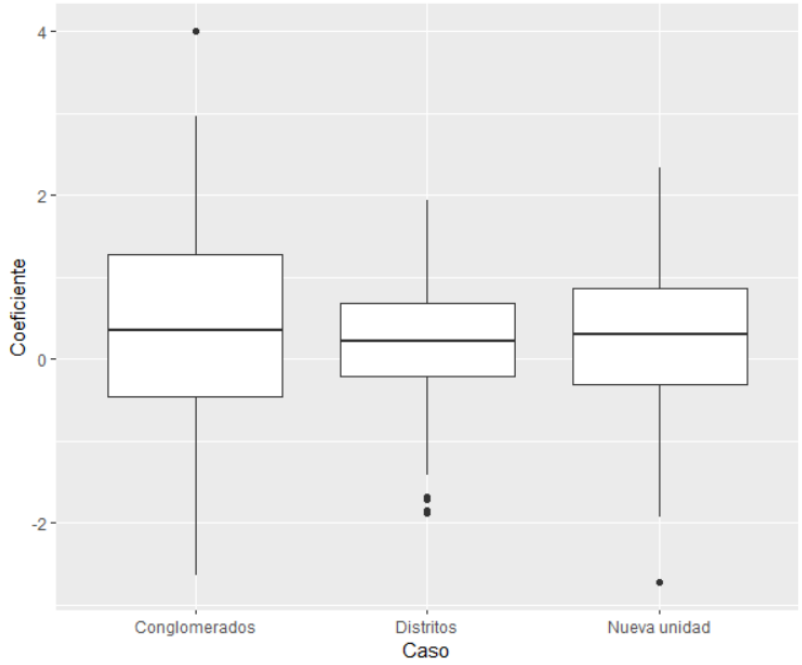


Figura 4.2.39. Criterio de información de Akaike Watanabe según magnitud del coeficiente estimado, unidad geográfica resultante del análisis de conglomerados.



Seguidamente se presentan los gráficos asociados a los coeficientes de regresión para las tres magnitudes del coeficiente según la unidad geográfica utilizada para la estimación del modelo.

Figura 4.2.40. Coeficiente de regresión magnitud baja según unidad geográfica.



Al analizar los resultados para las tres agrupaciones geográficas se puede denotar que el caso de la unidad de conglomerados presenta más variabilidad para el coeficiente de regresión asociado a la exposición de plaguicidas, pero en todos los casos la mediana toma valores muy similares entre sí. Además, se puede evidenciar la presencia de valores extremos para los coeficientes en casi todos los escenarios

Figura 4.2.41. Coeficiente de regresión magnitud media según unidad geográfica.

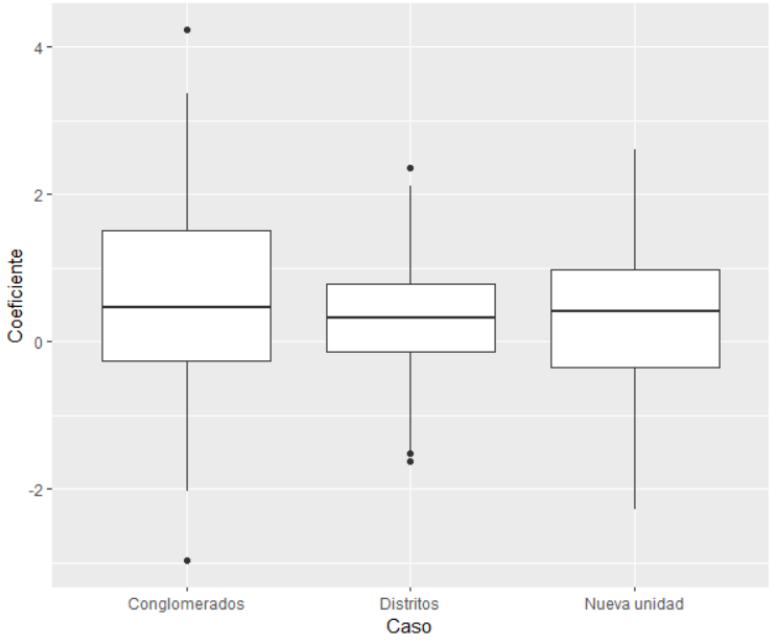
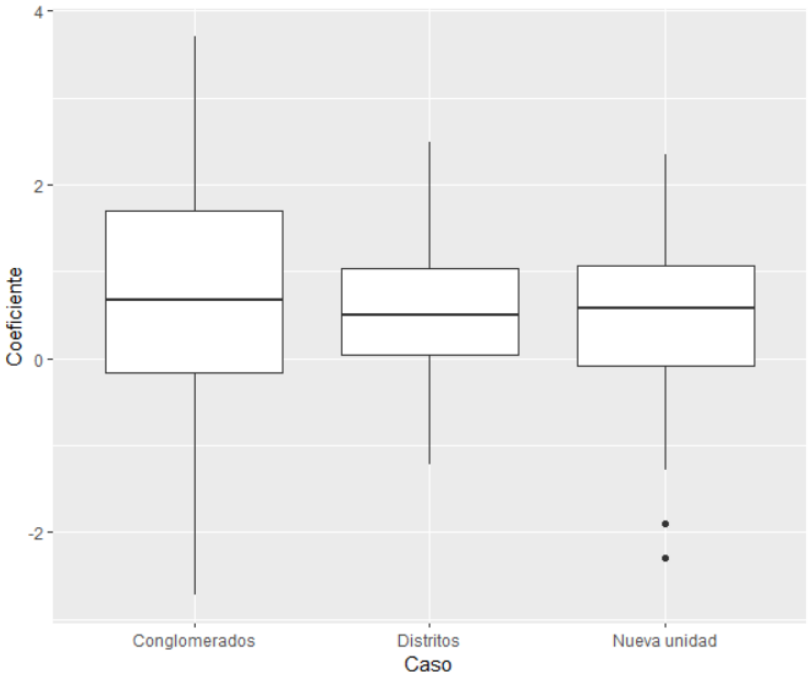


Figura 4.2.42. Coeficiente de regresión magnitud alta según unidad geográfica.



CAPÍTULO 5: Discusión y Conclusiones

El objetivo general de esta tesis consistió en identificar un método unificado para el análisis de datos con información jerárquica, espacial y ecológica sobre cáncer de mama y próstata a partir de un modelo lineal jerárquico binomial en Costa Rica, basado en los resultados de la simulación. Para ello en una primera instancia, se realizó un estudio empírico de los datos, para conocer la relación entre las variables a utilizar en el modelado de estos y para la construcción del índice de exposición de plaguicidas.

Es importante destacar que, al comparar los coeficientes de regresión asociados a la exposición a plaguicidas para el modelo jerárquico espacial con criterio de reina, los modelos resultantes de la unidad geográfica de conglomerados presentaron mayor variabilidad que el caso de la agrupación dentro de distritos y de la unidad de distritos.

Sin embargo, de la estimación de los modelos bajo las tres agrupaciones geográficas, se obtuvo que tanto el modelo con la agrupación de los distritos dentro de las provincias, así como el caso cuando se realizó la agrupación de los distritos dentro de provincias usando el análisis de conglomerados tuvieron un mejor desempeño que el agrupamiento original por distritos, esto se pudo determinar mediante la log verosimilitud marginal, WAIC, DIC, así como una mayor precisión, en especial para la primera agrupación.

Se generaron distintos escenarios para el diseño del estudio de simulación que permitiera identificar el modelo más adecuado, la simulación se realizó a partir de los resultados de los obtenidos en el análisis empírico de los datos.

Para la determinación de los escenarios, se varió la unidad geográfica que agrupa los datos en la unidad geográfica original correspondiente a los distritos, agrupamiento de los distritos dentro de las provincias respectivas y finalmente a partir del análisis de conglomerados se agruparon los distritos dentro de las provincias que tuvieran un comportamiento muy similar entre ellos y que tuvieran plausibilidad geográfica, es decir, que compartieran una frontera geográfica. Además, para la generación de escenarios se varió el valor del coeficiente de regresión asociado al índice de exposición a plaguicidas.

Los valores se establecieron a partir de los resultados del estudio empírico. Generándose finalmente nueve escenarios a evaluar y cada uno de ellos evaluando 4 modelos distintos.

