UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

MAPEO DE ENFERMEDADES CON ENFOQUES BAYESIANOS Y EVALUACIÓN DEL RIESGO RELATIVO PARA LA SALUD PÚBLICA EN COLOMBIA

MARLON RICARDO RUIZ FERNANDEZ

Facultad de Ingeniería
Ingeniería Catastral y Geodesia
Bogotá, Colombia
2020

Mapeo de enfermedades con enfoques Bayesianos y evaluación del riesgo relativo de la salud pública en Colombia

Marlon Ricardo Ruiz Fernandez

Cod. 20151025100

Trabajo de grado presentado como requisito parcial para optar al título de:
Ingeniero Catastral y Geodesta

Director(a):
Ph.D Carlos Eduardo Melo Martinez

Línea de Investigación:
Estadística espacial

Universidad Distrital Francisco José de Caldas
Facultad de Ingenieria
Bogotá, Colombia
2020

Dedicatoria

A mis padres

"La Teoría de la probabilidad en el fondo sólo es sentido común reducido a cálculo; nos hace apreciar con exactitud lo que las mentes razonables toman por un tipo de instinto, incluso sin ser capaces de darse cuenta..."

Laplace (1820)

Nota de aceptación

Firma del Director: Ph.D Carlos Eduardo Melo Martinez

Firma del Revisor: M.Sc Hector Javier Fuentes Lopez

Agradecimientos

Quiero agradecer especialmente al director del presente trabajo al profesor Carlos Eduardo Melo Martínez por la orientación brindada y sus oportunas observaciones para el desarrollo del proyecto. A la Universidad Distrital Francisco José de Caldas, directamente al proyecto curricular de Ingeniería catastral y geodesia por su formación académica y profesional.

También a aquellos que oportunamente solucionaron las dudas presentadas, a los doctores Ph.D Yu (Ryan) Yue y Ph.D Haavard Rue quienes amablemente me ayudaron con la implementación de la librería INLA. Al doctor Ph.D Alejandro Feged-Rivadeneira quien oportunamente facilitó la base de datos inicial para la implementación de la metodología.

Finalmente a mi compañero de carrera José Enrique Gomez Gomez con quien inicialmente se planteo el proyecto y que durante su ejecución aporto al mismo.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Antecedentes

Los datos de salud pública brindan información para identificar problemas, además de prevenir y controlar un amplio número de enfermedades y comportamientos relacionados con la salud. El mapeo de enfermedades y la estimación de riesgos son de gran interés en el área de la salud pública.

Los datos georreferenciados de la incidencia de una enfermedad puede acercar a la comprensión de los orígenes y causas de muchas enfermedades, de los primeros ejemplos de la importancia del análisis geográfico de enfermedades fue el análisis de brotes de cólera en el este de Londres por Snow (1857) en donde el efecto del suministro de agua sobre el cólera, determino en primera medida que el suministro de agua no potable, generaba una mortalidad seis veces mayor a la que se generaba en los suministros de agua potable. En ese momento los métodos de propagación y naturaleza del cólera eran desconocidos, al observar los pacientes y planear donde vivían los infectados se determinó que el cólera podría propagarse mediante agua infectada. Su mapa de puntos de las residencias de las víctimas mostró un gran grupo al lado de una bomba de agua en donde mas adelante se dieron indicios que la bomba fue contaminada por material fecal de un caso de cólera. Posterior a esto, la empresa local de agua se encargó de mejorar la calidad del suministro de agua.

Otro ejemplo que podría ilustrase es el de Holden (1880) el cual estudio la distribución comunitaria de muertes prevenibles durante las epidemias tifoideas de 1870, estableciendo que esta fiebre estaba asociada a la ausencia de un servicio de alcantarillado con algunas características geográficas como vertientes y valores de altura cercanos al nivel del mar.

Los anteriores determinan casos de estudio que abrieron la posibilidad de un enfoque estadístico a la ocurrencia de enfermedades o a su proliferación, así pues, el área de estudio de estos se limitaba a algún distrito como máximo, y aun no se consideraba el factor tiempo.

Los estudios más recientes como el de Sloan et al. (2015) que examinaron los riesgos no solamente espaciales si no espacio-temporales, estimaron el riesgo excesivo de cáncer testicular mediante la detección de ubicaciones de exposición compartida entre pacientes con cáncer testicular. Diseñó un estudio de casos y controles en donde se especifican los datos del historial residencial y el ajuste por factores de riesgo personales para probar la hipótesis nula de riesgo uniforme con estadísticas basadas en el vecino más cercano.

Los datos de cáncer danés incluyen un total de 3297 casos diagnosticados entre 1991 y 2003, además se tuvieron en cuenta dos conjuntos adicionales como fecha de nacimiento y grupos de control, junto con los registros de historias residenciales. Los factores de riesgo fueron examinados en modelos de regresión logística condicional a nivel individual y socioeconómico. El único predictor significativo en el aumento del riesgo fue el historial de cáncer testicular.

Kandhasamy y Ghosh (2017) en su estudio Riesgo relativo para el VIH en India, un estimado usando modelos condicionales auto-regresivos con aproximación Bayesiana efectúan una aplicación directa de los modelos CAR para modelar la enfermedad del VIH, además de una interesante especificación de disimilitud dentro de la matriz de pesos espaciales que es capaz de identificar limites o subregiones en las que el cambio del fenómeno de interés sea abrupto. Teniendo en cuenta la función de correlación entre los efectos ($_k$; $_j j$ $_{kj}$) en el modelo Leroux, dada por:

$$(\ _k; \ _j j \ \ _{kj}) = \frac{W_{kj}}{\sqrt{(\ \sum_{i=1}^{n} W_{ki} + 1 \ \ )(\ \sum_{i=1}^{n} W_{ki} + 1 \ \ )}}$$

Se consideran los elementos de matriz W de tal forma que cada uno de sus elementos mida la disimilitud entre dos regiones k y j dadas q variables $z_{ij} = (z_1; z_2; :::; z_3)$

$$W_{kj}(\ ) = \begin{cases} 1 & si \quad e^{(\ \sum_{i=1}^{q} z_{kj} \ \ i)} \\ 0 & si \qquad no \end{cases}$$

Adicionalmente, se hace uso del criterio DIC para la selección del mejor modelo de entre todas las especificaciones mencionadas.

Se tomaron datos referidos al año 2011 de 28 estados y 6 territorios unidos de India para un total de 34 áreas de estudio. Los registros de VIH son tomados del gobierno Indio y corresponden con el conteo de casos positivos de VIH para cada región referidos al año 2011. Las covariables empleadas para describir la variabilidad del VIH fueron:

- Proporción de mujeres trabajadoras
- Proporción de consumidores de doras vía intravenosa

- Taza de alfabetización.

En cada modelo se estimaron con distintos conjuntos de covariables debido a la existencia de correlación entre estas. Una vez estimados se obtiene el valor de DIC, según el cual el modelo que mejor explica la correlación es el BYM sin la variable de consumidores de drogas intravenosas construido con una matriz de pesos espaciales *W* basada en mínimas distancias, por lo que se intuye que la autocorrelación espacial es alta y existe una parte independiente de relación que no es capturada por las variables consideradas. Aplicando el I de Moran a los residuos del modelo se detecta aun presencia de auto-correlacion entre estos, confirmando así que el modelo no describe plenamente los datos.

Las conclusiones se centran en la comparación del riesgo relativo estimado con el propuesto en los informes publicados por el gobierno, lo mas resaltable es que el riesgo estimado muestra un total de 19 regiones propensas al VIH de las cuales solo 5 corresponden con los establecidos en estudios anteriores. Asimismo, la región de Maharashtra es estimada en bajo riesgo según el modelo pero es considerada altamente propensa al VIH por los estudios gubernamentales.

Los últimos dos casos, son estimados con enfoques bayesianos, que están ligados a realizar simulaciones y dependen de la capacidad de la estación de trabajo en gran medida, estos recurren a métodos convencionales de simulación como MCMC o muestreos de Gibbs. Una solucion alternativa a este tipo de enfoque, es la aproximación de Laplace anidada integrada (INLA), por sus siglas en inglés, la cual es el punto de interés de este trabajo.

La gran mayoría de aplicaciones en las que se emplea el método INLA son análisis de carácter espacio-temporal, pues al parecer facilita ciertos aspectos del computo contrario a lo que la intuición podría sugerir. Naeeim y Rahman (2017) Estimando el riesgo relativo del dengue en Malacia peninsular usando INLA en busca de modelar el riesgo relativo del dengue, una de las enfermedades infecciosas mas comunes y de fácil transmisión entre humanos y una de las principales preocupaciones para la salud publica de Malacia, se implementa la aproximación INLA para superar las limitantes del tiempo de computo y posibles grandes errores de MCMC. En adición a los modelos de riesgo previamente definidos se establece un modelo de tendencia temporal e interacción espacio-temporal de la forma:

$$ln(R_i) = \quad + \ _i + \ '_i + \ _i + \ _i$$

Donde se interpreta como el riesgo global, $_i$ se refiere a la componente espacial del modelo, $_i$ y $_i$ capturan los efectos estructurados y el ruido relativo a la componente temporal respectivamente, por ultimo $_i$ es el parámetro de interacción espacio-temporal. Todas las especificaciones son de tipo Leroux.

La interacción espacio-temporal puede presentarse en varios tipos:

- Tipo I: Donde todos los son independientes y no existe una estructura de correlación espacio-temporal.
- Tipo II: La interacción de depende únicamente del tiempo
- Tipo III: La interacción de depende únicamente del espacio
- Tipo IV: Todos los están completamente correlacionados y dependen de una estructura espacio-temporal.

El conjunto de datos consiste en la incidencia en 86 distritos de la parte peninsular de Malacia para cada mes el año 2015, obtenida del ministerio de salud pública del gobierno. No se considera ninguna covariable.

Se realizaron 8 modelos con distintos tipos de interrelación son estimados y comparados mediante sus valores de DIC. Los modelos que no contemplan la dimensión espacio-temporal muestran el peor ajuste, se logra determinar que la inclusión del término no es muy significativa para el modelo por lo que no hay presencia de ruido en la componente temporal. El mejor modelo es aquel con una componente espacial ajustada al modelo Leroux, una temporal con especificación RW1 y el tipo II de interacción espacio-temporal. La principal ventaja de este tipo de modelos es que se puede analizar las componentes de riesgo por separado, es decir, el riesgo asociado al espacio y e asociado al tiempo.

Gracias a la discriminación de las componentes que ofrecen los modelos espacio temporales, fue posible determinar zonas que muestran significativo riesgo a través del tiempo, para este caso corresponden con territorios cercanos a Kuala Lumpur, la capital de Malacia, los cuales son Petaling, Sepang, Hulu Langat,Gombak, Klang y Hulu Selangor. Por otro lado fue posible determinar las zonas que a través del tiempo muestran tendencia a aumentar su riesgo relativo como Seberang Perai Tengah, Kinta, Kuantan y Johor Bahru. Este tipo de resultados pueden ser aplicados en el control de la enfermedad, orientando las campañas de control de hacia la misma, de tal forma que en el principio y el final de año los esfuerzos de las entidades sanitarias deberían apuntar sus esfuerzos a las zonas de Kota Setar, Kota Bharu and Kuantan.

Capítulo 2

Justificación

El siguiente proyecto tiene como objetivo determinar las zonas en las cuales se presenta el mayor riesgo de contraer la enfermedad de Malaria a partir de un conjunto de variables ambientales, sociales y otras enfermedades. Esto se va a realizar a partir de enfoques bayesianos teniendo en cuenta el método computacional INLA.

INLA facilita el algoritmo computacional del enfoque bayesiano, y es comúnmente utilizado para datos de epidemiología, además la cantidad de datos a estimar es muy robusta y los reportes de las enfermedades en Colombia no son muy precisos dados por la falta de canales de comunicación y establecimiento de políticas de control por parte del ministerio de salud.

Los enfoques bayesianos tienen la cualidad de poder ingresar a un estudio información extra muestral, dado que el INS trata de realizar un control de enfermedades, se puede apoyar este proyecto basado en los boletines epidemiológicos suministrados por el INS, posteriormente confrontarlos y determinar la ocurrencia de una enfermedad sobre el territorio colombiano basándose en variables de planificación o que tengan una estrecha relación con las políticas de ordenamiento del territorio.

la idea es identificar una metodología viable para el control de enfermedades epidemiológicas en el país teniendo en cuenta variables de tipo social, así discutiendo una premisa para establecer políticas de planificación sobre los municipios que obtengan el mayor riesgo de contraer casos de una enfermedad.

Seguramente el estudio tendrá un déficit en cuanto a los datos y la selección de variables sobre el mismo intervalo temporal, sin embargo, el presentar este tipo de estudios puede generar que se acrecenté la metodología y finalmente sea adoptada por los organismos de control de enfermedades.

2.1. Problema

En Colombia los datos de presencia de enfermedades transmitidas por vectores tienen un problema de sobredispersión, esto quiere decir que más de un 70 % de los municipios no tienen casos reportados de las enfermedades o simplemente no los ingresan en el aplicativo SIVIGILA del Instituto Nacional de Salud (INS), también dada la baja operatividad en el sistema de salud en el país, los registros de casos para algunos municipios se pueden concentrar en las grandes urbes, esto quiere decir que casos registrados en grandes áreas municipales como por ejemplo Bogotá en realidad no poseen un recuento de la enfermedad, si no este conteo hace parte de las áreas cercanas con deficiencia en el sistema de salud donde en realidad se presenta esta enfermedad. Así pues, la estimación de el riesgo que pueda tener un municipio en Colombia de que se prolifere una enfermedad puede determinarse mediante el algoritmo INLA teniendo en cuenta estas relaciones espaciales y la presencia de atípicos dado el sistema ineficiente de salud. Los enfoques bayesianos tienen la cualidad de que no solamente se realiza un modelamiento a partir de los datos en cuestión, si no que tiene en cuenta la información que un investigador experimentado pueda facilitar al proyecto.

2.2. Objetivos

Implementar modelos bayesianos con dependencia espacial para analizar el comportamiento de enfermedades en el territorio Colombiano a partir de regresiones ecológicas en datos regionalizados a nivel municipal.