A partir del estudio de simulación se comprobó que, de manera general, el modelo simple muestra ajuste el ajuste más bajo para estimar, es decir, el modelo lineal generalizado binomial que no considera la estructura jerárquica ni la espacial. Se determinó a partir de medidas de posición como el promedio y mediana que subestimaban el valor del coeficiente de plaguicidas utilizado para realizar la simulación de los datos.

Además, generaba la presencia de valores extremos, su valor para la log verosimilitud marginal en todos los casos era el más alto, al compararlo con los otros modelos estimados. De igual manera al analizar medidas como el error cuadrático medio, la raíz del error cuadrático medio, así como el error absoluto medio, de manera general era el modelo que presentaba peores medidas de desempeño.

Seguidamente, se evidenció que entre el modelo jerárquico y los modelos espaciales el comportamiento fue muy similar. En términos de la log verosimilitud marginal, el modelo jerárquico era levemente mejor que las versiones espaciales, sin embargo, al comparar el criterio de información de Akaike Watanabe y el criterio de información de la devianza se obtuvo que el comportamiento para los tres modelos era muy similar.

Es importante resaltar que a pesar de la similitud que existe entre los modelos jerárquico y jerárquico espacial, se debe de tener en cuenta el objetivo por el cual se está incluyendo este término dentro del modelo de regresión. Anselin (1998) comenta que se puede considerar la dependencia espacial ya que se tiene interés en la evaluación y cuantificación de la interacción espacial, o bien porque se desea corregir el potencial sesgo cuando se introducen datos con esta estructura.

En este caso específico, dado que se quería controlar el posible sesgo en términos de la estimación del error ante la presencia de datos especialmente correlacionados, se considera que los modelos jerárquicos espaciales, a pesar de su similitud con el modelo jerárquico, representan un mejor acercamiento para los datos, aunque la estimación de este modelo implica una mayor complejidad.

Respecto a las tres medidas de bondad de ajuste, en algunos casos se indica que para las estimaciones bayesianas el WAIC es mejor que el AIC y el DIC ya que este utiliza densidad posterior completa. Gelman (2015) explica que, a pesar de la dificultad asociada al cálculo del WAIC, este es preferible a los otros criterios antes mencionados, ya que, algunas derivaciones del AIC y DIC asumen que los residuales son independientes dada la estimación puntual del parámetro.

Respecto a las medidas de bondad de ajuste, en algunos casos se indica que para las estimaciones bayesianas el WAIC es mejor que el AIC y el DIC ya que este utiliza densidad posterior completa. Gelman (2015) explica que, a pesar de la dificultad asociada al cálculo del WAIC, este es preferible a los otros criterios antes mencionados, ya que, algunas derivaciones del AIC y DIC asumen que los residuales son independientes dada la estimación puntual del parámetro.

Cuando se analizaron los resultados de los modelos estimados utilizando los agrupamientos que se generaron de las unidades geográficas originales, se determinó a partir de los tres criterios de bondad de ajuste analizados y de las medidas de exactitud que los modelos jerárquico y jerárquico espaciales tienen un mejor desempeño que el modelo denominado simple.

Al evaluar las medidas de exactitud como el error cuadrático medio, la raíz del error cuadrático medio y el error absoluto medio para el agrupamiento original de las unidades geográficas se tuvo que con el ECM y RMS el mejor modelo a seleccionar era el simple, mientras que con el MAE se debía de seleccionar a los modelos jerárquico y jerárquico espaciales.

Sin embargo, para los escenarios que contemplaron la agrupación de las unidades geográficas, las tres medidas de exactitud seleccionadas indicaban que el peor modelo para recuperar el coeficiente de regresión asociado a la exposición a plaguicidas era el modelo simple. Mientras que los otros modelos lo recuperaban de mejor manera.

Seguidamente al comparar la LML, DIC y WAIC para los tres tipos de unidades geográficas, se determinó que el mejor agrupamiento es el resultante del análisis de

conglomerados, ya que, para los cuatro tipos de modelos analizados presentó valores inferiores en los tres indicadores, indicando de esta manera un mejor ajuste.

Asimismo, al comparar los modelos espaciales, el agrupamiento de tipo reina genera una leve mejora en medidas de bondad de ajuste, así como en las medidas de exactitud respecto al modelo cuyo criterio es de torre. Por el comportamiento descrito respecto a la bondad de ajuste, se considera que el mejor modelado para el conjunto de datos analizado es con el modelo espacial con criterio de vecindad de reina cuando se utiliza la nueva unidad geográfica resultante del agrupamiento de distritos dentro de las provincias haciendo uso de los conglomerados.

Al comparar los resultados de los modelos estimados para diferentes unidades geográficas, se observa que el coeficiente de regresión asociado a la tasa de exposición a plaguicidas muestra un patrón consistente a los mismos modelos utilizando distribuciones previas. Sin embargo, los modelos basados en la unidad original de distritos y la nueva unidad geográfica presentan un coeficiente negativo, lo cual es inesperado dado que se esperaría una relación positiva con la letalidad del cáncer. Este resultado podría atribuirse a que estas agrupaciones no capturan adecuadamente la relación esperada, a diferencia del modelo basado en conglomerados, donde se espera que la inclusión de determinantes sociales de la salud resulte en una relación positiva e importante entre la prevalencia del cáncer y su letalidad.

Es importante destacar que, si bien estos determinantes de la salud no se incluyeron en la presente investigación, esto se debió a que el objetivo principal de la misma era establecer un acercamiento inicial para determinar el mejor modelo para analizar estos datos teniendo en cuenta su estructura espacial y jerárquica. Sin embargo, se reconoce la necesidad profundizar en el análisis, incluyendo una gama más amplia de variables relacionadas con la salud, así como un modelo que se ajuste a la estructura de los datos.

De igual manera, es importante destacar que el cáncer es una enfermedad que tiene un alto costo para la salud del país y por lo tanto el análisis de sus determinantes cobra vital importancia. Teniendo en cuenta la relación con dirección positiva que se pudo capturar en

los distintos modelos estimados, vale la pena resaltar la trascendencia que pueden tener los plaguicidas en la propensión de muerte de una persona diagnosticada con esta enfermedad.

En línea con el modelo planteado en esta investigación y para poder estimar modelos de una mayor calidad, se debe de instar a las instituciones encargadas del registro de plaguicidas a realizar un control más estricto de los tipos y cantidad de plaguicidas utilizados en el país.

5.1 Limitaciones

Entre las principales limitaciones del estudio se encuentran las fuentes de información disponibles, especialmente en relación con los datos sobre la exposición a plaguicidas. Los censos agropecuarios realizados por el INEC se llevaron a cabo en 1984 y 2014, lo que deja una ventana de 30 años sin una recolección nacional de esta información. Además, el acceso a otras fuentes de información es limitado.

Los datos disponibles sobre la exposición a plaguicidas están desagregados a nivel cantonal, es decir, en unidades geográficas de gran tamaño, lo cual puede no representar adecuadamente los patrones de exposición de las personas que residen en unidades geográficas más pequeñas.

En cuanto a la información proporcionada por el Registro Nacional de Tumores, se pierden datos debido a la falta de un identificador único que permita analizar si una persona ha fallecido o no. Esta problemática también impide realizar un análisis que permita evaluar la presencia o ausencia de la enfermedad de manera precisa.

Respecto a algunos individuos y su región de residencia, como se mencionó en la sección de metodología, no se dispone de la información del cantón y distrito de residencia. Para no perder estos datos, se asignan proporcionalmente a los cantones y distritos correspondientes, lo que podría introducir sesgos en los resultados.

Otra limitación de esta investigación es la no inclusión de determinantes sociales de la salud para evaluar su relación con la probabilidad de muerte de personas diagnosticadas

con cáncer de mama o próstata. Se sugiere que futuras investigaciones consideren la inclusión de estos factores para obtener una comprensión más completa de las variables que influyen en la mortalidad por estas enfermedades.