2.2.1. Objetivos específicos

- Identificar la información a priori que sea significativa para la optimización de los resultados, así mismo como establecer nueva información para posteriores estudios
- Encontrar un modelo adecuado que se ajuste adecuadamente a los datos de casos registrados para cada una de las enfermedades, así como las covariables que den respuesta al riesgo que presente la enfermedad a nivel municipal.
- Determinar las políticas a sugerir mediante la significancia de las covariables en el modelado
- Comparar los resultados con los boletines epidemiológicos entregados por el INS

Capítulo 3

Marco teórico

3.1. Elementos básicos de probabilidad

Según Ross (2013), los elementos básicos de probabilidad están dados por:

- Espacio muestral (S): Es la colección de todos los resultados posibles de un experimento que no se ha realizado de antemano.
- Evento: Un evento puede definirse como los resultados posibles del experimento de interés y estos siempre pertenecen a (S). Dado que el resultado del experimento debe estar en el espacio muestral (S), se deduce que (S^c) no contiene ningún resultado y, por lo tanto, no puede ocurrir.
- Probabilidad de un evento: Esta definido por los axiomas de la probabilidad, esta representa la medida de ocurrencia de un evento y puede variar entre 0 y 1, lo que representa, en sus extremos, una certeza completa de que el evento está ocurriendo (probabilidad = 1) o no está ocurriendo (probabilidad = 0); cualquier cosa entre estos dos valores proporciona el grado de incertidumbre sobre si el evento es verdadero.

De acuerdo con Ruiz y Peña (2007) existen diferentes formulaciones de probabilidad, existe el enfoque Frecuentista y el enfoque Subjetivo que es el de mayor interés ya que formula la base para la estadística bayesiana.

Las ideas iniciales de la probabilidad surgieron relacionadas con los juegos de azar y su interpretación es básicamente frecuentista. Esta formulación frecuentista trabaja bien en muchas situaciones, pero no en todas. Una caracterıstica distintiva de la estadıstica

bayesiana es que tiene en cuenta de forma explıcita la información previa y se involucra en el análisis en forma de distribución, llamada distribución a priori. La teorıa clásica la considera básicamente para determinar tamaños muéstrales, el diseño de experimentos y, a veces, como forma de crıtica de los resultados hallados.

El enfoque de la llamada estadística frecuentista no permite incorporar de manera coherente en el análisis estadıstico la información extra-muestral disponible, se apoya únicamente en datos muéstrales observados. Si no hay datos, la estadística frecuentista esta imposibilitada para operar. Si hay muy pocos datos, la estadıstica frecuentista presenta fuertes problemas también, pues muchos de sus métodos se apoyan en resultados asintóticos, tales como la ley de los grandes números, el teorema central del lımite, y sus consecuencias, y por ello por lo general requiere de muestras “grandes” para que sus resultados sean “confiables”. En cambio, la Estadıstica Bayesiana aprovecha tanto la información que nos proporcionan los datos muéstrales ası como la información extra-muestral disponible, entendiendo por esto ultimo, de manera informal, toda aquella información relevante, además de los datos, que nos ayude a disminuir nuestra incertidumbre o ignorancia en torno a un fenómeno aleatorio de interés. (Correa Morales, 2013)

- Probabilidad condicional: Dado un espacio muestral S y dos eventos A y B que pertenecen a este espacio, se asocia una medida de probabilidad de que ocurra el evento A dado que el evento B ya se ha observado. Esta medida de probabilidad se llama probabilidad condicional. En otras palabras, si se considera una probabilidad condicional, se centra el interés en el subconjunto del espacio muestral donde pueden aparecer tanto A como B y como base para la comparación, es de interés solo el subconjunto donde B puede ocurrir en lugar de considerar el todo el espacio muestral S. (Bayes, 1763) Por lo tanto, la probabilidad condicional se puede especificar de la siguiente manera:

$$P(AjB) = \frac{P(A \setminus B)}{P(B)} \tag{1}$$

3.2. Teorema de bayes

La probabilidad condicional juega un papel importante en la estadística bayesiana, el teorema de Bayes se deduce naturalmente de esta. Teniendo en cuenta lo expresado en la ecuación (1), es posible escribir la probabilidad de la intersección entre los eventos A y B

de la siguiente manera:

$$P(A \setminus B) = P(B) \quad P(AjB) \tag{2}$$

Si se quiere establecer la probabilidad de ocurrencia de el evento B dado el evento A se aplica la ecuación (1) como:

$$P(BjA) = \frac{P(B \setminus A)}{P(A)} \tag{3}$$

Así pues reemplazando la ecuación (2) en la ecuación (3):

$$P(BjA) = \frac{P(B) \quad P(AjB)}{P(A)} \tag{4}$$

Si ahora esto se extiende para k eventos $(B_1; :::; B_k)$, y se establece que cada evento es mutuamente excluyente $B_i \setminus B_j = 0$, entonces, $\sum_{i=1}^{k} P(B_i) = 1$, para que la probabilidad de A $P(A)$ pueda ser escrita como:

$$P(A) = \sum_{i=1}^{k} P(A \setminus B_i) \tag{5}$$

Así pues reemplazando en la ecuación (4), el teorema de bayes queda expresado como:

$$P(B_ijA) = \frac{P(B_i) \quad P(AjB_i)}{\sum_{i=1}^{k} P(B_i) \quad P(AjB_i)} \tag{6}$$

Esta expresión esta dada para datos de tipo discretos, sin embargo es fácil intuir el cambio a variables continuas simplemente cambiando la sumatoria por una integral.

3.3. Distribuciones a priori y posteriori

Todos los parámetros dentro de los modelos bayesianos son estocásticos y se les asignan distribuciones de probabilidad apropiadas. Por lo tanto, un valor de parámetro único es simplemente una realización posible de los valores posibles del parámetro, cuya probabilidad está definida por la distribución a priori. La distribución a priori es una distribución asignada al parámetro antes de ver los datos. También que una interpretación de las distribuciones a priori es que proporcionan “datos” adicionales para un problema y, por lo tanto, pueden usarse para mejorar la estimación o identificación de parámetros. Para un solo parámetro, , la distribución a priori puede denotarse $g(\ )$, mientras que

para un vector de parámetros, , la distribución a priori conjunta es $g(\ )$. (A. Lawson, 2013)

La distribución posterior contiene toda la información del parámetro de interés. Sin embargo, es deseable resumir la información posterior de manera concisa. Las distribuciones posteriores son bastante menos problemáticas que las a priori. Después de todo, están completamente determinados por la elección de la a priori (junto con la verosimilitud). Sin embargo, comprender la posteriori no siempre es sencillo. Un problema común es que la posteriori $p(\ jy)$ suele ser multivariante, a veces de alta dimensión. Esto es difícil de visualizar y, por lo tanto, a menudo queremos contemplar la distribución marginal de un solo parámetro, $_i$ (Wang, Ryan, y Faraway, 2018).

3.3.1. Inferencia bayesiana

La obtención de la distribución posteriori es el proceso de inferencia bayesiana y hace una distinción fundamental entre cantidades observables y desconocidas. Considere una variable aleatoria Y, su incertidumbre se modela utilizando una distribución de probabilidad o una función de densidad (según si Y es una variable aleatoria discreta o continua, respectivamente) indexada por un parámetro genérico , entonces:

$$L(\ ) = P(Y = yj\ ) \tag{7}$$

La ecuación (7) hace referencia a la función de verosimilitud. La variabilidad en Y depende solo de la selección de muestreo (variabilidad de muestreo). En otras palabras, se supone que los datos son una muestra aleatoria de la población de estudio y la incertidumbre se origina por el hecho de que solo observamos esa muestra en lugar de todas las otras posibles. Por el contrario el parámetro es una cantidad desconocida y modelada a partir de la información a priori $P(\ )$. (Lesa re y Lawson, 2012)

Cuando la información a priori depende de una dependencia espacial o temporal, o simplemente de una estructura jerárquica, el conocimiento de se da a través de los Hiper-parámetros ( ).

El proceso para obtener la distribución marginal $p(y)$ es similar a lo que se muestra en la ecuación (6), donde se aplica la ley de probabilidades para eventos mutuamente excluyentes. En otras palabras la ecuación (6) puede significar:

Posterior / Verosimilitud apriori

Esta es la probabilidad marginal de los datos bajo el modelo. Calcular integrales como

estas es la razón por la cual la simplicidad conceptual de la inferencia bayesiana requiere bastante trabajo del que parece. Como la distribución posterior es una combinación de a priori y de verosimilitud, siempre está en algún punto entre estas dos distribuciones.

3.4. Elección de distribución a priori

Al realizar la inferencia bayesiana, la elección de la distribución a priori es un tema muy importante, ya que representa la información disponible para los parámetros de interés. En particular, hay dos aspectos que deben tenerse en cuenta: (A. Lawson, 2013)

(i) El tipo de distribución, que debe ser representativo de la naturaleza de los parámetros

(ii) Los hiper-parámetros, que harían la distribución más o menos informativa, proporcionando así el nivel de información (o ignorancia) disponible para los parámetros.

3.4.1. Tipo de distribución

De acuerdo con Blangiardo y Cameletti (2015) el tipo de distribución depende de la naturaleza del parámetro de interés, por ejemplo si se tiene datos proporcionales la incertidumbre sobre el parámetro debe estar representada por una distribución que varía entre 0 y 1, si el parámetro de interés es una variable simétrica continua esta debe a los R en el intervalo $[1; 1]$, Si el parámetro de interés es una variable positiva continua esta debe a los N en el intervalo $[0; 1]$. Existen modelos de uso común, que describen la elección típica de distribución previa y el proceso inferencial bayesiano que conduce a la distribución posterior en los cuales se puede encontrar: Poisson–Gamma, Normal–Normal y Binomial-Beta

3.4.2. Conjugada

Algunas combinaciones de distribuciones a priori junto con la verosimilitud, conducen a la misma familia de distribución en la posteriori que para la distribución a priori. Esto puede facilitar la inferencia ya que la forma de la posteriori se deducirá de la especificación a priori.

Ruiz y Peña (2007) simplifica en una tabla algunas de las familias conjugadas:

Familia paramétrica	Familia conjugada
Bernoulli( )	Beta (j ,)
Poisson( )	Gamma(j ,)
Geométrica( )	Beta (j ,)
Exponencial( )	Gamma(j ,)
Binomial(,)	Beta (, | ,)

Cuadro 3.1: Distribución de Familias conjugadas

3.4.3. Informativa o No informativa a priori

Una vez que se ha especificado la forma funcional de la distribución a priori, la definición de sus parámetros debe ser informada por cualquier conocimiento disponible. Este siempre ha sido un tema crítico en la inferencia bayesiana y una fuente de grandes críticas por parte de la escuela frecuentista.

No informativa: La propiedad de permitir ingreso de información extra muestral en un análisis estadístico es una característica muy particular de la estadística bayesiana, aunque, a veces los investigadores no pueden o no desean hacer uso de información a priori.

Esta información que expresa falta de información o ignorancia según Lesa□re y Lawson (2012) la información a priori que expresa falta de información o ignorancia se llamó inicialmente una distribución a priori no informativa. Hoy en día, se ha sostenido que la información a priori siempre lleva algo de información, incluso si pretende representar ignorancia, y que en el mejor de los casos uno puede esperar que la información equivalente sea mínima.

Como cada a priori implica que cierta información externa se infiltra en el análisis, uno debe mostrar en la práctica que el prior elegido es de hecho mínimamente informativo. Esto no siempre es fácil, especialmente cuando se trata de un modelo complejo que involucra muchos parámetros.

Cuando un a priori incorrecto hace que el posteriori sea incorrecto, el análisis bayesiano está en problemas, ya que el posteriori ya no puede proporcionar medidas de resumen.

Informativa: Box y Tiao (1973) argumentaron que casi nunca estamos en un estado de ignorancia absoluta. El desafío es incorporar ese conocimiento previo aunque sea escaso en un marco probabilístico. La pregunta es si incluir las creencias de un experto o de una comunidad de expertos. Esto no es importante para la aplicación del teorema de Bayes, pero es importante para la aceptación y la utilidad del análisis bayesiano.

Lo que se puede realizar con la información a priori obtenida puede llegar a ser el

formalizar el uso de datos históricos como información previa utilizando el poder a priori, el uso de antecedentes clínicos, que son distribuciones a priori basadas en datos históricos o en conocimiento experto además de antecedentes que se basan en reglas formales que expresan escepticismo y optimismo a priori. El conjunto de distribuciones a priori que representan el conocimiento previo se denominan antecedentes subjetivos o informativos.

3.5. Aproximación de Laplace anidada integrada (INLA)

Rue et al. (2009) muestra que, al usar una aproximación de Laplace anidada integrada y su versión simplificada, se puede calcular directamente aproximaciones muy precisas a la distribución marginal posterior. El principal beneficio de estas aproximaciones es computacional, cuando los algoritmos MCMC necesitan horas y días para ejecutarse, las aproximaciones proporcionan estimaciones más precisas en segundos y minutos. Otra ventaja del enfoque es su generalidad, que permite realizar análisis bayesianos de forma automática y simplificada, y calcular criterios de comparación de modelos y diversas medidas predictivas para que los modelos puedan compararse y el modelo en estudio pueda ser cuestionado.