BIBLIOGRAFÍA

- Acevedo, R. (2008). Cuaderno Metodológico 1. Los modelos jerárquicos lineales: fundamentos básicos para su uso y aplicación. San José, CR.: Instituto de Investigaciones Psicológicas, Universidad de Costa Rica.
- Anselin, Luc (1988), Spatial Econometrics: Methods and Models, Kluwer Academic, Dordrecht.
- Aparicio, A., Morera, M. (2007). La conveniencia del análisis multinivel para la investigación en salud: una aplicación para Costa Rica Población y Salud en Mesoamérica, vol. 4, núm. 2, enero-junio, 2007, p. 0 Universidad de Costa Rica San José, Costa Rica.
- Arriaga, E. E. (1996). Comentarios sobre algunos índices para medir el nivel y el cambio de la mortalidad. *Estudios Demográficos y Urbanos*, *11*(1 (31)), 5–30. http://www.jstor.org/stable/40315373
- Bassil, K. L., Vakil, C., Sanborn, M., Cole, D. C., Kaur, J. S., & Kerr, K. J. (2007). Cancer health effects of pesticides: systematic review. *Canadian family physician Medecin de famille canadien*, *53*(10), 1704–1711.
- Beard, J. D., Umbach, D. M., Hoppin, J. A., Richards, M., Alavanja, M. C., Blair, A., Sandler, D. P., & Kamel, F. (2014). Pesticide exposure and depression among male private pesticide applicators in the agricultural health study. Environmental health perspectives, 122(9), 984–991. https://doi.org/10.1289/ehp.1307450
- Bivand, R. Keitt, T. and Rowlingson, B. (2019). rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.4-8.
- https://CRAN.R-project.org/package=rgdal
- Bivand, R. (2019). classInt: Choose Univariate Class Intervals. R package version
- 0.3-3. https://CRAN.R-project.org/package=classInt
- Bravo-Durán, V., Berrocal-Montero, S.-E., Ramírez-Muñoz, F., de-la-Cruz-Malavassi, E., Canto-Mai, N., Tatis-Ramírez, A., Mejía-Merino, W., & Rodríguez-Altamirano, T. (2015). Pesticides import and health hazards. Central America, 2005 - 2009. *Uniciencia*, *29*(2), 84-106. https://doi.org/10.15359/ru.29-2.6

- Bruno A., W., & Moore, J. (2005). The concepts of bias, precision and accuracy, and their use in testing. Ecography 28 , 815-829.
- Camus M, Band P. Relationship between mortality and Cancer incidence in Montreal compared to Canada excluding Québec, 1984–1994, Montreal, Public Health Department, Montreal Health and Social Services Agency 2005.
- Cavalier, H., Trasande, L., & Porta, M. (2023). Exposures to pesticides and risk of cancer: Evaluation of recent epidemiological evidence in humans and paths forward. International journal of cancer, 152(5), 879–912. https://doi.org/10.1002/ijc.34300
- Centro Centroamericano de Población. Proyecciones de Población, 1970-2030.
- Corella, D., Herranz, C., Calatayud, A., Font, G., Celma, C., & Laborda, R. (2000). Cancer Mortality and Exposure to Chemical Carcinogens in the Work Place: An Ecological Study in the Valencian Community, Spain (1981-1995). <i>European Journal of Epidemiology, 16 (5), 401-409. Recuperado de: http://www.jstor.org/stable/3582114
- de Corso Sicilia, G. B., & Pinilla Rivera, M. (2017). Métodos gráficos de análisis exploratorio de datos espaciales con variables espacialmente distribuidas. Cuadernos Latinoamericanos De Administración, 13(25), 92–104. https://doi.org/10.18270/cuaderlam.v13i25.2417
- de la Cruz, F. (2008). Modelos Multinivel Revista Peruana de Epidemiología, vol. 12, núm. 3, diciembre, pp. 1-8 Sociedad Peruana de Epidemiología Lima, Perú. Recuperado de: https://www.redalyc.org/articulo.oa?id=203120335002
- Dich, J., Zahm, S. H., Hanberg, A., & Adami, H.-O. (1997). Pesticides and Cancer. *Cancer Causes & Control*, *8*(3), 420–443. http://www.jstor.org/stable/3552701
- Dirección de investigación y extensión de la caña de azúcar. (2000). Censo de variedades de caña de azúcar de Costa Rica 2000. https://servicios.laica.co.cr/laica-cv-biblioteca/index.php/Library/download/WEgUsTBVlNfQkOMmUaMsuhvkuNeEtDFT

- Dobson, A., & Barnett, A. (2008). An Introduction to Generalized Linear Models (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780367807849
- Dunn, P., Smyth, G. Generalized Linear Models With Examples in R. Springer texts in Statistics. https://doi.org/10.1007/978-1-4419-0118-7
- En Cheng, MD, PhD, Dong Hoon Lee, ScD, Rulla M Tamimi, ScD, Susan E Hankinson, ScD, Walter C Willett, MD, DrPH, Edward L Giovannucci, MD, ScD, A Heather Eliassen, ScD, Meir J Stampfer, MD, DrPH, Lorelei A Mucci, ScD, Charles S Fuchs, MD, MPH, Donna Spiegelman, ScD, Long-Term Survival and Causes of Death After Diagnoses of Common Cancers in 3 Cohorts of US Health Professionals, *JNCI Cancer Spectrum*, Volume 6, Issue 2, April 2022, pkac021, https://doi.org/10.1093/jncics/pkac021
- Fajardo, A. (2017). Medición en epidemiología: prevalencia, incidencia, riesgo, medidas de impacto. Revista Alergia México. 64(1):109-120.
- Faraway, J. (2016). Extending the linear model with R. Generalized Linear, Mixed effects, and nonparametric regression models. 2nd edition. CHAPMAN & HALL/CRC
- Garrido, C., Murillo, F. (2014). Programas para la realización de Modelos Multinivel. Un análisis comparativo entre MLwiN, HLM, SPSS y Stata. Revista Electrónica de Metodología Aplicada. 19 (2): 1-24.
- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018
- Gower, J.C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27: 857–874. https://doi.org/10.2307/2528823
- Grolemund, G., Wickham, H. (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.
- Gujarati, D. N. (2010). *Econometría* (4 ed.). (G. Arango Medina, Trad.) México D.F: McGraw Hill.

- Haneuse, S., Bartell, S. (2011). Review Article: Designs for the Combination of Group- and Individual-level Data. Epidemiology, 22 (3), 382–389. Recuperado de: http://www.jstor.org/stable/23047606
- Hoeting, J. A. (2009). The Importance of Accounting for Spatial and Temporal Correlation in Analyses of Ecological Data. Ecological Applications, 19 (3), 574–577. Recuperado de : http://www.jstor.org/stable/27645996
- Holmes, W., Bolin, J., Kelley. K. (2014). Multilevel Modeling Using R. Taylor and Francis Group.
- Howard J (2014) Minimum latency & types or categories of cancer. Centers for Disease Control and Prevention. November 7, 2014. https://www.cdc.gov/wtc/pdfs/policies/wtchpminlatcanc er2014-11-07-508.pdf
- Instituto Nacional de Estadsística y Censos. (1984). Censo Nacional Agropecuario.
- Instituto Nacional de Estadística y Censos. (2016). Clasificación Geográfica con fines Estadísticos.
- Instituto Nacional de Estadística y Censos. 2023. Encuesta Nacional de Hogares, Resultados Generales.
- Jiang, J., Nguyen, T. (2021). Linear and Generalized Linear Mixed Models and Their Applications. 2nd edition. Springer Series in Statistics
- Jin, X., Carlin, B. P., & Banerjee, S. (2005). Generalized Hierarchical Multivariate CAR Models for Areal Data. *Biometrics*, *61*(4), 950–961. http://www.jstor.org/stable/3695906
- Juliana Oliveira Fernandes, Cassio Cardoso-Filho , Maria Beatriz Kraft , ´ Amanda Sacilotto Detoni , Barbara Narciso Duarte , Julia Yoriko Shinzato , Diama Bhadra Vale , ´ Differences in breast cancer survival and stage by age in off-target screening groups: a population-based retrospective study, AJOG Global Reports (2023), doi: https://doi.org/10.1016/j.xagr.2023.100208
- Kutner, M., Christopher, J., Neter, J., William, L. (2004). Applied Linear Statistical Models. (5). McGraw-Hill Irwin.

- Lunn, D., Thomas, A., Spiegelhalter, D., Jackson, C., & Best, N. (2012). The BUGS Book: A Practical Introduction to Bayesian Analysis (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b13613
- March, Guillermo Juan. (2014). Agricultura y plaguicidas: un análisis global. - 1a ed. - Rio Cuarto : FADA - Fundación Agropecuaria para el Desarrollo de Argentina.
- Mattiuzzi, C., & Lippi, G. (2019). Current Cancer Epidemiology. *Journal of epidemiology and global health*, *9*(4), 217–222. https://doi.org/10.2991/jegh.k.191008.001
- Martino, S. and Riebler, A. (2020). Integrated Nested Laplace Approximations (INLA). In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). https://doi.org/10.1002/9781118445112.stat08212
- Mendenhall, W., Wackerly, D., Scheaffer, R. (2010). Estadística matemática con aplicaciones. (7th ed.). Cenage Learning.
- Ministerio de Salud. Registro Nacional de Tumores 1980-2015.
- Mink, Pamela J; Adami, Hans-Olov; Trichopoulos, Dimitrios; Britton, Nicole L.; Mandel, Jack S. 2008. Pesticides and prostate cancer: a review of epidemiologic studies with specific agricultural exposure information. European Journal of Cancer Prevention 17(2):p 97-110. DOI: 10.1097/CEJ.0b013e3280145b4c
- Morgenstern H. (1995). Ecologic studies in epidemiology: concepts, principles, and methods. *Annual review of public health*, *16*, 61–81. https://doi.org/10.1146/annurev.pu.16.050195.000425
- Monsalve, N. (2013). Modelos jerárquicos bayesianos espaciales en epidemiología agrícola. Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universidad Politécnica de Valencia.
- Morris, R. D., & Munasinghe, R. L. (1993). Aggregation of existing geographic regions to diminish spurious variability of disease rates. *Statistics in medicine*, *12*(19-20), 1915–1929. https://doi.org/10.1002/sim.4780121916

- Mostafalou, S., & Abdollahi, M. (2013). Pesticides and human chronic diseases: evidences, mechanisms, and perspectives. Toxicology and applied pharmacology, 268(2), 157–177. https://doi.org/10.1016/j.taap.2013.01.025
- Muir, K., Rattanamongkolgul, S., Smallman, M., Thomas, M., Downer, S., Jenkinson, Crispin. (2004). Breast cancer incidence and its possible spatial association with pesticide application in two counties of England. Public health. 118. 513-20. 10.1016/j.puhe.2003.12.019.
- Nelder, J., Wedderburn, W. (1972). Journal of the Royal Statistical Society. Series A (General). Vol. 135, No. 3 (1972), pp. 370-384 (15 pages)
- Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer
- Neter, J., Wasserman, W., & Kutner, M. H. (1990). Applied linear statistical models : regression, analysis of variance, and experimental designs. 3rd ed. Burr Ridge (Ill.): Irwin.
- Oaks, J., Cobb, K., Minin, V., & Leaché, A. (2019). Marginal Likelihoods in Phylogenetics: A Review of Methods and Applications. Systematic biology, 68(5), 681–697. https://doi.org/10.1093/sysbio/syz003
- OECD (2017), OECD Reviews of Health Systems: Costa Rica 2017, OECD Reviews of Health Systems, OECD Publishing, Paris, https://doi.org/10.1787/9789264281653-en.
- Organización Mundial de la Salud. (2018). Residuos de plaguicidas en los alimentos. Recuperado de https://www.who.int/es/news-room/fact-sheets/detail/pesticide-residues-in-food
- Organización Mundial de la Salud. 2020. Las 10 principales causas de defunción. https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death
- Organización Panamericana de la Salud. 2022. ONU alerta a Costa Rica sobre alto costo del uso de plaguicidas en la salud. Recuperado de: https://www.paho.org/es/noticias/27-5-2022-onu-alerta-costa-rica-sobre-alto-costo-uso-plaguicidas-salud

- Pardo, L.A., Beane Freeman, L.E., Lerro, C.C. et al. Pesticide exposure and risk of aggressive prostate cancer among private pesticide applicators. Environ Health 19, 30 (2020). https://doi.org/10.1186/s12940-020-00583-0
- Partanen, Timo, Monge, Patricia, & Wesseling, Catharina. (2009). Causas y prevención del cáncer ocupacional. *Acta Médica Costarricense*, *51*(4), 195-205. Recuperado de: http://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S000160022009000400003&lng=en&tlng=es
- Press, J. (2003). Subjective and Objective Bayesian Statistics: principles, models and aplications. DOI:10.1002/9780470317105
- Ramirez, F., Luna, S., Orozco, M., Williamson, S. 2016. Proyecto: Alternativas para la reducción y eliminación del uso de los "plaguicidas Altamente Peligrosos". Serie Informes Técnicos Instituto Regional de Estudios en Sustancias Tóxicas.
- Rogers, R. G., Everett, B. G., Onge, J. M., & Krueger, P. M. (2010). Social, behavioral, and biological factors, and sex differences in mortality. Demography, 47(3), 555–578. https://doi.org/10.1353/dem.0.0119
- Rosero-Bixby, L. (1991). Socioeconomic development, health interventios, and mortality decline in Costa Rica. Scandinavian Journal of Social Supplement 46: 33-42.
- Rue, H., Riebler, A., Sørbye, S., Illian, J., Simpson, D., Lindgren, F. (2016). Bayesian Computing with INLA: A Review. Annual Review of Statistics and Its Application 2017 4:1, 395-421. https://doi.org/10.1146/annurev-statistics-060116-054045
- Sánchez, L. (2012). Alcances y límites de los métodos de análisis espacial para el estudio de la pobreza urbana. Papeles de Población, 18(72),147-179. ISSN: 1405-7425. https://www.redalyc.org/articulo.oa?id=11223536007
- Santamaría, C. (2009). The impact of pesticide exposure on breast cancer incidence. Evidence from Costa Rica. Población y Salud en Mesoamérica, 7 (1). DOI 10.15517/PSM.V7I1.1091 Santamaría-Santamaría, C., Valverde, C. (2019).

Inequality in the Incidence of Cervical Cancer: Costa Rica 1980-2010. Frontiers in oncology, 8, 664. https://doi.org/10.3389/fonc.2018.00664

- Santamaría-Ulloa, C., & Valverde-Manzanares, C. (2019). Inequality in the Incidence of Cervical Cancer: Costa Rica 1980-2010. Frontiers in oncology, 8, 664. https://doi.org/10.3389/fonc.2018.00664
- Siabato, Willington, y Jhon Guzmán-Manrique. 2019. La autocorrelación espacial y el desarrollo de la geografía cuantitativa. Cuadernos de Geografía: Revista Colombiana de Geografía 28 (1): 1-22. doi: 10.15446/rcdg.v28n1.76919.
- Silva, L., Benavides, A., & Vidal, C. (2003). Análisis espacial de la mortalidad en áreas geográficas pequeñas: El enfoque bayesiano. *Revista Cubana de Salud Pública*, *29*(4). http://ref.scielo.org/6z6vhc
- Tribunal Supremo de Elecciones. Padrones electorales, 1990- 2014.
- Tobler, Waldo. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." Economic Geography 46 (Junio): 234-240. doi: 10.2307/143141
- Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M., & Fortin, M.-J. (2018). Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs*, *88*(1), 36–59. https://www.jstor.org/stable/26598521
- Waterhouse J. y Col. (Eds.). Cancer incidence in five continents. Lyon, IARC, 1976.
- Wesseling, C., D. Antich, C. Hodgsted, A.C. Rodríguez, and A. Ahlbom. (1999). Geographical differences of cancer incidence in Costa Rica in relation to environmental and occupational pesticide exposure. International Journal of Epidemiology 28: 365-374
- West, B., Welch, K., Gatecki, A. (2006). Linear Mixed Models, A practical guide using statistical software. Chapman and Hall/CRC.
- Wesseling, C., Ahlbom, A., Antich, D., Rodriguez, A., Castro, R. (1996). Cancer in Banana Plantation Workers in Costa Rica. *International Journal of Epidemiology*, Volumen 25, Número 6, Pages 1125–1131, Recuperado de: https://doi.org/10.1093/ije/25.6.1125

- West, B.T., Welch, K.B., & Galecki, A.T. (2006). Linear Mixed Models: A Practical Guide Using Statistical Software (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420010435
- Wickham, H., François, R., Henry, L., and Müller, K. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.6.https://CRAN.R-project.org/package=dplyr
- Xu, H. (2014). COMPARING SPATIAL AND MULTILEVEL REGRESSION MODELS FOR BINARY OUTCOMES IN NEIGHBORHOOD STUDIES. *Sociological*
- Xu, X., Dailey, A. B., Talbott, E. O., Ilacqua, V. A., Kearney, G., & Asal, N. R. (2010). Associations of Serum Concentrations of Organochlorine Pesticides with Breast Cancer and Prostate Cancer in U.S. Adults. *Environmental Health Perspectives*, *118*(1), 60–66. http://www.jstor.org/stable/30249906

ANEXOS

Anexo 1. Agrupamiento de los distritos.