3.5.1. Aproximación de Laplace

Un enfoque alternativo para la integración de MC basada en simulación es la aproximación analítica con el método de Laplace, esto puede ser visto con más detalle en Barndor□-Nielsen y Cox (1989). Suponga que se esta interesado en calcular la siguiente integral:

$$\int f(x)dx = \int exp(log f(x))dx \tag{8}$$

Donde

- $f(x)$: función de densidad de una variable aleatoria x
- $log f(x)$: será representada por la serie de Taylor en x_0

$$log f(x) \quad log f(x_0) + (x \quad x_0)\frac{@log f(x)}{@x}j_{x=x_0} + \frac{(x \quad x_0)^2}{2}\frac{@^2 log f(x)}{@x^2}j_{x=x_0} \tag{9}$$

Si x_0 se aproxima a, $x^* = argmax_x \log f(x)$, entonces $\frac{@log f(x)}{@x} j_{x=x^*} = 0$ la aproximación se convierte en:

$$log f(x) \quad log f(x^*) + \frac{(x \quad x^*)^2}{2} \frac{@^2 log f(x)}{@x^2} j_{x=x^*} \tag{10}$$

La ecuación (8) se reescribe así:

$$\int f(x)dx \quad \int exp(log f(x^*) + \frac{(x \quad x^*)^2}{2} \frac{@^2 log f(x)}{@x^2} j_{x=x^*})dx$$

$$\int f(x)dx = exp(log f(x^*)) \int exp(\frac{(x \quad x^*)^2}{2} \frac{@^2 log f(x)}{@x^2} j_{x=x^*})dx \tag{11}$$

así pues el integrando se puede asociar a la función de densidad de probabilidad de una distribución Normal.

$$;^2(x) = \frac{1}{p\overline{2}} \int_{1}^{x} exp(\frac{u}{2^2})du$$

Estableciendo $^2 = 1=\frac{@^2 log f(x)}{@x^2} j_{x=x^*}$, se obtiene de la ecuación (11):

$$\int f(x)dx \quad exp(log f(x^*)) \int exp(\frac{(x \quad x^*)^2}{2^2})dx$$

Donde el integrando es el núcleo de una distribución Normal con media igual a x^* y varianza ^{2}, más precisamente, la integral evaluada en el intervalo $(\ ; \ )$ se aproxima por:

$$\int f(x)dx \quad f(x^*) p\overline{2^2}(\ (\ ) \quad (\ )) \tag{12}$$

Donde $(:)$ es la función de densidad normal$(x^* ; ^2)$

3.5.2. Entorno de INLA

El algoritmo se basa en la definición de la distribución de probabilidad de la variable observada $Y = (y_1; y_2; :::; y_n)$, mas específicamente se asigna una distribución dada por $'_i$ a cada y_i la cual será descrita por un predictor $_i$, sin embargo esta relación no será directa sino que se plantea a través de una función de enlace de tal forma que $f('_i) = _i$ y el predictor tendrá la forma:

$$_i = _0 + \sum_{m=1}^{M} _m x_{mi} + \sum_{j=1}^{J} f_1(z_{li}) \tag{13}$$

Donde $_0$ representa el intercepto del modelo, los coeficientes representan la relación lineal entre el predictor y un conjunto de variables explicativas $X_i = (x_1; x_2; x_3 ::: ; x_n)$, finalmente el componente $f_n(z_n)$ es una o varias funciones en términos de un subconjunto de covariables $Z_i = (z_1; z_2 ::: ; z_n)$ las cuales, dependiendo de su definición, pueden capturar relaciones no lineales, tendencias temporales o efectos espaciales, luego todos los efectos observables y no observables serán capturados en un vector $= f_0; ; fg$. Esta es la razón por la cual el método INLA puede ser aplicado en diversas disciplinas. (Blangiardo y Cameletti, 2015)

Adicionalmente, para la estimación se precisa la definición de un conjunto de hiperparametros $= (_1; _1; ::: ; _k)$ asociados a distribución *a priori* bajo la cual se llevará a cabo el algoritmo. De tal forma que si se asume independencia en las n observaciones la distribución de probabilidad estará dada por:

$$P(Y j \; ; \;) = \prod_{i=1}^{n} P(y_i j \; _i; \;) \tag{14}$$

Bajo esta perspectiva, el objetivo de la estimación Bayesiana con el método INLA se centra en encontrar la distribución marginal posterior de cada uno de los elementos del vector

$$\int P(\; _i j y) = P(\; _i j \; ; Y) P(\; \; j Y) d(\; \;) \tag{15}$$

Para lo cual el algoritmo desarrolla dos tareas, la estimación de $P(\; \; j Y)$ donde se emplean los hiper-parámetros de la distribución *a priori*, para posteriormente obtener $P(\; _i j \; ; Y)$ y así calcular finalmente $P(\; _i j y)$.

El enfoque INLA explota los supuestos del modelo para producir una aproximación numérica a los posteriores de interés basada en el método de aproximación de Laplace introducido en la sección previa.

3.6. Modelos

3.6.1. Modelos lineales generalizados

Los modelos lineales generalizados (GLM), proporcionan una familia unificadora de modelos lineales que se usa ampliamente en el análisis práctico de regresión. Los GLM generalizan la regresión lineal ordinaria al permitir que los modelos se relacionen con la variable respuesta a través de una función de enlace y al permitir que la magnitud de la

varianza de cada medición sea una función de su valor predicho. Por lo tanto, estos modelos permiten describir variables respuesta que tienen una distribución de error diferente a la normal. Evitan tener que seleccionar ciertas transformaciones de los datos para lograr los objetos posiblemente conflictivos de normalidad, linealidad y=o homogeneidad de varianza. Los GLM comúnmente utilizados incluyen regresión logística para datos binarios y regresión de Poisson o regresión binomial negativa para datos de conteo. (Wang et al., 2018)

3.6.1.1. Modelo lineal generalizado de Poisson

Blangiardo y Cameletti (2015) explica que esta se usa cuando la variable resultado representa datos de conteo de modo que puede asumir valores entre (0*; 1*). El parámetro de interés es el número promedio de eventos $E(y_i) = \ _i$ y la función de enlace es el logaritmo, de modo que:

$$\begin{aligned} _i &= log(\ _i) = X_i \\ _i &= Exp(X_i\ ) \end{aligned} \tag{16}$$

En otras palabras, la función exponencial transforma los valores continuos obtenidos X_i aplicado en el rango de valores de X_i

$$\begin{aligned} y_i \quad & Poisson(\ _i); \\ _i = log(\ _i) = \ _0 &+ \sum_{m=1}^{M} \ _m X_{im} \end{aligned} \tag{17}$$

Para completar el modelo, se especifican los antecedentes de típicamente como distribuidos normalmente, caracterizada por una gran variabilidad si no hay información de estudios previos u opiniones de expertos.

Cuando se utiliza la regresión Poisson el interés se basa en Tasas o Riesgos relativos más que en el número promedio de casos indicados como $_i$ en la ecuación (17). Para cambiar la escala se puede usar un desplazamiento como factor de corrección en la especificación del modelo.

$$_i = log(\ _i) = \ _0 + \sum_{m=1}^{M} \ _m X_{im} + log(offset)_i \tag{18}$$

Así que el logaritmo del riesgo relativo se obtiene como:

$$log(\frac{i}{offset}) = \ _0 + \sum_{m=1}^{M} \ _m X_{im} \tag{19}$$

Y los coeficientes pueden interpretarse en la escala de riesgo en lugar de escala absoluta. En este caso exponiendo el intercepto devuelve la tasa de referencia ($_0$) mientras que la exponencial de los demás representa un cambio en el riesgo relativo para un cambio de unidad en el predictor.

3.6.1.2. Modelo lineal generalizado binomial para datos de conteo

Suponga que se observan proporciones como respuesta $y_1; :::; y_N$ de poblaciones binomiales con proporciones $_1; :::; \ _N$ y sus correspondientes tamaños muestrales $_1; :::; \ _N$. Asociado con la i-ésima observación hay un vector de covariables x_i y la proporción $_i$ es encadenada a las covariables xi por medio del modelo logístico. Por lo tanto, se puede suponer que dada la probabilidad de un caso $_i$, y_i se distribuye independientemente como: (Correa Morales, 2013)

$$\begin{aligned} y_i \quad & bin(\ _i; \ _i) \\ log(\frac{i}{1 \quad _i}) &= x_i^T \end{aligned} \tag{20}$$

La verosimilitud viene dada por:

$$L(y_i j \ _i; \ _i) = \prod_{i=1}^{m} \binom{i}{y_i} \ _i^{y_i} (1 \quad _i)^{\ i \ y_i} \tag{21}$$

Es habitual que se elija una función de enlace adecuada para la probabilidad a un predictor lineal. Lo más común sería un enlace logit para que se cumpla: (A. Lawson, 2013)

$$_i = \frac{exp(\ _i)}{1 \quad exp(\ _i)} \tag{22}$$

Aquí, se prevé la especificación del modelo dentro de $_i$, para incluir componentes espaciales y no espaciales.

3.6.1.3. Modelo lineal generalizado binomial negativo (NB)

De acuerdo con A. Zuur et al. (2009) en algunos conjuntos de datos al modelarlos con una distribución Poisson se encuentran problemas de sobre dispersión, en consecuencia, todos los errores estándar se corrigieron multiplicándolos con la raíz cuadrada del modelo cuando se aplica un modelo cuasi-Poisson. Un enfoque alternativo es aplicar el modelo binomial negativo. Al igual que para los GLM de Gauss y Poisson, se especifica el modelo con tres pasos. El NB GLM viene dado por:

$$y_i = NB(\ _i; k) \tag{23}$$

y_i tiene función de distribución binomial negativa con media $_i$ y parámetro k. Por definición, la varianza de y_i es $_i + \frac{^2_i}{k}$

$$log(\ _i) = \ (x_{i1}; :::; x_{iq}) \tag{24}$$

La parte sistemática está dada por $(X_{i1}; :::; X_{iq}) = \ + \ _1X_{i1} + ::: + \ _qX_{iq}$. Existe un vínculo logarítmico entre la media de y_i y la función de predicción $(x_{i1}; :::; x_{iq})$. El enlace logarítmico asegura que los valores ajustados sean siempre no negativos.

3.6.2. Modelos de ceros inflados

En estadística, a menudo se usa una regresión inflada a cero para el problema de modelar datos con ceros en exceso. El modelo se basa en una distribución de probabilidad inflada a cero, es decir, una distribución que permite observaciones frecuentes de valor cero. La situación de inflación cero ocurre a menudo en los datos de conteo de modelado. INLA permite a los usuarios ajustar modelos inflados a cero con distribuciones de Poisson, binomial, binomial negativo y beta-binomial. (Wang et al., 2018)

Los modelos de cero inflado proporcionan un enfoque de modelado mixto para modelar los ceros en exceso además de permitir la sobredispersión. En particular, hay dos posibles procesos de generación de datos para cada observación, en los que se utiliza el resultado de un ensayo de Bernoulli para determinar el proceso. Por sí solos, los datos truncados a cero no son necesariamente un problema. Es la suposición subyacente de Poisson y las distribuciones binomiales negativas lo que puede causar un problema ya que estas distribuciones permiten ceros dentro de su rango de valores posibles. Si la media es pequeña y la variable respuesta no contiene ceros, entonces los parámetros estimados y los errores estándar obtenidos por GLM pueden estar sesgados. Si la media de la variable respuesta

es relativamente grande, ignorando el problema de truncamiento, es poco probable que la aplicación de un modelo lineal generalizado (GLM) de Poisson o binomial negativo (NB) cause un problema. En tales casos, los parámetros estimados y los errores estándar obtenidos por Poisson GLM y Poisson GLM truncado tienden a ser similares (lo mismo ocurre con los modelos binomiales negativos). (A. Zuur et al., 2009)

En la investigación ecológica, debe buscar mucho para encontrar datos truncados a cero. La mayoría de los datos de conteo son cero inflados. Esto significa que la variable respuesta contiene más ceros de lo esperado, en función de la distribución binomial negativa o de Poisson. Un histograma simple o diagrama de frecuencia con un pico grande en cero da una advertencia temprana de una posible inflación cero.

3.6.2.1. Modelo Poisson para ceros inflados

Para hablar de modelos ZIP (por sus siglas en inglés) es necesario hacer una referencia a la función de distribución de Poisson dada por:

$$P(Y = yj\ ) = \frac{^{y}e}{y!} \tag{25}$$

Esta distribución esta pensada para describir la probabilidad para la abundancia de observaciones dada una media conocida del fenómeno estudiado. Como características principales los valores de la media y la varianza de la distribución son iguales y corresponden con $E(y) =$ y $var(y) =$.

Sin embargo en la practica es desconocida por lo que se debe modelar a partir de un conjunto de covariables lo que motiva la estimación de modelos lineales generalizados Poisson (A. F. Zuur, Ieno, y Saveliev, 2017).

$$\begin{aligned} y_i &\quad Poisson(\ _i) \\ E(y_i) = \ _i &= var(y_i) \\ Log(\ _i) = \ _0 + \ _1x_1 + \ _2x_2 + :::: + \ _nx_n & \end{aligned} \tag{26}$$

Para extender este tipo de modelos a datos con exceso de ceros, se parte de una distribución Bernoulli, donde el éxito se define como la obtención de un cero en las observaciones y el fracaso cualquier otro valor, la probabilidad de éxito estará dada por $_i$. Así se modelan por separado las componentes del modelo, con una probabilidad asociada a la obtención de ceros, dada por el valor de $_i$ y otra probabilidad asociada a un proceso Poisson que explique los valores de conteo encontrados en las observaciones, entonces

tenemos una distribución ZIP de la forma (A. F. Zuur et al., 2017):

$$P(Y_i = y_i | \ _1; \ _2; \ _1) = \begin{cases} \ _i + (1 \quad \ _i)e^{\ \ i} & y_i = 0 \\ (1 \quad \ _i) f_{poisson}(Y | \ _1; \ _2) & y_i > 0 \end{cases} \tag{27}$$

Siguiendo el esquema de la regresión Poisson se construye el modelo de regresión ZIP

$$y_i \quad ZIP(\ _i) \tag{28}$$

$$E(y_i) = (1 \quad) \ _i$$

$$var(y_i) = (1 \quad)(\ _i + \quad ^2_i)$$

$$Log(\ _i) = \ _0 + \ _1 x_1 + \ _2 x_2 + \cdots + \ _n x_n$$

3.7. Modelamiento Espacial y Mapeo de Enfermedades

A continuación, se consideran las formas básicas de datos que surgen en los estudios de mapeo de enfermedades y se describen modelos comunes para la variación en el riesgo de enfermedad. De acuerdo con A. B. Lawson et al. (2016) con los datos de eventos de caso, los datos toman la forma de coordenadas geográficas, como la latitud y la longitud, de una residencia. Sin embargo, a menudo la coordenada de residencia (x, y) puede no estar disponible debido a restricciones de confidencialidad médica. En ese caso, los datos de recuento agregados suelen ser la única forma disponible. Para algunos estudios, en particular aquellos que se ocupan de supuestos peligros para la salud, el objetivo es una resolución espacial fina, por lo que las ubicaciones de las direcciones pueden estar en el nivel de resolución esencial para que el estudio continúe. En lo que sigue, se deben examinar modelos de probabilidad comunes tanto para eventos de casos como para datos de recuento y también se debe considerar cómo estos pueden simplemente extenderse para incluir heterogeneidad adicional en el riesgo de enfermedad.

La representación de los casos de incidencia de una enfermedad pueden ser vistos como mapas de puntos y recuentos por área hasta complejos modelos que describen la estructura del evento de una enfermedad. Para evaluar el estado de un área con respecto a la incidencia de la enfermedad, es conveniente intentar primero evaluar qué incidencia de la enfermedad debe ser “esperada” localmente en el área y luego comparar la incidencia observada con la incidencia “esperada”. Este enfoque se ha utilizado tradicionalmente para el análisis de recuentos dentro de los tratados y también se puede aplicar a los mapas de eventos de casos. (A. Lawson et al., 1999)

3.7.1. Datos de área

De acuerdo con Cressie (1992) define los datos de área como un enrejado o "lattice" lo cual evoca una idea de puntos espaciados en R^d vinculados a los vecinos más cercanos indistintamente del grado de estos. Estos no siempre siguen un comportamiento predecible en su desplazamiento y sus relaciones no siempre son obvias por su geometría. La perspectiva lattice o grilla, es la estructura predominante en la econometrıa espacial, así como la mas usada en ciencias sociales, requiere la formación de la estructura de vecindad para cada observación, es decir, la topologıa u ordenación espacial de los datos en la forma de una matriz de pesos espaciales.

Por medio de análisis espacial, se interpretaran los datos de área que se tienen: municipios de Colombia (datos espaciales), debido a que están referidos a una superficie, haciendo hincapié a un fenómeno específico en este caso conteo de casos para enfermedades epidemiológicas; siendo la idea mostrar el comportamiento de los mismos para tomar decisiones futuras y revelar patrones y anomalías que no son inmediatamente obvias.

3.7.1.1. Matrices de vecindad espacial

La estructura espacial suele expresarse formalmente a través de una matriz de interacciones espaciales, también llamada "matriz de pesos, ponderaciones, distancias o contactos espaciales". En esta matriz, cada unidad espacial se representa a la vez mediante una fila y una columna. En cada fila, los elementos no nulos de las columnas se corresponden con las unidades espaciales contiguas. (Yrigoyen, 2003)

El concepto de vecindad espacial o matriz de proximidad es útil en la exploración de datos de área. El elemento (i,j) de una matriz de vecindad espacial W, denotado por w_{ij}, conecta espacialmente las áreas i y j de alguna manera, $(i; j)$ $(1; :::; n)$. W define una estructura de vecindario en toda la región de estudio, y sus elementos pueden verse como pesos. Más peso se asocia con j^0s más cerca de i que aquellos más lejos de i. La definición de vecindad más simple la proporciona la matriz binaria donde w_{ij} = 1 si las regiones i y j comparten algún límite común, tal vez un vértice, y w_{ij} = 0 de lo contrario. Habitualmente, w_{ij} se establece en 0 para i=1,...,n . Tenga en cuenta que esta elección de definición de vecindad da como resultado una matriz de vecindad espacial simétrica. (Moraga, 2019)

3.7.2. Mapeo de enfermedades

El mapeo de enfermedades se usa comúnmente con datos de área para evaluar el patrón espacial de una enfermedad en particular e identificar áreas caracterizadas por un riesgo relativo inusualmente alto o bajo. Los datos en este caso son de naturaleza discreta, ya que son recuentos de enfermedades o muertes en cada área.(Blangiardo y Cameletti, 2015)

La detección y el control efectivos de las enfermedades en humanos y animales por parte de las autoridades sanitarias deben tener en cuenta los patrones espaciales de la aparición de la enfermedad y los factores de riesgo asociados. Esto incluye una recopilación, gestión y análisis de datos eficientes. La integración de la funcionalidad SIG en la mayoría de los sistemas modernos de información sobre enfermedades refleja el reconocimiento de la importancia de la dimensión espacial del control de enfermedades. La funcionalidad analítica de tales sistemas está típicamente restringida a la producción de mapas descriptivos, a menudo basados en agregaciones de datos a nivel de alguna área administrativa, como el distrito o la provincia. Al aplicar métodos de análisis espacial como parte del manejo de la enfermedad en lugar de como una herramienta de investigación, los resultados deben interpretarse con cierta precaución, particularmente debido a posibles errores y sesgos. (Pfei□er et al., 2008)

3.7.2.1. Tasas estandarizadas

Un enfoque simplista para los datos de recuento de enfermedades podría considerar el cálculo de la razón estandarizada de mortalidad (o morbilidad) o índice de radio estandarizado.

La justificación para el uso de SIR puede respaldarse mediante el análisis de modelos de verosimilitud con la multiplicación del riesgo esperado.

SIR es simplemente la relación entre el número de casos observados y_i y el número de casos esperados E_i en el área i-ésima ($i = 1;:::;n$), típicamente calculado usando tasas de referencia estandarizadas por edad y sexo r_j ($j = 1;:::;J$ combinaciones de categorías de edad y sexo) y la población del censo cuenta Pop_{ij}. (Waller y Gotway, 2004)

$$r_j = \frac{\sum_{i=1}^{n} y_{ij}}{\sum_{i=1}^{n} Pop_{ij}} \tag{29}$$

$$E_i = \sum_{j=1}^{J} Pop_{ij} r_j \tag{30}$$

$$SIR_i = \frac{y_i}{E_i} \tag{31}$$

3.7.3. Análisis ecológico

Aunque los SIR pueden ser útiles en algunos entornos, en regiones con poblaciones pequeñas o enfermedades raras, los recuentos esperados pueden ser muy bajos y los SIR pueden ser engañosos e insuficientemente confiables para informar. Por lo tanto, se prefiere estimar el riesgo de enfermedad mediante el uso de modelos que permitan tomar prestada información de áreas vecinas e incorporar información de covariables que resulte en la suavización o reducción de valores extremos basados en tamaños de muestra pequeños. (Moraga, 2019)

De acuerdo con A. B. Lawson et al. (2001) el término “análisis ecológico” se usa a menudo para describir investigaciones epidemiológicas que asocian la aparición de una enfermedad y los posibles factores de riesgo. Aumentar la conciencia pública sobre las implicaciones para la salud en el medio ambiente ha llevado a un creciente número de estudios de regresión ecológica en la literatura sobre epidemiología.

El análisis y la interpretación pueden ser muy complicados por varias características inherentes a los datos y diseño de estudio.

Cuando los factores de riesgo están disponibles y el objetivo del estudio es evaluar su efecto sobre el riesgo de enfermedad (o muerte), se pueden especificar modelos de regresión ecológica, simplemente extendiendo el procedimiento descrito en la sección anterior para el mapeo de la enfermedad. (Blangiardo y Cameletti, 2015)

Capítulo 4

Metodología

El desarrollo del proyecto esta orientado a distintas etapas, y cada una de estas subdivididas en procesos los cuales se enfrentan en conjunto a la resolución del objetivo general del proyecto.

- Como primera etapa se puede encontrar la delimitación del área de estudio, allí se determinará el nivel geográfico a trabajar, ya sea local, regional o nacional. Esto tendrá gran influencia en el resultado de los datos puesto que si el nivel es sumamente detallado los resultados serán de igual forma además si el nivel esta ampliamente generalizado se podría incurrir en errores en cuanto a la selección del Datúm.
- Para la segunda etapa esta orientada a la construcción de la base de datos, esta etapa condiciona automáticamente el nivel de detalle del estudio puesto que el nivel geográfico estará condicionado por la recolección de la información. Dentro de esta etapa se debe realizar un filtrado y limpieza de los datos, para que la información se localice tanto en el mismo espacio como en la misma temporalidad.
- La tercera etapa esta encaminada a el tratamiento estadístico de los datos, esta siendo la de mayor importancia puesto que es la que validará los distintos modelos y el enfoque bayesiano para el mapeo de enfermedades. Allí se realizará tanto un análisis exploratorio como uno confirmatorio y será el pilar para poder determinar la incidencia de covariables sobre cada una de las enfermedades.
- La cuarta etapa es la interpretación de los resultados cuantitativos entregados por la etapa anterior. Aquí se debe esclarecer la incidencia de las enfermedades y como mediante políticas públicas se puede disminuir o atenuar el número posible de contagios de las enfermedades.

4.1. Área de estudio

La región en la cual se determinará el riesgo relativo de infección por enfermedades epidemiológicas es la República de Colombia en su parte continental, el problema se tratará a una escala Municipal.

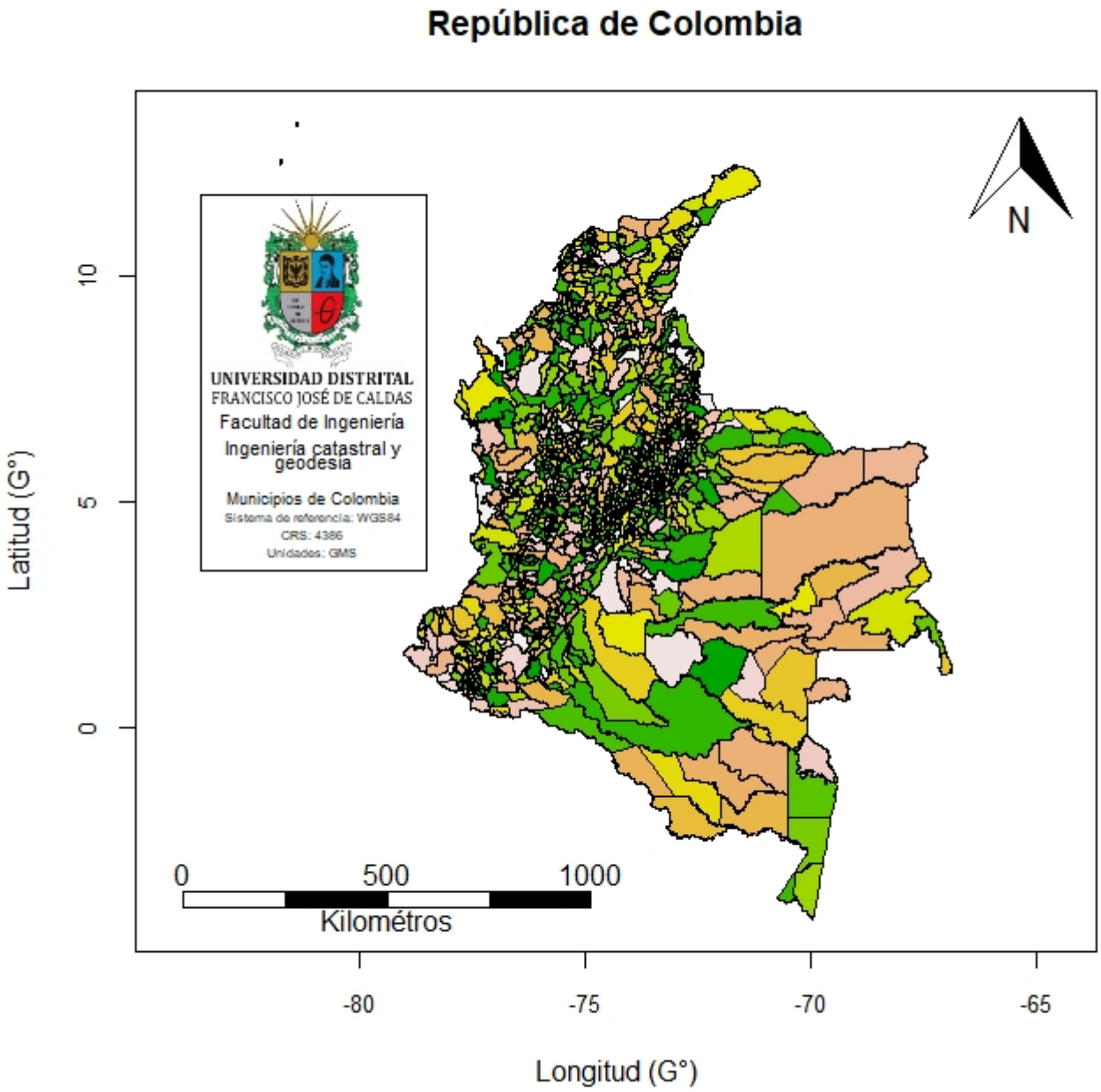


Figura 4.1: Área de estudio, división política de Colombia, software R

4.2. Base de datos

La adquisición de la base de datos se obtuvo de distintas fuentes, en donde lo importante es poder unificarlas bajo la misma extensión espacial, para esto se tuvo en cuenta el *Identificador* de cada municipio respecto al DANE y así pues fue como se consolido una fuente de información geográfica.

Variable	Tipo	Descripción	Fuente de información	Intervalo de tiempo
Cod_Dane	Ordinal	Código Dane del municipio	Dane	N.A
Nom_Depto	Nominal	Nombre del departamento	Dane	N.A
Nom_Mpio	Nominal	Nombre del municipio	Dane	N.A
Pob_hombre	Discreta	Población de hombres	Proyecciones constantes	2015
Pob_mujer	Discreta	Población de mujeres	Proyecciones constantes	2015
Pob_urbana	Discreta	Población urbana	Dane proyecciones	2015
Pob_rural	Discreta	Población rural	Dane proyecciones	2015
Pob_total	Discreta	Población total	Dane proyecciones	2015
Sex_mujer_malari	Discreta	Mujeres infectadas con malaria	SIVIGILA	2007 2015
Sex_hombre_malari	Discreta	Hombres infectados con malaria	SIVIGILA	2007 2015
Sex_NoRep_malari	Discreta	Infectados con malaria sin reporte	SIVIGILA	2007 2015
Sex_NoDef_malari	Discreta	Infectados con malaria No definidos	SIVIGILA	2007 2015
Mala_indigena	Discreta	Indígenas infectados con malaria	SIVIGILA	2007 2015
Mala_afro	Discreta	Afrodescendientes infectados con malaria	SIVIGILA	2007 2015
Mala_raizal	Discreta	Raizales infectados con malaria	SIVIGILA	2007 2015
Mala_pale	Discreta	Palenqueros infectados con malaria	SIVIGILA	2007 2015
Mala_rom	Discreta	Romanies infectados con malaria	SIVIGILA	2007 2015
Malaria_vivax	Discreta	Cantidad de infectados con malaria vivax	SIVIGILA	2007 2015
Malaria_falciparum	Discreta	Cantidad de infectados con malaria falciparum	SIVIGILA	2007 2015
Malaria_complicada	Discreta	Cantidad de infectados con malaria complicada	SIVIGILA	2007 2015
Malaria	Discreta	Cantidad de infectados con malaria	SIVIGILA	2007 2015
Dengue	Discreta	Cantidad de infectados con Dengue	INS	2007 2017
Chicu_hom	Discreta	Cantidad de hombres infectados con chikungunya	INS	2014 2015
Chicu_muj	Discreta	Cantidad de mujeres infectadas con chikungunya	INS	2014 2015
chikungunya	Discreta	Cantidad total de infectados con chikungunya	INS	2014 2015
Zika	Discreta	Cantidad de infectados con Zika	INS	2015 2017
Cober_alcanta	Razón	Cobertura de alcantarillado	SIGOT	2005 2015
Cober_acued	Razón	Cobertura de acueducto	SIGOT	2008 2015
Bosque	Razón	Dinámica de cobertura de Bosque media	IDEAM	2010 2017
Precipitación	Categórica	Precipitación mensual	IDEAM	2007 2015
Altura_media	Continua	Altura media sobre el nivel del mar	SRTM	N.A

Cuadro 4.1: Descripción de la Base de datos

Como primera fuente de información, se acudió a el Dane, allí se extrajo los datos correspondientes a las poblaciones proyectadas para cada municipio, a nivel urbano, rural y total, además de eso de acuerdo con el censo de 2005 se determino una proporción de población masculina y femenina y de acuerdo con estos porcentajes, se establecieron también para el año 2015. Así pues se tiene una proyección constante de hombres y mujeres respecto a su total.