A continuación, se presenta el agrupamiento realizado de los distritos utilizando los códigos que representan la provincia, cantón y distrito, así como en el nuevo código a utilizar el cual consiste en un número consecutivo.

Nuevo agrupamiento de distritos	Nuevo código
10101-10102	1
10103-10104-10105	2
10106-10111	3
10107-10109-11304	4
10108-10110	5
10201-10203-10903	6
10202-10902	7
10301-10311	8
10302-10307	9
10303-10304-10607	10
10305-10312	11
10306-10501-12005	12
10308-11702	13
10309-10604	14
10310-10313	15
10401-10405-10408	16
10403-10407	17
10404-10705	18
10402-10406	19
10409-11605	20
10502-10503	21
10601-11003	22
10602-11201	23
10603-10606	24
10605-11205	25
10701-10702-10707	26
10703-11203	27
10704-10706	28
10801-10803	29
10802	30
10804-11504	31
10805-10807	32
10806-11101-11102	33
10901-10904	34
10905-10906	35
11001-11005	36
11002-11004	37
11103-11105-11402	38
11104-11403	39
11202-11204	40
11301-11305	41
11303-11401	42
11501-11502-11503	43
11601-11602	44
11603-11604	45
11701-11703	46
11901-11906	47

Nuevo agrupamiento de distritos	Nuevo código
11902-11903	48
11904-11911	49
11905-11908	50
11907-11912	51
11909-11910	52
12001-12006	53
12002-12003-12004	54
11801-11802	219
20101-20110	55
11803-11804	220
20102-20112	56
20103-20106-20107	57
20104-20105	58
20108-20109	59
20111-20113	60
20114-21601	61
20201-20203	62
20202-20206	63
20204-20212	64
20205-20209	65
20207-20211	66
20208-20210-20213-20214	67
20301-20303	68
20302-20304-20308	69
20305-20307	70
20401-20402	71
20403-20404	72
20501-20505	73
20502-20508	74
20503-20504	75
20506-20507	76
20601-20608	77
20602-20607	78
20603-20606	79
20604-20605	80
20701-20703-20707	81
20702-20704	82
20705-20706	83
20801-20804-20805	84
20802-20803	85
20901-20902-20903	86
20904-20905	87
21001-21106	88
21002-21003-21008	89
21004-21009	90
21005-21203	91
21006-21011	92
21007-21010	93
21012-21013	94
21101-21104	95
21102-21103-21105-21107	96
21201-21205	97
21202-21204	98
21301-21304	99
21302-21306	100
21303-21308	101
21305-21307-21402	102
21401-21403-21404	103
21501-21502	104

21503-21504	105
21602-21603	106
Nuevo agrupamiento de distritos	Nuevo código
30101-30102	107
30103-30108	108
30104-30110	109
30105-30802	110
30106-30111	111
30107-30512-30804	112
30109-30203-30803	113
30201-30205	114
30301-30302	116
30303-30308	117
30304-30306	118
30305-30307	119
30401-30509	120
30402-30403	121
30501-30511	122
30502-30506	123
30503-30510	124
30504-30505	125
30507-30508	126
30601-30603	127
30602-30704	128
30701-30702	129
30703-30705	130
30801	131
50101-50401	132
50102-50105	133
50103-50104	134
50201-50202	135
50203-50204	136
50205-50207	137
50206-50303-50306	138
50301-50307	139
50302-50501	140
50304-50305	141
50308-50309	142
50402-50403	143
50404-50602	144
50502-50503-50504	145
50601-50805	146
50603-50701-50704	147
50604-50605	148
50702-50703	149
50801-50802-50807	150
50803-50808-50903	151
50804-50806-50901	152
50902-50904	153
50905-50906	154
51001-51004	155
51002-51003	156
51101-51104-51105	157
51102-51103	158
40101-40102	159
40103-40104	160
40105-41003	161
40201-40204	162
40202-40203-40206	163
40205-40504-40505	164

40301-40306	165
40302-40305	166
40303-40307	167
40304-40308	168
Nuevo agrupamiento de distritos	Nuevo código
40401-40404	169
40402-40403	170
40405-40406	171
40501-40502-40503	172
40601-40604	173
40602-40603	174
40701-40702-40703	175
40801-40803	176
40901-40902	177
41001-41004	178
41002-41005	179
60101-60102	180
60103-60106	181
60104-60111	182
60105-60107-60109	183
60108-60205-60403	184
60112-60115	185
60113	186
60114-60116	187
60201-60202	188
60203	189
60204	190
60206-61102	191
60601-60602-60603-60901-61101	192
60501-60502-60504	193
60301-60303-60307	194
60304-60305-60306	195
60302-60309	196
60308-60305	197
60401-60402	198
60503-60506-60702	199
60505-60701-60703	200
60704-61004	201
60801-60804	202
60802-60806	203
60803-61001	204
61002-61003	205
70101-70103	206
70102-70401-70404	207
70104-70501-70503	208
70201-70207	209
70202-70601	210
70203-70204-70205-70206	211
70301-70307	212
70302-70502	213
70303-70306	214
70304-70305	215
70402-70403	216
70602-70603	217
70604-70605	218

Anexo 2. Agrupación según conglomerados dentro de las provincias

Distritos agrupados	Nuevo código de los distritos agrupados según conglomerados
10101-10102-10103-10104-10105-10106-10107-10108-10109-10110-10111-10201-10203-10301-10303-10304-10305-10310-10311-10312-10313-10801-10802-10803-10805-10807-10901-10903-10906-11001-11002-11004-11005-11104-11301-11302-11303-11304-11305-11401-11403-11501-11502-11503-11504-11801-11802-11803-11804	1
11101-11102-11103-11105-11402	2
10701-10904-10905	3
11901-11903	4
10502-11701	5
11703-11902-11904-11905-11909-11910-11911	6
11906	7
11907	8
11912	9
10501-12001-12006	10
11702-12005	11
12004	12
10503	13
10306	14
10308	15
10606	16
11605	17
11602-11604	18
10704-10705-10706-10707	19
10307-	20
11601	21
10403	22
10401	23
10202-10302-10601-10607-11003	24
10806	25
10804	26
11603	27
10407	28
10405-10406	29
10402-10404-10408-10409-10605-11202-11204-11205-12002-12003	30
10309-10602	31
10603-10604-10702-10703-10902-11201-11203	32
11908	33
21302-21303-21306	34
21301-21304-21307	35
21305	36
21308	37
21402-21504	38
21502	39
21503	40
21501	41
21010-21012-21404	42
21004-21005-21006-21011-21013-21401-21403	43
20213-21008	44
21001-21002-21007	45
20107-20114-21009-	46
20101-20102-20104-20109	47
20305	48
20805	49

20802-20803	50
20504	51
20905	52

Distritos agrupados	Nuevo código de los distritos agrupados según conglomerados
20901-20903	53
20702-20707	54
20703-20706	55
20701	56
20704	57
20602-20607-20705	58
70202-70203-70205-70604-70605	59
20502-20503	60
20201-20203-20211	61
20209	62
20601-20605	63
21101-21105	64
20404	65
20402-20508	66
20103-20106-20110	67
20105-20108-20111-20112-20113-20501-20505-20801-20804	68
20301-20302-20303-20304-20307-20506-20507	69
21102-21103-21107	70
20204-20205-20208-20210-20603-20604-20606	71
20202-20206-20207-20401-20403-20902-20904	72
20212-20214	73
21003-21106	74
20308-21201-21202-21204-21205	75
20608	76
21104	77
21203-21601-21602-21603	78
30512	79
30507-30508	80
30502	81
30510	82
30503	83
30401-30501-30506-30509-30511	84
30505	85
30504	86
30602	87
30601-30603-30704	88
30202-30204-30402-30403	89
30804	90
30107	91
30101-30102-30103-30104	92
30301-30303-30305-30307	93
30105-30106-30108-30109-30111-30201-30203-30205-30701-30702-30703-30705-30801-30802-30803	94
30302-30304-30306-30308	95
30110	96
41004-41005	97
40105-40405-41001-41002-41003	98
40202-40206-40404	99
40402-40406	100
40203	101
40802	102
40308	103
40401-40403	104
40504-40505-40602-40603	105
40304	106

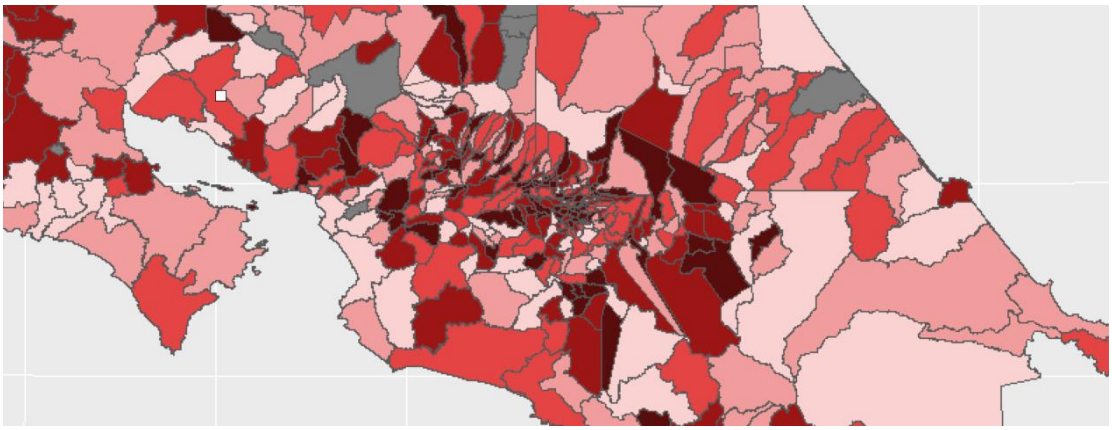
40101-40102-40103-40104-40201-40204-40205-40301-40302-40303-40305-40306-40307-40501-40502-40503-40601-40604-40701-40702-40703-40801-40803-40901-40902	107
51001-51002-51003-51004	108
50102-50105	109

Distritos agrupados	Nuevo código de los distritos agrupados según conglomerados
50103	110
50403	111
50602	112
50803-50808	113
50703	114
50604-50605	115
50302	116
50306	117
50309	118
51105	119
51102-51103-51104	120
50101-50104-50201-50301-50305-50501-50502-50503-50504	121
50601	122
50801	123
50804-50806-50807	124
50802	125
50702-50805-	126
50603-50701-50704	127
50203-50307-50401-50402	128
50202-50901-50902-50904-51101	129
50205-50206	130
50204	131
50207	132
50303	133
50304-50308	134
50404	135
50903-50905	136
50906	137
60104-60105-60111	138
60103-60106	139
60102	140
60109	141
60107	142
60113	143
60115-60401	144
60101-60309	145
60402	146
60108-60112-60114	147
60204	148
60116-60201-60202	149
60203	150
60205	151
61101	152
61004	153
60806	154
60506	155
60503-6061-60602	156
60403	157
60304-60305	158
60301-60308	159
60206	160
60302	161
60307-60804-60805	162

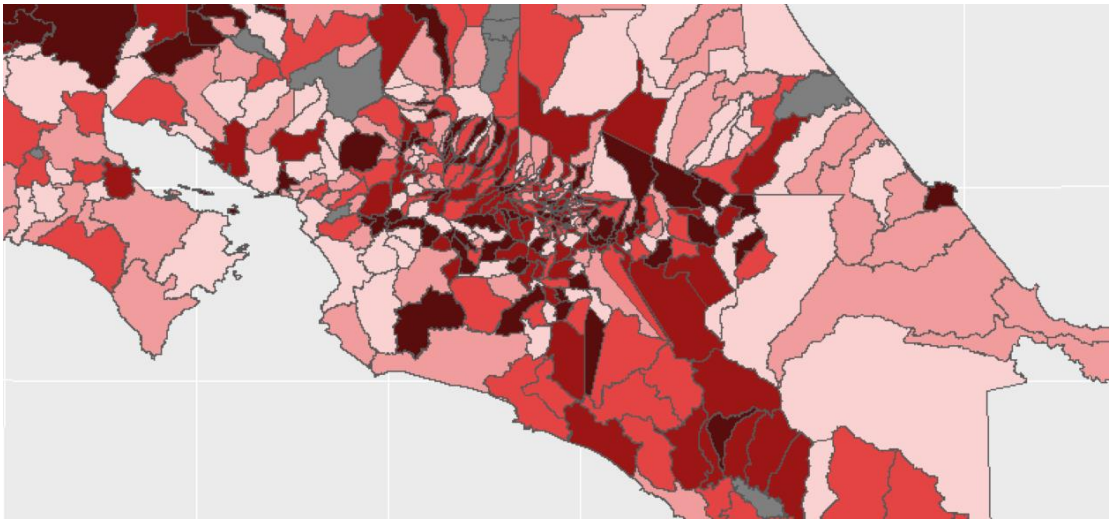
60306-60803	163
60704-60801-60802	164
60603-60702-60901-61002	165
60703-61001-61003	166
60303-60501-60502-60504	167
60505	168
60701	169

Distritos agrupados	Nuevo código de los distritos agrupados según conglomerados
70404	170
70201-70206	171
70602	172
70305	173
70306	174
70204	175
70104	176
70207-70601	177
70603	178
61102	179
70101-70102-70103-70301-70401-70402-70403-70501-70502-70503	180
70302-70303-70304	181

Anexo 3. Distribución por distritos de las tasas de cáncer de mama. 2011-2015.



Anexo 4. Distribución por distritos de las tasas de cáncer de próstata. 2011-2015.



Anexo 5.

Cuadro 1: Distribución de frecuencias de la edad de diagnóstico de cáncer por sexo, 2011-2015.

Clases	Hombres				Mujeres			
	Frecuencia	%	Frecuencia acumulada	Porcentaje acumulado	Frecuencia	%	Frecuencia acumulada	Porcentaje acumulado
[10,20[	0	0.0	0	0.0	2	0.0	2	0.0
[20,30[	1	0.0	1	0.0	53	0.9	55	1.0
[30,40[	5	0.1	6	0.1	386	6.8	441	7.7
[40,50[	112	2.3	118	2.4	1250	21.9	1691	29.6
[50,60[	814	16.6	932	19.0	1607	28.1	3298	57.7
[60,70[	1880	38.4	2812	57.4	1324	23.2	4622	80.9
[70,80[	1496	30.6	4308	88.0	746	13.1	5368	93.9
[80,90[	529	10.8	4837	98.8	305	5.3	5673	99.3
[90,100[	54	1.1	4891	99.9	42	0.7	5715	100.0
[100,110[	5	0.1	4896	100.0	1	0.0	5716	100.0

Anexos 6.

Resultados de modelos de las mujeres

Cuadro 6.1. Resultados de los modelos de regresión bayesianos estimados para la unidad geográfica de distritos entre la letalidad por cáncer de mama y la edad, tasa de cáncer, Índice de Exposición a Plaguicidas en mujeres (N=5716).

Modelo	Criterio de vecindad	Coeficientes				Criterios de bondad de ajuste			Precisión	
		Intercepto	Edad	Tasa de cáncer	Índice de Exposición a Plaguicidas	Log verosimilitud marginal	DIC	WAIC	Estructura jerárquica	Estructura espacial
De un nivel	-	-3.515*	0.035*	-0.410*	0.390*	-2737.06	5438.79	5439.04		
jerárquico no espacial	-	-3.515*	0.035*	-0.411*	0.390*	-2737.13	5437.31	5438.85	21571.07	
jerárquico espacial	Reina	-3.516*	0.035*	-0.413*	0.392*	-2920.81	5437.7	5438.82	18283.03	18512.42
jerárquico espacial	Torre	-3.516*	0.035*	-0.413*	0.392*	-2930	5437.72	5438.84	17945	18461.78

Cuadro 6.2. Resultados de los modelos de regresión bayesianos estimados para la unidad geográfica resultante de la agrupación de distritos dentro de las provincias entre la letalidad por cáncer de mama y la edad, tasa de cáncer, Índice de Exposición a Plaguicidas en mujeres (N=5716).