Posteriormente se obtuvo información de la bodega de datos de SISPRO, en donde se accedía a los datos del Sistema Nacional de Vigilancia en Salud Pública para poder descargar los datos que conciernen a todo lo referente con la enfermedad de Malaria

como se puede observar en la tabla (4.1)

Como tercera fuente de información podemos observar los datos abiertos del Instituto nacional de salud, allí se encuentran los datos de tres enfermedades epidemiológicas como lo son Dengue, Chikungunya y Zika.

Como fuentes de información geográficas se acude a el SIGOT que es el sistema de información para ordenamiento territorial manejado por el *IGAC*, de allí se descargaron los datos correspondientes a las coberturas en infraestructura de acueducto y alcantarillado. Además del SIGOT, se descarga un DEM a 90 metros de precisión para todo Colombia, este es facilitado por La Misión Topográfica Shuttle Radar.

Capítulo 5

Análisis exploratorio de datos espaciales

5.1. Estadísticas descriptivas

En el análisis de datos, se contrarestra modelos en los cuales se quiere conocer si la variable dependiente es explicada por un conjunto de covariables establecidas, sin embargo, en muchas ocasiones estas covariables no son del todo "determinantes", es por esto que el estudio se denomina exploratorio y se deben utilizar ciertas técnicas para encontrar algún sentido en los datos. (Yrigoyen, 2003)

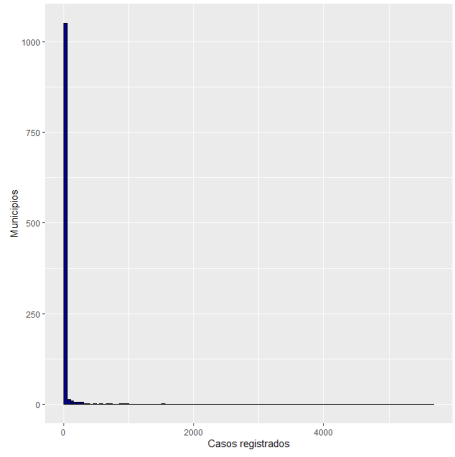


Figura 5.1: Histograma de casos reportados de Malaria para 2015, software R

En la figura (5.1) se puede observar la naturaleza de la variable y asociarla a un conteo (como se había mencionado en la base de datos).

La variable respuesta cuyo comportamiento se evidencia en la tabla (5.1) tiene un porcentaje cercano al 75 % de observaciones con valor 0, sin embargo, tiene un valor máximo extremadamente alto, de 5722. No vale la pena interpretar las medidas de tendencia central por las características de la variable, es por esto que se propone un modelo para ceros inflados como alternativa para lograr resultados adecuados.

Mínimo	1er cuartil	Mediana	Media	3er cuartil	Máximo
0	0	0	35.99	1	5722

Cuadro 5.1: Resumen estadísticas descriptivas para la incidencia de malaria 2015.

Las características mencionadas no son altamente informativas, por lo que se procede a analizar la distribución espacial de la variable. En la figura (5.2) se evidencia que los municipios con alta incidencia de malaria se agrupan, así como aquellos con cero casos "observados" lo que puede indicar un comportamiento de tipo cluster. Las zonas que presentan conteos mas altos se ubican en la costa pacifica, los llanos orientales y la parte de la Amazonia, mientras que en la región Andina no se aprecia parecencia de la enfermedad.

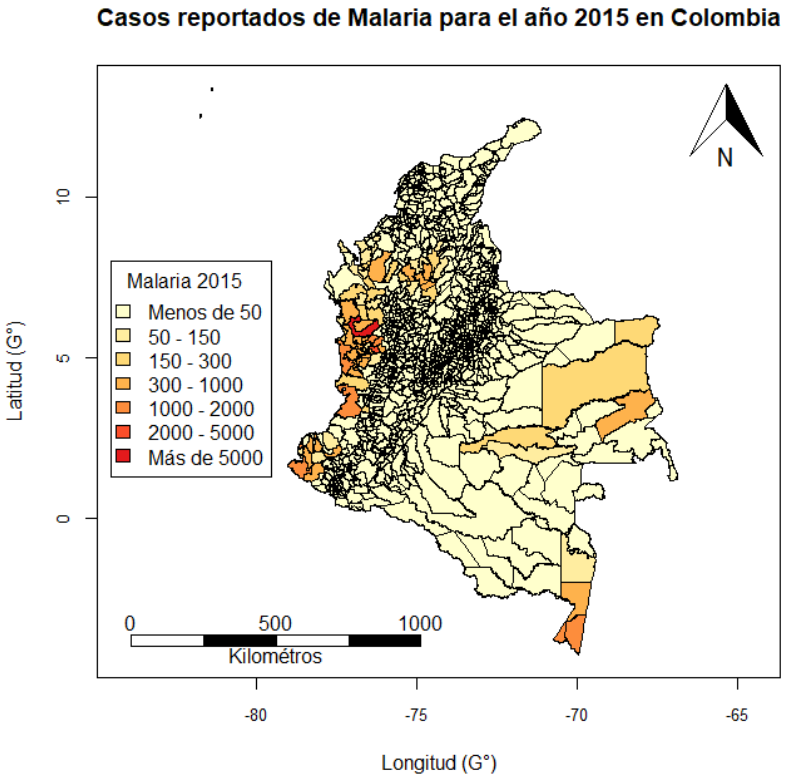


Figura 5.2: Box map de casos reportados de Malaria para 2015 en el territorio colombiano a nivel municipal, software R

Dentro de este análisis se tendrá en cuenta la cantidad de casos reportados por las demás enfermedades dado que si estas se encuentran en la misma situación de cluster que la variable respuesta Malaria, se procede a verificar la relación de estas con diagramas de dispersión.

En la figura (5.3) se puede observar la naturaleza de la variable asociada con los casos reportados de Dengue y asociarla a un conteo (como se había mencionado en la base de datos). También se observa que el histograma es asimétrico a la derecha o positivo, esto puesto que la media es superior a la mediana además implica que hay más valores distintos a la derecha de la media, también observando una gran concentración de valores cero muy cercanos al 50 %.

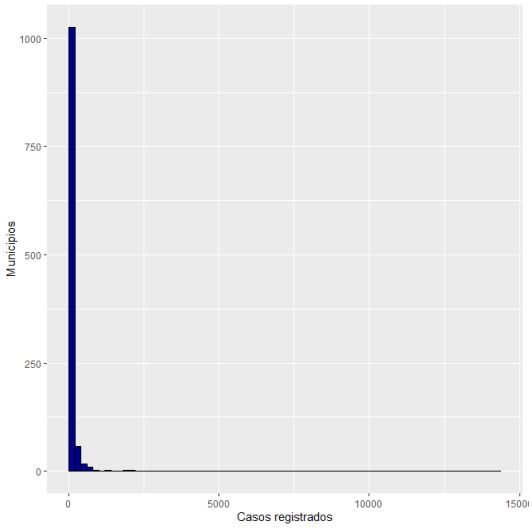


Figura 5.3: Histograma de casos reportados de Dengue para 2015, software R

Para esclarecer el anterior histograma se observa en la tabla (5.2) el comportamiento de la variable Dengue esta tiene un porcentaje cercano al 50 % de observaciones con valor 0, sin embargo, tiene un valor máximo extremadamente alto, de 14535, con un rango mucho más amplio de casos reportados que la enfermedad de interés Malaria.

Mínimo	1er cuartil	Mediana	Media	3er cuartil	Máximo
0	0	6	85.56	42.75	14535

Cuadro 5.2: Resumen estadísticas descriptivas para la incidencia de dengue 2015.

A diferencia de lo observado en la figura (5.2), se puede apreciar la asociación espacial de la enfermedad Dengue y esta tiene una gran concentración en la costa Caribe, llanos orientales y ciertos casos bajo la región Andina con fuerte presencia sobre las cordilleras.

En este caso el sur del país contiene registros muy bajos con relación a la enfermedad de interés que es Malaria. En la figura (5.4), se observa que el número de casos de la enfermedad es muy superior a la variable de interés y se observa mucho más aleatorizada en cuanto a sus asociaciones espaciales.

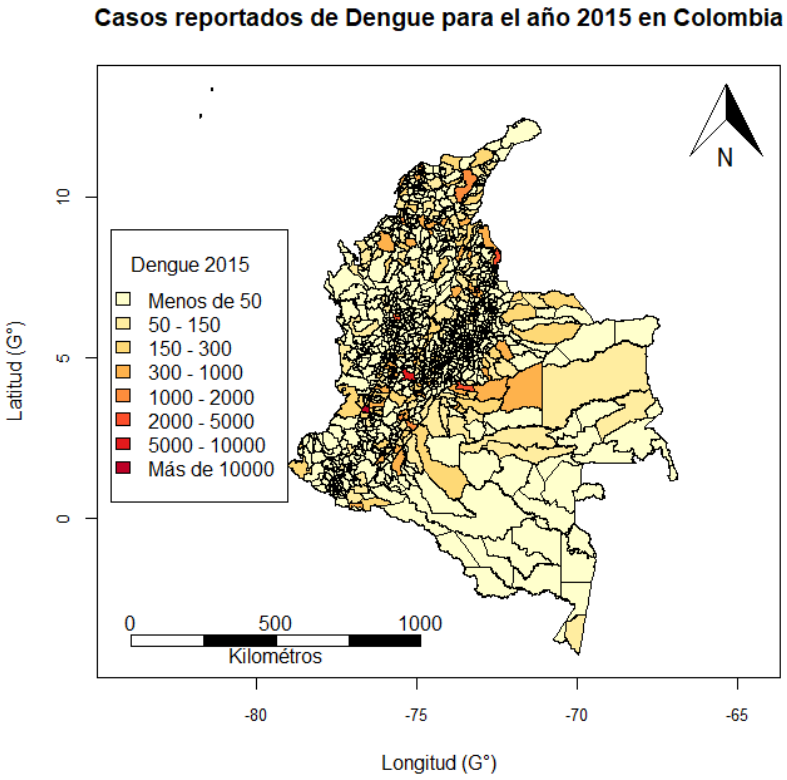


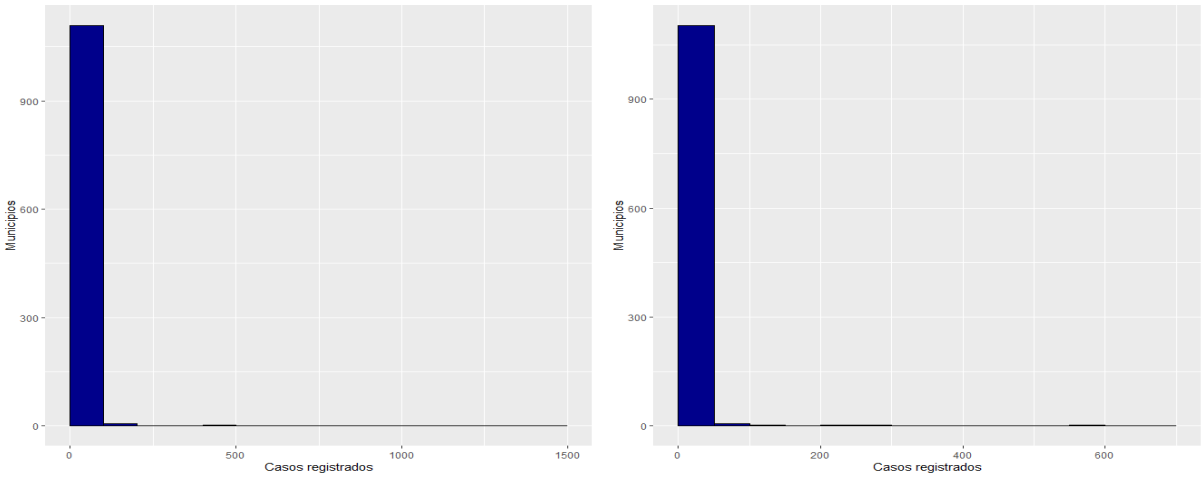
Figura 5.4: Box map de casos reportados de Dengue para 2015 en el territorio colombiano a nivel municipal, software R

Bajo la temporalidad establecida en un principio (año 2015) se añade el comportamiento de dos enfermedades más como lo son Zika y Chikungunya, que a diferencia de las dos primeras enfermedades los registros de estas solo se encuentran en periodos máximos de 3 años.

Para el tiempo de referencia establecido Año 2015, el comportamiento de estas enfermedades es muy similar a las anteriormente descritas, aquí se observa que tienen un exceso de casos reportados con 0 incidencias en un gran porcentaje sobre el territorio colombiano.

De acuerdo con la figura (5.5), se puede apreciar que ambos histogramas son asimétricos positivos dado que más del 75 % de los datos tienen recuento igual a cero, esto más adelante

se observará si es por falta de canales de transmisión o simplemente la ausencia de la enfermedad sobre Colombia.



(a) Comportamiento de la variable Zika (b) Comportamiento de la variable Chikungunya

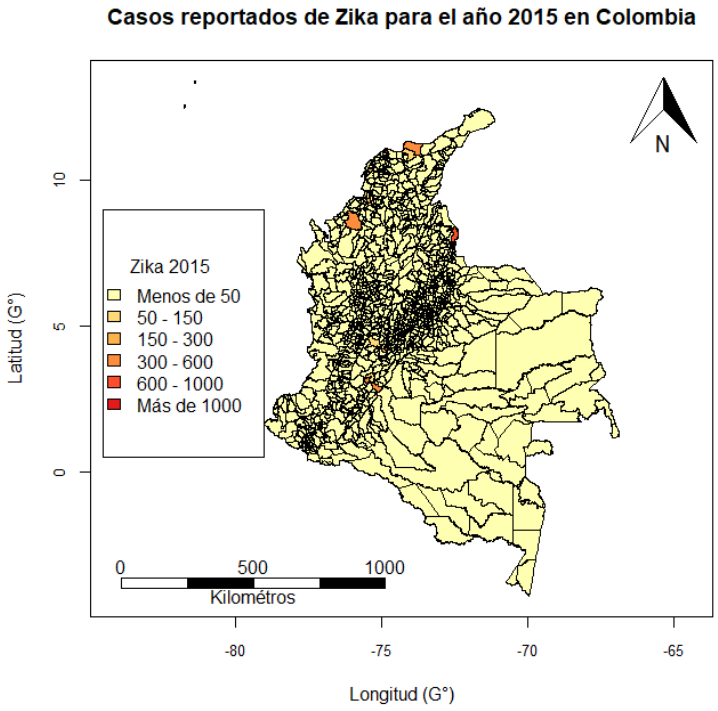
Figura 5.5: Histograma de casos reportados de Zika y Chikungunya para 2015, software R

Se puede apreciar en la tabla (5.3) que ambas enfermedades superan el exceso de ceros superior a un 75 % de los datos, también valores máximos muy grandes que pueden hacer referencia a los cluster que presenta la enfermedad Malaria que es objeto de estudio.