Modelo	Criterio de vecindad	Coeficientes				Criterios de bondad de ajuste			Precisión	
		Intercepto	Edad	Tasa de cáncer	Índice de Exposición a Plaguicidas	Log verosimilitud marginal	DIC	WAIC	Estructura jerárquica	Estructura espacial
De un nivel	-	-3.517*	0.035*	-0.498*	0.468*	-2735.84	5436.54	5436.83		
jerárquico no espacial	-	-3.517*	0.035*	-0.499*	0.468*	-2735.94	5436.05	5436.75	19684.48	
jerárquico espacial	Reina	-3.517*	0.035*	-0.5*	0.468*	-2757.92	5436.14	5436.68	23616.33	22051.31
jerárquico espacial	Torre	-3.517*	0.035*	-0.5*	0.468*	-2759.09	5436.1	5436.66	20583.58	21790.07

Cuadro 6.3. Resultados de los modelos de regresión bayesianos estimados para la unidad geográfica de distritos utilizando distribuciones previas entre la letalidad por cáncer de mama y la edad, tasa de cáncer, Índice de Exposición a Plaguicidas en mujeres (N=5716).

Modelo	Criterio de vecindad	Distribución previa		Coeficientes				Criterios de bondad de ajuste			Precisión	
		Jerárquica	Espacial	Intercepto	Edad	Tasa de cáncer	Índice de Exposición a Plaguicidas	Log verosimilitud marginal	DIC	WAIC	Estructura jerárquica	Estructura espacial
Jerárquico espacial	Reina	Normal(0,0.0001)	Loggamma(0.5, 0.0005)	-3.530*	0.035*	-0.422*	0.402*	-2924.06	5437.07	5437.69	44.01	948.38
Jerárquico espacial	Torre	Normal(0,0.0001)	Loggamma(0.5, 0.0005)	-3.530*	0.035*	-0.424*	0.402*	-2933.41	5437.85	5437.73	43.81	1086.55
Jerárquico no espacial	-	Normal(0,0.0001)	-	-3.523*	0.035*	-0.415*	0.393*	-2740.3	5437.92	5438.15	5.03E+09	
Jerárquico espacial	Reina	-	Loggamma(0.5, 0.0005)	-3.521*	0.035*	-0.417*	0.4*	-2920.88	5437.44	5438.7	18388.92	796.07
Jerárquico espacial	Torre	-	Loggamma(0.5, 0.0005)	-3.520*	0.035*	-0.419*	0.4*	-2930.03	5437.47	5438.78	18464.27	763.19

Cuadro 6.4. Resultados de los modelos de regresión bayesianos estimados para la unidad geográfica resultante de la agrupación de distritos dentro de las provincias utilizando distribuciones previas entre la letalidad por cáncer de mama y la edad, tasa de cáncer, Índice de Exposición a Plaguicidas en mujeres (N=5716).

Modelo	Criterio de vecindad	Distribución previa		Coeficientes				Criterios de bondad de ajuste			Precisión	
		Jerárquica	Espacial	Intercepto	Edad	Tasa de cáncer	Índice de Exposición a Plaguicidas	Log verosimilitud marginal	DIC	WAIC	Estructura jerárquica	Estructura espacial
Jerárquico espacial	Reina	Normal(0,0.0001)	Loggamma(0.5, 0.0005)	-3.522*	0.035*	-0.511*	0.472*	-2761.82	5435.45	5436.13	1090.12	948.38
Jerárquico espacial	Torre	Normal(0,0.0001)	Loggamma(0.5, 0.0005)	-3.523*	0.035*	-0.512*	0.472*	-2762.84	5435.23	5436.15	1286.84	1086.55
Jerárquico no espacial	-	Normal(0,0.0001)	-	-3.517*	0.035*	-0.5*	0.466*	-2739.75	5435.9	5436.84	3.3E+16	
Jerárquico espacial	Reina	-	Loggamma(0.5, 0.0005)	-3.523*	0.035*	-0.512*	0.474*	-2757.76	5435.53	5436.21		
Jerárquico espacial	Torre	-	Loggamma(0.5, 0.0005)	-3.523*	0.035*	-0.512*	0.474*	-2758.9	5435.5	5436.17		

Cuadro 6.5. Resultados de los modelos de regresión bayesianos estimados para la unidad geográfica resultante del análisis de conglomerados entre la letalidad por cáncer de mama y la edad, tasa de cáncer, Índice de Exposición a Plaguicidas en mujeres (N=5716).

Modelo	Criterio de vecindad	Coeficientes				Criterios de bondad de ajuste			Precisión	
		Intercepto	Edad	Tasa de cancer	Índice de Exposición a Plaguicidas	Log verosimilitud marginal	DIC	WAIC	Estructura jerárquica	Estructura espacial
De un nivel	-	-3.471*	0.035*	-0.975	0.273*	-2737.05	5441.42	5441.61		
jerárquico no especial	-	-3.469*	0.035*	-1.013	0.269*	-2737.09	5440.54	5441.28	20831.72	
jerárquico espacial	Reina	-3.469*	0.035*	-1.016	0.269*	-2751.42	5440.74	5441.22	19698.93	22091.78
jerárquico espacial	Torre	-3.469*	0.035*	-1.019	0.268*	-2752.33	5440.6	5441.23	18194.07	18208.45

Cuadro 6.6. Resultados de los modelos de regresión bayesianos estimados para la unidad geográfica resultante del análisis de conglomerados usando distribuciones previas entre la letalidad por cáncer de mama y la edad, tasa de cáncer, Índice de Exposición a Plaguicidas en mujeres (N=5716).

Modelo	Criterio de vecindad	Distribicuón previa		Coeficientes				Criterios de bondad de ajuste			Precisión	
		Jerarquica	Espacial	Intercepto	Edad	Tasa de cáncer	Índice de Exposición a Plaguicidas	Log verosimilitud marginal	DIC	WAIC	Estructura jerárquica	Estructura espacial
Jerarquico espacial	Reina	Normal(0,0.0001)	Loggamma(0.5, 0.0005)	-3.467*	0.036*	-1.156	0.245	-2754.36	5439.56	5440.35	78.53	1022.43
Jerarquico espacial	Torre	Normal(0,0.0001)	Loggamma(0.5, 0.0005)	-3.467*	0.036*	-1.156	0.245	-2755.24	5440.23	5440.39	80.03	963.19
Jerarquico no espacial	-	Normal(0,0.0001)	-	-3.468*	0.036*	-1.082	0.254	-2739.87	5440.79	5440.97	5.47E+12	
Jerarquico espacial	Reina	-	Loggamma(0.5, 0.0005)	-3.468*	0.036*	-1.115	0.26	-2751.47	5440.2	5440.96	18774.63	655.98
Jerarquico espacial	Torre	-	Loggamma(0.5, 0.0005)	-3.468*	0.035*	-1.11	0.26	-2752.26	5440.29	5441.02	18058.38	611.97

Anexos 7

7.1 Resultados simulación agrupación original usando distribuciones previas

A continuación, se presentan los resultados de los modelos de regresión jerárquico y jerárquicos espaciales utilizando distribuciones previas para los componentes jerárquico y jerárquico espacial. Por ende, no se muestran los resultados para el modelo lineal generalizado.

Cuadro 7.1.1. Medidas de posición y variabilidad del coeficiente de regresión de exposición a plaguicidas según el modelo estimado, para la unidad geográfica de distritos usando distribuciones previas.

Modelo	Promedio	Cuartil 1	Mediana	Cuartil 3	Máximo	Mínimo	Desviación estándar
Bajo, β=0.2							
Jerárquico	0.259	-0.29	0.25	0.818	2.084	-1.548	0.788
Torre	0.261	-0.291	0.255	0.815	2.089	-1.554	0.789
Reina	0.261	-0.91	0.263	0.817	2.087	-1.542	0.89
Medio, β=0.35							
Jerárquico	0.325	-0.234	0.277	0.822	2.491	-1.433	0.766
Torre	0.327	-1.433	0.289	0.823	2.494	-1.433	0.777
Reina	0.327	-1.434	0.289	0.831	2.503	-1.434	0.777
Alto, β=0.50							
Jerárquico	0.544	0.091	0.541	0.937	2994	-1.328	0.741
Torre	0.547	0.093	0.542	0.951	2.999	-1.328	0.742
Reina	0.546	0.094	0.543	0.953	2.999	-1.331	9.742

Cuadro 7.1.2. Medidas de bondad de ajuste según el modelo estimado, para la unidad geográfica de distritos usando distribuciones previas.

Modelo	LML	WAIC	DIC
	Bajo, β=0.2		
Jerárquico	-2262.79	4186.703	4219.159
Torre	-2457.47	4186.601	4218.968
Reina	-2448.24	4186.586	4218.962
	Medio, β=0.35		
Jerárquico	-2265.61	4191.704	4223.817
Torre	-2460.39	4191.584	4223.638
Reina	-2450.86	4191.584	4223.653
	Alto, β=0.50		
Jerárquico	-2269.32	4197.363	4229.9
Torre	-2464.07	4197.195	4229.675
Reina	-2454.79	4197.195	4229.688

Cuadro 7.1.3. Medidas de exactitud para el coeficiente de regresión de exposición a plaguicidas según el modelo estimado, para la unidad geográfica de distritos usando distribuciones previas.