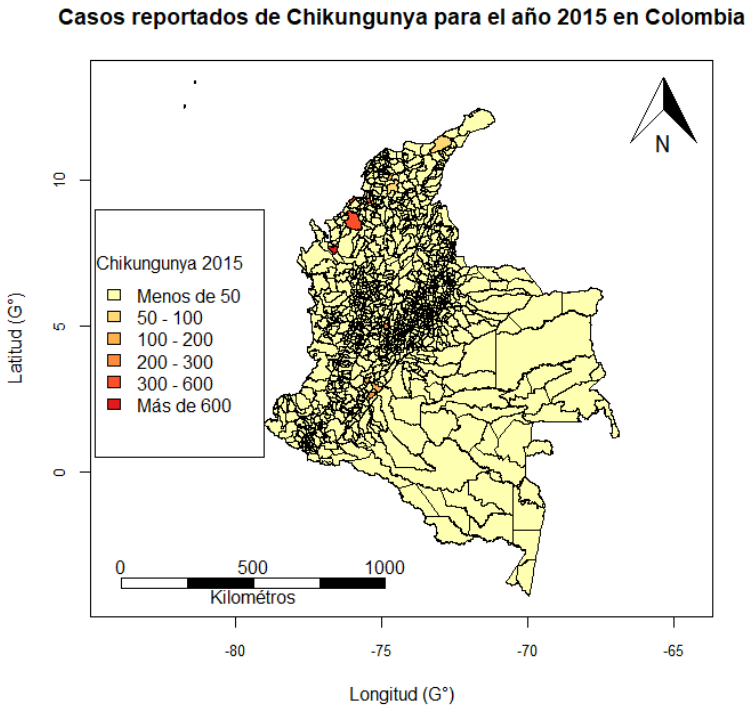
Enfermedad	Mínimo	1er cuartil	Mediana	Media	3er cuartil	Máximo
Zika	0	0	0	6.893	0	1366
Chikungunya	0	0	0	4.039	0	645

Cuadro 5.3: Resumen estadísticas descriptivas para la incidencia de Zika y Chikungunya 2015.

Para observar las asociaciones espaciales de estas enfermedades, se debe remitir a la figura (5.6) la cual muestra el pequeño recuento de estas enfermedades, en donde se observa que dentro del territorio colombiano no existen casos reportados significativos en comparación con las dos enfermedades principales como lo son Malaria y Dengue.



(a) Comportamiento espacial de la variable Zika



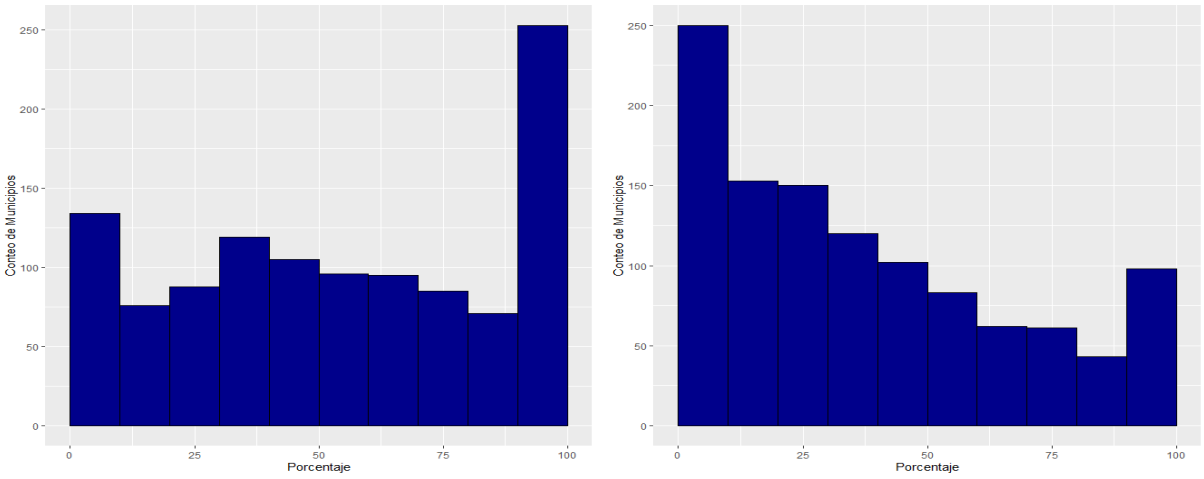
(b) Comportamiento espacial de la variable Chikungunya

Figura 5.6: Box map de casos reportados de Zika y Chikungunya para 2015 en el territorio colombiano a nivel municipal, software R

En la tabla (4.1) se observan más variables a parte de las relacionadas con enfermedades, allí se observan variables de tipo ambiental y otras que aluden a presencia y gestión gubernamental como lo es la cobertura de acueducto y alcantarillado.

Estas dos variables, son porcentajes en el cubrimiento de la prestación del servicio en cada municipio, como se observa en la figura (5.7) la cobertura de acueducto en el país es mayor que la cobertura de alcantarillado, estas son dos variables de mucho interés dado que las enfermedades de Dengue y Malaria pueden verse asociadas.

La cobertura de estos servicios idealmente debería ser cercana al 100 %, sin embargo, existen zonas como la Amazonia donde su cobertura se hace nula dadas las condiciones de protección que se tienen.



(a) Cobertura de Acueducto (b) Cobertura de Alcantarillado

Figura 5.7: Histograma de cobertura de acueducto y alcantarillado para 2015, software R

A partir de esto, el histograma de la cobertura de acueducto figura (5.7a) parece que en algún momento seguirá una distribución uniforme continua, dado algunas excepciones como lo es los valores de cubrimiento del 100 % al contrario de la variable alcantarillado figura (5.7b) que si es un histograma asimétrico positivo.

Cobertura	Mínimo	1er cuartil	Mediana	Media	3er cuartil	Máximo
Acueducto	0.00	28.61	53.85	54.65	85.89	100.00
Alcantarillado	0.00	12.40	30.50	37.23	58.48	100.00

Cuadro 5.4: Resumen estadísticas descriptivas para la cobertura de acueducto y alcantarillado, año 2015.

En la tabla (5.4), se puede apreciar el comportamiento casi uniforme de la variable

antes mencionada, acueducto, y también el comportamiento asimétrico de la variable alcantarillado dado que su media es superior a su segundo cuartil.

Para observar las asociaciones espaciales de cada una de las variables, se utilizan los Box Map aquí se puede observar en la figura (5.8). Como se menciono anteriormente, la parte del Amazonia tiene cobertura nula de ambos servicios así mismo como lo es la parte de la Orinoquía, esto puede influir drásticamente el modelo puesto que, estas áreas tienen presencia de la variable Malaria la cual es el objeto de este estudio. Además de eso, la parte del Pacifico esta con una cobertura superior al 80 % en cuanto al servicio de acueducto, con algunas excepciones, sin embargo, en la cobertura de alcantarillado se localizan municipios los cuales tienen baja cobertura y esto podría influenciar la presencia de las enfermedades.

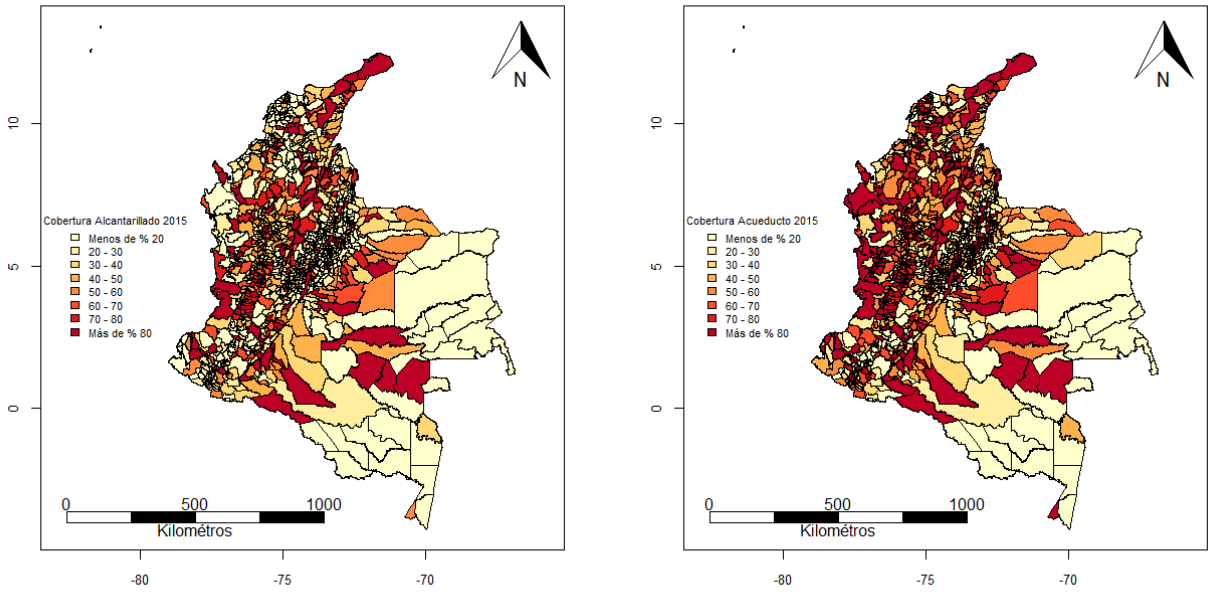
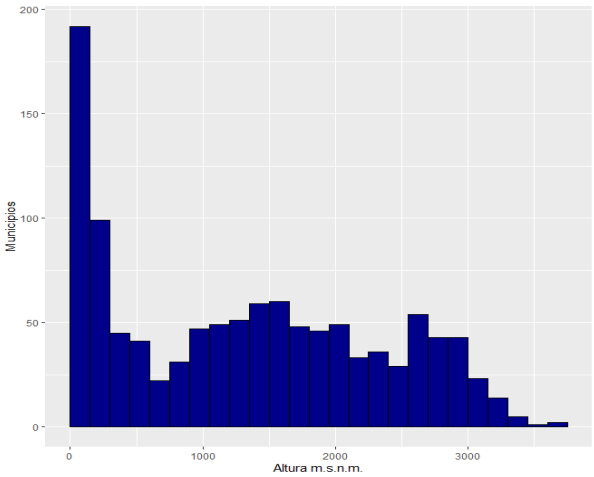


Figura 5.8: Mapas de cobertura de alcantarillado y acueducto en Colombia, año 2015, software R

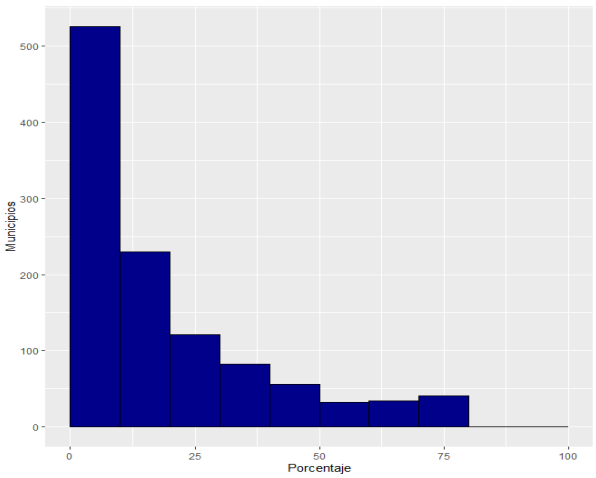
Por último, las variables ambientales se tendrán en cuenta para el estudio. Estas son variables con un sesgo dado que la recolección de estas se da mediante sensores remotos y su adaptación a los datos de área se da mediante el promedio de datos sobre las regiones.

Como en la mayoría de estudios de epidemiología para Malaria tienen en cuenta la presencia o ausencia de las poblaciones de mosquitos, pero para este caso, la regionalización de este tipo de variables y encontrarlas al nivel de detalle utilizado es de difícil acceso o

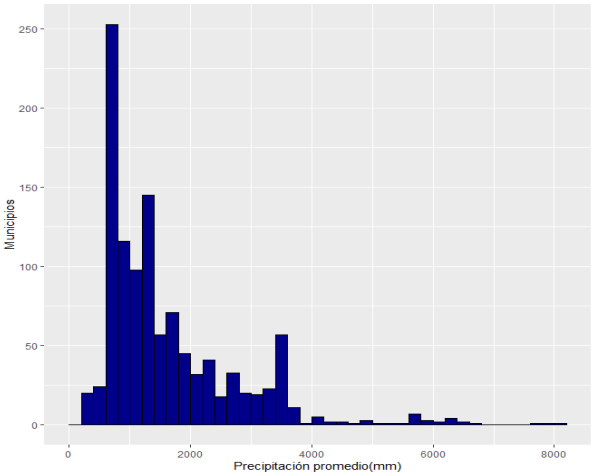
no se encuentra para el territorio colombiano, se opta por utilizar las variables que pueden ser determinantes en que la población de mosquitos este presente en cada uno de los municipios. Para esto se tiene en cuenta la precipitación, la altura media y el porcentaje de bosque natural, cada una de estas variables fue compartida por el IDEAM.



(a) Altura media s.n.m.



(b) Porcentaje de Bosque Natural

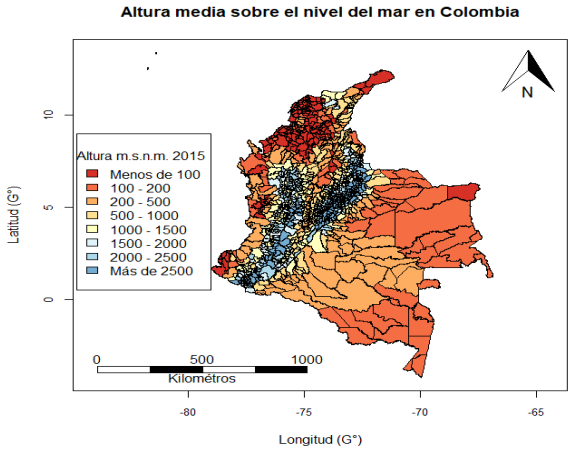


(c) Precipitación media

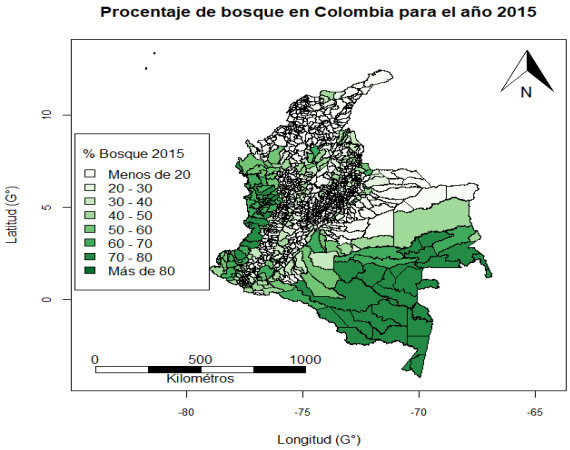
Figura 5.9: Histograma de variables ambientales para el año 2015, software R

Aquí se observa que el territorio colombiano esta comprendido en su mayoría por una altura media cercana a cero, dado que mas de 150 municipios están sobre las costas (Caribe y Pacifica), además también cuenta con el sistema de cordilleras que hace que sus alturas entre un municipio y otro sean muy distintas. En cuanto a la proporción de bosque natural, Colombia es un país rico en cuanto a sus recursos, siendo así fuertemente un

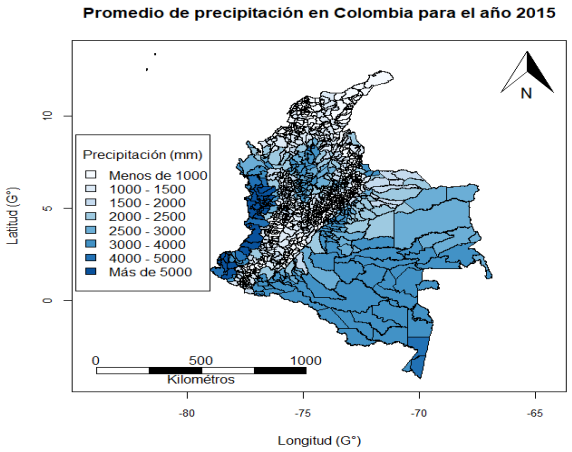
pulmón para todo el mundo, sin embargo, no todos los municipios tienen esta cobertura como se puede apreciar en la figura (5.9b). Observando la precipitación, esta para el año 2015 esta cercana a los 1000mm en promedio, que es una de las características propia de la costa Caribe y de los municipios sobre las cordilleras.



(a) Altura media s.n.m.



(b) Porcentaje de Bosque Natural



(c) Precipitación media

Figura 5.10: Mapa de caja de las variables ambientales para el año 2015, software R

Como se observa en la figura (5.10), tanto la variable de bosque como la precipitación se podrían relacionar con las enfermedades dada su alta presencia en el Pacifico y el Amazonia. Mientras que la altura media sobre el nivel del mar no es del todo visible en los clusters antes definidos por las variables de epidemiología.

5.2. Matrices de contigüidad espacial

Anselin (1998) dice que la noción misma de dependencia espacial implica la necesidad de determinar qué otras unidades en el sistema espacial influyen en la unidad particular bajo consideración. Formalmente, esto se expresa en las nociones topológicas de vecindario y vecino más cercano. Según esta noción, la estructura subyacente de los vecinos se expresa mediante valores 0 1. Si dos unidades espaciales tienen un borde común de longitud distinta de cero, se considera que son contiguas y se asigna un valor de 1. (Arbia, 2014)

Si no se analizan diferentes tipos de contigüidad no se tendrıa en cuenta el principio de alotopía que establece que a menudo lo que ocurre en una región se encuentra relacionado con otro fenómeno localizado en otras partes diferentes y remotas del espacio. (Ancot et al., 1982)

La elección de la matriz de pesos espaciales (SWM), se llevó a cabo mediante una revisión de las conexiones presentadas en las matrices generadas por los efectos reina, torre, k- vecinos más cercanos, distancia de Gabriel y Triangulación Delaunay.

Criterio		Enlaces No nulos	% W No nulos	Promedio de Enlaces
Torre	Orden 1	6394	0.5079102	5.698752
	Orden 2	15528	1.233473	13.83957
	Orden 3	26672	2.118702	23.77184
	Orden 4	38762	3.079076	34.54724
	Orden 5	51492	4.090289	45.89305
	Orden 6	63636	5.054953	56.71658
Reina	Orden 1	6570	0.5218908	5.855615
	Orden 2	16054	1.275256	14.30838
	Orden 3	27570	2.190035	24.57219
	Orden 4	39954	3.173763	35.60963
	Orden 5	52936	4.204994	47.18004
	Orden 6	65198	5.179032	58.10873
Métodos basados en distancias				
K-vecinos	k=1	1122	0.08912656	1
	k=2	2244	0.1782531	2
	k=3	3366	0.2673797	3
	k=4	4488	0.3565062	4
	k=5	5610	0.4456328	5
	k=6	6732	0.5347594	6
Distancia de Gabriel		5068	0.4025788	4.516934
Triangulación de Delaunay		6706	0.532694	5.976827

Cuadro 5.5: Conexiones presentadas en los criterios de contigüidad evaluados, Software R

Observando la tabla (5.5) el número de enlaces nulos crece en cuanto se utiliza un orden mayor para cada uno de los efectos, esto se da puesto que se están utilizando un total de 1120 regiones (municipios). También se observa la cantidad porcentual de enlaces establecidos, sin embargo, como la variable respuesta es una enfermedad que no tiene registro de casos sobre el territorio nacional, no es pertinente escoger el de mayor porcentaje como podría ser lo más conveniente en una variable continua. El promedio de enlaces puede ser más intuitivo para la determinación de la matriz de pesos dado que se establecen un mínimo de enlaces. Con base a lo anterior el promedio de enlaces llega a ser el más significativo en cuanto a una interpretación visual de las matrices binarias, así pues, de acuerdo con la figura (5.2) y la tabla (5.5), el numero de relaciones entre cada uno de los municipios es aproximadamente del orden menor a 15, así se podría optar por el uso de cualquiera de los criterios (Torre 1 y 2, reina 1 y 2 o cualquiera de los métodos basados en distancias).

Dray et al. (2006) explora un enfoque llamado coordenadas principales de matrices vecinas (PCNM) para crear predictores espaciales que se pueden incorporar fácilmente en modelos de regresión, proporcionando una herramienta flexible para la elección de la matriz de pesos espacial.

Este enfoque depende de la maximización de los mapas de vectores propios de Moran (MEM) que se derivan de la matriz espacial de ponderación (SMW) o matriz de pesos espaciales. Se generan variables MEM (también denominadas predictores espaciales o vectores propios espaciales) por la diagonalización de una matriz de ponderación espacial doblemente centrada (SWM) W, la optimización para la selección se realiza maximizando el R cuadrado ajustado (R^2_{Adjust}) o minimizando la autocorrelación espacial residual.

Bauman et al. (2018) observo que optimizar la elección del SWM conducía a tasas de error de tipo I infladas si no se aplicaba un control explícito del número de SWM probadas. Por lo tanto se debe aplicar una corrección de Šidák (1967) para múltiples pruebas al *P_value* de la prueba global de cada SWM (es decir, el modelo que integra todo el conjunto de predictores espaciales). La corrección de Sidak se calcula como $P_{corregido} = 1 \quad (1 \quad P)^{1=n}$ donde n es el número de pruebas realizadas, P es el *P_value* observado y $P_{corrected}$ es el nuevo valor p después de la corrección. El valor p se calcula primero usando n permutaciones y luego se corrige de acuerdo con el número total de SWM probadas.

Los resultados de esta elección pueden observarse mejor en la tabla (5.6), en donde se observa lo antes mencionado, las matrices que podrían considerarse son torre 1 y 2, reina 1 y 2 o cualquiera de los métodos basados en distancias.

Criterio		R2Adj.pos	R2Adj.neg	Pvalue.pos	Pvalue.neg	N.var	R2Adj.select
Reina	1	0.390778	-0.578802	0.003992	1.000000	30	0.337734
	2	0.170878	-0.184571	0.091641	1.000000		
	3	0.109236	-0.097589	0.587589	1.000000		
	4	-0.020113	0.064955	1.000000	0.985862		
	5	-0.022330	0.087213	1.000000	0.836746		
	6	0.013909	0.080506	1.000000	0.955197		
Torre	1	0.390852	-0.568278	0.003992	1.000000	36	0.355027
	2	0.183512	-0.203211	0.065794	1.000000		
	3	0.118128	-0.109914	0.458097	1.000000		
	4	0.000908	0.033643	1.000000	0.999999		
	5	-0.007466	0.060882	1.000000	0.995122		
	6	0.012216	0.067462	1.000000	0.986417		
K vecinos	1	0.349156	-0.261317	0.031506	1.000000	16	0.346117
	2	0.353745	-0.383388	0.027621	1.000000	7	0.347253
	3	0.257762	-0.315669	0.050703	1.000000		
	4	0.410966	-0.604202	0.003992	1.000000	27	0.407728
	5	0.500380	-0.820199	0.003992	1.000000	35	0.413964
	6	0.408820	-0.743497	0.003992	1.000000	27	0.347199
Distancia de Gabriel		0.378266	-0.476845	0.003992	1.000000	32	0.285089
Triangulación de Delaunay		0.430865	-0.642939	0.003992	1.000000	41	0.326207

Cuadro 5.6: Parámetros de elección de la matriz de contigüidad, software R

Para el test, se selecciono un nivel de significancia ( ) de 5 %, así pues se establece el primer filtro para la elección de la matriz donde quedan únicamente el efecto reina de orden 1, el efecto torre de orden 1, y cualquiera de los métodos basados en distancias a excepción de k vecinos con 3 vecinos más cercanos.

El criterio que más valor tiene para la selección definitiva de la matriz de contigüidad es el R cuadrado ajustado (R^2_{Adjust}), que minimiza la autocorrelación. Por esto la matriz elegida será K vecinos con 5 vecinos más cercanos.

5.3. Información a priori

Para establecer la información a priori se debe establecer el tipo de distribución que pueden adoptar los datos para posteriormente establecer los hiper-parámetros que expresarán la información inicial. Para esto se debe tener en cuenta también el tipo de familia conjugada de la distribución del parámetro de interés, en este caso es el número de registros por municipio de la enfermedad seleccionada.

Este paso resulta ser el de mayor importancia dentro del estudio dado que es información externa que se infiltra en el análisis ya que de este depende si la estimación del riesgo de contraer una enfermedad en cada municipio sea correcta o no. Para esto existen

informaciones extra muéstrales que se denominan *Informativas* o *No-informativas*, esto se dará puesto si existen antecedentes subjetivos que daten con cierta precisión la información a priori o si se evidencia el estado de ignorancia absoluta.

(A. Lawson et al., 1999) describe que uno de los modelos más usados en epidemiología son aquellos que tienen que ver con datos de conteo (distribuciones Poisson), sin embargo, Blangiardo y Cameletti (2015) comenta que pese a que la distribución de Poisson es muy usada en el mapeo de enfermedades, esta se puede justificar como un modelo binomial cuando la enfermedad es rara o su probabilidad binomial es pequeña, así pues esta puede ser usada en esta ocasión o para establecer la tasa de prevalencia de la enfermedad.

5.3.1. A priori no-informativa

A menudo se supone que las distribuciones a priori no hacen preferencias fuertes sobre los valores de las variables. A veces se conocen como distribuciones a priori vagas, de referencia o planas o no informativas. (A. Lawson, 2013)

Je□reys (1998) sugirió que un criterio razonable para desarrollar distribuciones a priori que deberían implicar que las probabilidades hechas sobre las variables aleatorias observables deberían permanecer invariables bajo los cambios en la parametrización del problema.

En términos generales, la construcción de esta clase consiste en buscar simultáneamente invariancia ante transformaciones y proveer la menor información a priori en relación a la información muestral, vía la información de Fisher.

Para satisfacer este criterio, Je□reys demostró que un parámetro de vector aleatorio : (p x l), o un parámetro de matriz aleatorio : (p x p) debe recibir una densidad a priori de la forma:

$$g(\ ) \ / \ jJj^{0:5} \tag{31}$$

Donde $J \quad j_{ij}$ denota la matriz de información de Fisher cuadrada asociada con la función de probabilidad de los datos. Es decir, si $f(x;:::;x_n)$ denota una función de probabilidad para los datos posiblemente p-dimensionales para cada i (p podría ser igual a *l*), los componentes de *J* están definidos por: (Press, 1989)

$$j_{ij} \qquad E\left[\frac{@^2 log f(x_1;:::;x_n)}{@\ _i @\ _j}\right] \tag{32}$$

A partir de la regla de Je□rey se determinaron las distribuciones a priori para las distribuciones Poisson y binomial dado que estas son las de interés para el presente trabajo.

Partiendo de la ecuación (32) entonces:

Poisson: si y_i $Poisson(\ )$ entonces:

$$f(y_i j\ ) = {}^{P\ y_i} e^{\ n}\ \prod_{i=1}^{n} \frac{1}{y_i!} \tag{33}$$

Que se deriva de la ecuación (25), así pues:

$$log f(y_i j\ ) = \sum_{i=1}^{n} y_i log(\ )\quad n\quad log\left[\sum_{i=1}^{n} y_i!\right] \tag{34}$$

Derivando con respecto a se obtiene:

$$\frac{@ log f(y_i j\ )}{@} = \frac{@}{@}\sum y_i log\quad \frac{@}{@} n\quad \frac{@}{@} log\left[\sum_{i=1}^{n} y_i!\right]$$

$$= \sum y_i \frac{1}{-}\quad n$$

$$\frac{@^2 log f(y_i j\ )}{@\ ^2} = \frac{@}{@}\sum y_i\quad {}^{1} + \frac{@}{@} n$$

$$\frac{@^2 log f(y_i j\ )}{@\ ^2} = \quad \frac{\sum y_i}{2} \tag{35}$$

Reemplazando en la (32)

$$E\left[\frac{@^2 log f(y_i j\ )}{@\ ^2}\right] = E\left[\frac{\sum y_i}{2}\right]$$

$$= \frac{1}{2} E\left[\sum y_i\right]$$

$$= \frac{n}{2}$$

$$E\left[\frac{@^2 log f(y_i j\ )}{@\ ^2}\right] = \frac{n}{-} \tag{36}$$

Así pues lo que satisface el criterio de Je□rey en la ecuación (31) se da mediante:

$$g(\ ) \ / \ j\frac{n}{-}j^{0:5} \tag{37}$$

n es simplemente una constante así que la distribución a priori queda:

$$g(\ ) \ / \ \frac{1}{0{:}5} \tag{38}$$

De acuerdo con Press (1989), esta es una distribución impropia. Una distribución impropia es la que suma o integra a un valor diferente de uno, suponga K. Si K es finito, entonces la distribución impropia induce una distribución propia normalizando la función. Si K es infinito, entonces la distribución tiene un papel de ponderación o de herramienta técnica para llegar a una distribución posterior. Esto resulta ser un problema, sin embargo, según Ruiz y Peña (2007) la distribución a posteriori de :

$$(\ jy) \ / \ p(yj\ )g(\ ) \ / \ e^{\ n\ \ \sum y_j\ \ 1=2} \tag{39}$$

Lo anterior es el Kernel de la distribución gamma por lo que $(\ jy) = Gamma(\ j\sum y_i + 1=2; n)$, dado que la anterior es la distribución posterior y por la propiedad del modelo Poisson-Gamma se tiene: (A. Lawson, 2013)

$$\begin{array}{cccc} & Modelo & A\_priori & Posterior \\ y & Poisson(\ ) & G(\ ;\ ) & jy \quad G(\sum y_i + \ ; n + \ ) \end{array} \tag{40}$$