Modelo	ECM	RMS	MAE
	Bajo, β=0.2		
Jerárquico	0.0158	0.1259	0.0015
Torre	0.0158	0.126	0.0015
Reina	0.0158	0.126	0.0015
	Medio, β=0.35		
Jerárquico	0.0152	0.1235	0.0006
Torre	0.0153	0.1237	0.0005
Reina	0.0153	0.1237	0.0005
	Alto, β=0.50		
Jerárquico	0.01398	0.0016	0.0011
Torre	0.01402	0.0011	0.0011
Reina	0.01401	0.0016	0.0011

Cuadro 7.1.4. Precisión para el componente jerárquico y jerárquico espacial según el modelo estimado, para la unidad geográfica de distritos usando distribuciones previas.

Modelo	Precisión jerárquica	Precisión espacial
Bajo, β=0.2		
Jerárquico	0.268	-
Torre	41.2717	15061.65
Reina	0.2672	109066
Medio, β=0.35		
Jerárquico	0.268	-
Torre	0.267	10618.25
Reina	0.267	115786.7
Alto, β=0.50		
Jerárquico	0.265	-
Torre	0.265	14347.66
Reina	0.265	32504.29

7.2. Resultados de la simulación de los conglomerados usando distribuciones previas

Cuadro 7.2.1. Medidas de posición y variabilidad del coeficiente de regresión de exposición a plaguicidas según el modelo estimado, para la unidad geográfica resultante del análisis de conglomerados usando distribuciones previas.

Modelo	Promedio	Cuartil 1	Mediana	Cuartil 3	Máximo	Mínimo	Desviación estándar
Bajo, β=0.2							
Jerárquico	0.331	-0.474	0.357	1.337	3.336	-4.079	1.276
Torre	0.331	-0.5	0.342	1.326	3.337	-4.191	1.281
Reina	0.332	-0.474	0.349	1.341	3.34	-4.091	1.278
Medio, β=0.35							
Jerárquico	0.493	-0.256	0.443	1.269	3.674	-4.338	1.292
Torre	0.493	-0.25	0.436	1.263	3.674	-4.374	1.294
Reina	0.492	-0.253	0.44	1.264	3.671	-4.363	1.295
Alto, β=0.50							
Jerárquico	0.6589	-0.107	0.574	1.429	4.367	-3.591	1.2916
Torre	0.6553	-0.106	0.574	1.429	4.382	-3.602	1.2974
Reina	0.6617	-0.056	0.571	1.432	4.395	-3.797	1.3019

Cuadro 7.2.2. Medidas de bondad de ajuste según el modelo estimado, para la unidad geográfica resultante del análisis de conglomerados usando distribuciones previas.

Modelo	LML	WAIC	DIC
	Bajo, β=0.2		
Jerárquico	-2093.32	4000.81	4010.94
Torre	-2109.85	4000.76	4010.91
Reina	-2109.28	4000.76	4010.92
	Medio, β=0.35		
Jerárquico	-2099.32	4012.85	4023.98
Torre	-2115.81	4012.85	4022.98
Reina	-2115.12	4012.85	4022.98
	Alto, β=0.50		
Jerárquico	-2105.19	4025.17	4034.43
Torre	-2954	425.094	4034.43
Reina	-2124.81	4025.05	4034.35

Cuadro 7.2.3. Medidas de exactitud para el coeficiente de regresión de exposición a plaguicidas según el modelo estimado, para la unidad geográfica resultante del análisis de conglomerados usando distribuciones previas

Modelo	ECM	RMS	MAE
	Bajo, β=0.2		
Jerárquico	0.0417	0.204	0.0033
Torre	0.042	0.205	0.0033
Reina	0.0418	0.204	0.0033
	Medio, β=0.35		
Jerárquico	0.0428	0.2069	0.0036
Torre	0.0429	0.2072	0.0036
Reina	0.0429	0.2073	0.0036
	Alto, β=0.50		
Jerárquico	0.0428	0.0029	0.004
Torre	0.0432	0.0039	0.0039
Reina	0.0435	0.0029	0.0041

Cuadro 7.2.4. Precisión para el componente jerárquico y jerárquico espacial según el modelo estimado, para la unidad geográfica resultante del análisis de conglomerados usando distribuciones previas.

Modelo	Precisión jerárquica	Precisión espacial
Bajo, β=0.2		
Jerárquico	0.258	-
Torre	0.2628	1537.719
Reina	0.2625	2377660
Medio, β=0.35		
Jerárquico	0.2581	-
Torre	0.2619	1370.762
Reina	0.2473	1292.546
Alto, β=0.50		
Jerárquico	0.2633	-
Torre	0.2736	10735454
Reina	0.2935	3814238

7.3 Resultados de la simulación de la unión de distritos dentro de las provincias usando distribuciones previas

Cuadro 7.3.1. Medidas de posición y variabilidad del coeficiente de regresión de exposición a plaguicidas según el modelo estimado, para la unidad geográfica resultado de la agrupación de distritos usando distribuciones previas.

Modelo	Promedio	Cuartil 1	Mediana	Cuartil 3	Máximo	Mínimo	Desviación estándar
Bajo, β=0.2							
Jerárquico	0.246	-0.35	0.292	0.865	2.339	-2.73	0.968
Torre	0.248	-0.317	0.294	0.867	2.336	-2.735	0.9683
Reina	0.251	-0.316	0.305	0.868	2.335	-2.73	0.9684
Medio, β=0.35							
Jerárquico	0.339	-0.341	0.408	0.967	2.611	-2.277	0.946
Torre	0.341	-0.357	0.408	0.982	2.61	-2.278	0.947
Reina	0.341	-0.342	0.412	0.98	2.611	-2.277	0.948
Alto, β=0.50							
Jerárquico	0.498	-0.091	0.572	1.058	2.353	-2.295	0.853
Torre	0.502	-0.091	0.572	1.059	2.354	-1.289	0.855
Reina	0.501	-0.092	0.577	1.063	2.352	-2.294	0.856

Cuadro 7.3.2. Medidas de bondad de ajuste según el modelo estimado, para la unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

Modelo	LML	WAIC	DIC
	Bajo, β=0.2		
Jerárquico	-2167.06	4050.053	4069.006
Torre	-2191.56	4049.942	4068.895
Reina	-2190.26	4049.949	4068.905
	Medio, β=0.35		
Jerárquico	-2173.2	4063.173	4082.069
Torre	-2197.61	4063.082	4081.986
Reina	-2196.24	4063.08	4081.969
	Alto, β=0.50		
Jerárquico	-2172.99	4061.793	4080.582
Torre	-2197.25	4061.716	4080.582
Reina	-2196.29	4061.694	4080.504

Cuadro 7.3.3. Medidas de exactitud para el coeficiente de regresión de exposición a plaguicidas según el modelo estimado, para la unidad geográfica de distritos usando distribuciones previas

Modelo	ECM	RMS	MAE
	Bajo, β=0.2		
Jerárquico	0.0237	0.1542	0.0011
Torre	0.0238	0.1542	0.0012
Reina	0.0238	0.1542	0.0013
	Medio, β=0.35		
Jerárquico	0.0226	0.1506	0.00026
Torre	0.0227	0.1507	0.00021
Reina	0.0227	0.1509	0.00021
	Alto, β=0.50		
Jerárquico	0.0184	0.0019	0.00004
Torre	0.0185	0.00006	0.00006
Reina	0.0185	0.0019	0.00003

Cuadro 7.3.4. Precisión para el componente jerárquico y jerárquico espacial según el modelo estimado, para la unidad geográfica resultado de la agrupación de distritos dentro de las provincias.

Modelo	Precisión jerárquica	Precisión espacial
Bajo, β=0.2		
Jerárquico	0.26	
Torre	0.267	79546.29
Reina	0.265	1348.022
Medio, β=0.35		
Jerárquico	0.263	
Torre	0.269	9560.37
Reina	0.267	6101.36
Alto, β=0.50		
Jerárquico	0.261	
Torre	0.265	5770.28
Reina	0.265	5841.55