De acuerdo con la ecuación (40), los Hiper-parámetros y son respectivamente $(1=2; 0)$ por lo tanto la distribución a priori no informativa para los modelos Poisson será:

$$g(\ ) = G(1=2; 0) \tag{41}$$

Binomial: si se esta interesado en la prevalencia de mortalidad o incidencia de una enfermedad ( ) en n áreas, se sabe que $(y_1; :::; y_n)$ es el número de casos registrados en cada una de estas áreas y $(\ _1; :::; \ _n)$ es el número total de personas que viven en cada área, entonces $y_i \quad Binomial(\ _i; \ _i)$ con su respectiva función de densidad:

$$f(y_i j\ ) = \binom{\ }{y}\ ^y(1 \quad)\ ^y \tag{42}$$

Realizando el procedimiento de la ecuación (32) se tiene:

$$log f(y_i j\ ) = log\binom{\ }{y} + y log(\ ) + (\quad y) log(1 \quad) \tag{43}$$

Derivando se obtiene:

$$\frac{@ log f(y_i j \quad)}{@} = \frac{y}{} + (\quad y)\frac{1}{(1 \quad)}$$
$$= y \quad {}^{1} \quad (\quad y)(1 \quad)^{\ 1}$$

$$\frac{@^2 log f(y_i j \quad)}{@ \quad ^2} = \frac{y}{2} \quad \frac{y}{(1 \quad)^2} \tag{44}$$

Reemplazando en la (32) y teniendo en cuenta $E[y] =$, entonces:

$$E\left[\frac{@^2 log f(y_i j \quad)}{@ \quad ^2}\right] = \frac{}{2} + \frac{}{(1 \quad)^2}$$
$$= - + \frac{(1 \quad)}{(1 \quad)^2}$$
$$= - + \frac{}{(1 \quad)}$$
$$= \left[\frac{1 \quad +}{(1 \quad)}\right]$$

$$E\left[\frac{@^2 log f(y_i j \quad)}{@ \quad ^2}\right] = \frac{}{(1 \quad)} \tag{45}$$

Así pues lo que satisface el criterio de Je□rey en la ecuación (31) se da mediante:

$$g(\quad) \ / \ \sqrt{E\left[\frac{@^2 log f(y_i j \quad)}{@ \quad ^2}\right]}$$
$$p_{-}$$
$$g(\quad) \ / \ \frac{}{1=2(1 \quad)^{1=2}}$$

$$g(\quad) \ / \ p_{-}(\quad {}^{\frac{1}{2}}(1 \quad)^{\ \frac{1}{2}}) \tag{46}$$

La ecuación (46) es el kernel de la distribución Beta,

$$Beta(\quad ; \quad) = \frac{(\quad + \quad)}{(\quad)\ (\quad)} \quad {}^{1}(1 \quad)^{\ 1} \tag{47}$$

Por lo tanto la distribución a priori no informativa para la distribución binomial es $beta(\quad = 1=2; \quad = 1=2)$:

$$g(\quad) \ / \ beta(\frac{1}{2}; \frac{1}{2}) \tag{48}$$

Binomial negativo: En el trabajo aplicado, la regresión de Poisson es restrictiva en el análisis de los datos de conteo. Se reconoce que los recuentos a menudo muestran una considerable variación extra-Poisson o sobredispersión. La sobredispersión se refiere a la situación cuando la varianza de una variable dependiente observada excede la varianza nominal, dada la distribución supuesta respectiva. La suposición en el modelo de Poisson de que la media condicional y la varianza de Y dado X son iguales puede ser demasiado fuerte y, por lo tanto, no dar cuenta de la sobredispersión. La imposición inapropiada de esta restricción puede dar lugar a errores estándar estimados irracionalmente pequeños de las estimaciones de los parámetros. La regresión binomial negativa es quizás la forma más conveniente de relajar la restricción de Poisson y lidiar con la sobredispersión. (Wang et al., 2018)

Debido a que se esta trabajando para lograr un GLM, se presenta la Binomial negativa utilizada en los GLM. Se presenta en la literatura como una combinación de dos distribuciones, dando una distribución combinada de Poisson-gamma. Esto significa que primero se supone que las Y son Poisson distribuidas con la media que se supone sigue una distribución gamma. (A. Zuur et al., 2009)

Al igual que la distribución Binomial, se puede encontrar la a priori de Je□reys, sin embargo, al trabajar con campos aleatorios Gausianos de Markov (GMRF por sus siglas en inglés) la distribución beta como a priori no tiene mucho sentido. Para esto INLA tiene las a priori más indicadas para este tipo de modelos dada cada distribución especifica (para los modelos Poisson es una distribución Gaussiana y para la Binomial INLA no tiene a priori predeterminada) que han denominado a priori penalizadas complejas. (Simpson, Rue, Riebler, Martins, y Sørbye, 2017)

Para la distribución Binomial negativa la a priori penalizada compleja es una distribución Gamma de acuerdo con:

$$\begin{aligned} y_i &\quad Poisson(z\ ) \\ z_0 &\quad Gamma(\ ^1;\ ^1) \end{aligned} \tag{49}$$

z esta dado en escala logarítmica $z_0 = log(z)$, por ende el valor de z para que sea a priori no informativo debe ser igual a 2:718282 aproximadamente, para esto se trunca el valor de = 7.

Poisson inflado con ceros: A menudo cuando se trabajan unidades de análisis pequeñas o se intenta modelar una enfermedad rara, se realiza mediante distribuciones infladas de ceros. Como parten de dos distribuciones por lo general para datos discretos una

distribución es la Bernoulli y la otra puede ser tanto Poisson como Binomial.

De acuerdo con lo expuesto en (Liu y Powers, 2012), las distribuciones a priori no informativas para este tipo de caso se pueden determinar mediante el uso de la regla de Je□reys combinada, esto quiere decir que la distribución a priori no informativa para ambas distribuciones (Bernoulli y Poisson en este caso) serán utilizadas.

Como se ha mencionado, la distribución a priori no informativa para Poisson esta determinada por $g(\ ) = Gamma(1=2; 0)$, para la distribución Bernoulli, en este caso se puede establecer de manera sencilla con una distribución uniforme continua dado que los únicos valores que esta puede adoptar es 0 y 1 de acuerdo con la ecuación (27), la distribución indicada sería $uniforme(0; 1)$ (Mersad, Ganjali, y Rivaz, 2015).

A pesar de la solución encontrada, implementar esto en el algoritmo de INLA es algo complicado, por eso el paquete R-INLA proporciona una solución interesante la cual es la implementación de una distribución a priori Logit-Beta.

Suponga que el parámetro de interés se distribuye como un ZIP

$$y_i \quad ZIP(\ ) \tag{50}$$

El parámetro se determina mediante:

$$log(\frac{p}{1 \quad p}) \quad logit(p) \tag{51}$$

p es la proporción de ceros que se tienen y se distribuye mediante una función Beta.

$$p \quad Beta(\ ;\ ) \quad \frac{(\ +\ )}{(\ )\ (\ )} p^{\ 1}(1 \quad p)^{\ 1} \tag{52}$$

La ecuación (52) es la que establece los hiperparámetros de interés, sin embargo, la ecuación (51) es donde se determina que tan informativo o no es la información extra muestral.

- Si el $logit(p) = 0$ existe la misma cantidad de ceros como de números distintos a cero, y el parámetro será igual a cero.
- Si el $logit(p) > 0$ mayor cantidad de ceros, y el parámetro será mayor a cero.
- Si el $logit(p) < 0$ menor cantidad de ceros, y el parámetro será menor a cero (caso que no debería ocurrir puesto que se habla de conteos y no puede ser inferior a cero).

En este caso es de interés que exista la misma proporción de ceros como de números mayores a cero, así pues, el $logit(p) = 0$ debe ser la opción seleccionada para poder encontrar los hiperparámetros. Para que esto se de, p debe ser igual a 0.5, entonces se debe estimar los valores asociados para y .

Propiedades de la distribución Beta

1. $Beta(;) = Beta(;)$
2. $Beta(;1) = \frac{1}{}$
3. $Beta(+ 1;) = \frac{}{+} Beta(;)$
4. $E(x \quad Beta) = \frac{}{+}$

Por lo tanto para cualquier valor de = > 0, se cumplirá que p =0.5.

5.3.2. A priori informativa

La información extra muestral se puede extraer de estudios previos realizados, sin embargo, al nivel de detalle trabajado en el presente proyecto, resulta complicado la búsqueda de tal información. Dentro de los boletines epidemiológicos del Instituto Nacional de salud (INS) no se encuentra a detalle los datos puesto que se tiene como unidad de análisis los departamentos, esto sesga la información, sin embargo, es una fuente de información que se puede incluir dentro del estudio.

Dentro de la búsqueda de información a priori, se encontró algo muy particular que es de valiosa información para estimar el riesgo relativo de cada municipio en el país de contraer la enfermedad, esto fue el Índice Parasitario Anual de Malaria el cual expresa la relación de los casos de malaria y la población que vive en zonas de riesgo por cada mil habitantes (vivir a menos de 1600 metros por debajo del nivel del mar, en clima templado), da cuenta de la probabilidad de contraer la enfermedad entre la población en riesgo.

Estos datos se encuentran a nivel departamental con una periodicidad anual en una serie temporal desde el año 2000 hasta el año 2019 y se pueden encontrar en ÍNDICE PARASITARIO ANUAL DE MALARIA - GEOREFERENCIADO.

Para el presente estudio se toma el IPA anual de Colombia para el periodo 2000-2015

AÑO	2000	2001	2002	2003	2004	2005	2006	2007
IPA	5.90	5.60	6.30	5.00	5.00	4.80	8.40	11.10
AÑO	2008	2009	2010	2011	2012	2013	2014	2015
IPA	6.20	7.90	11.50	6.30	5.80	5.00	3.40	5.20

Cuadro 5.7: IPA para Colombia, software R

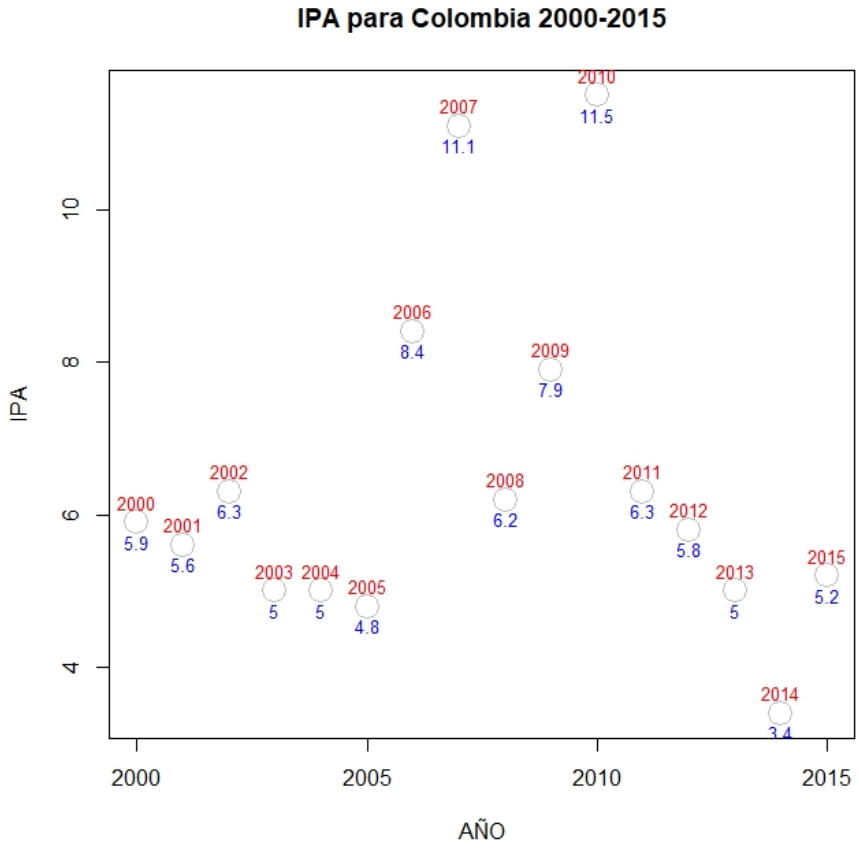


Figura 5.11: IPA para Colombia, software R

Como se puede apreciar en la tabla (5.7) y la figura (5.11), el IPA en Colombia tiene una tendencia media cercana a 5.5 aproximadamente, sin embargo, existen dos años en los cuales este índice se disparo de forma abrupta al igual que el crecimiento inesperado en otros dos, esto para el periodo comprendido entre 2006 a 2010, esto puede influir drásticamente en el acceso extra muestral. Cuando se quiere incluir esta información externa dentro del estudio, se tienen en cuenta dos principios los cuales son uno de centralidad, que equivale al valor medio o esperanza del parámetro de interés y el otro de dispersión el cual se da mediante la información que se encuentra dentro de los cuartiles.

El valor medio de IPA es de 6.4625, además se calcula un intervalo de confianza para determinar la dispersión de este, bajo una confiabilidad del 5%:

$$\begin{aligned} IC = Media \quad & t(1 \quad \frac{}{2}; n \quad 1) \quad \frac{p}{n \quad 1} \\ Limite\ inferior &= 5{:}236201 \\ Limite\ superior &= 7{:}688799 \end{aligned} \tag{53}$$

Este intervalo y media están estimados para IPA a nivel nacional, pero en realidad se necesita es la media para un nivel de detalle municipal, lo primero es calcular estos tres indicadores denotados por x (media, limite superior e inferior) en el número de casos confirmados de acuerdo con la población en riesgo:

$$Confirmados = Pob\_riesgo \quad \frac{x}{1000} \tag{54}$$

Así pues se hallan los casos confirmados para Colombia, sin embargo, es de interés que los datos sean a nivel municipio, que reflejen los valores que puede adoptar cada municipio en el país, por esto se divide lo anterior por 1120 (excluyendo el departamento de San Andrés) obteniendo así:

$$\begin{array}{ccc} Li & Media(\ ) & Ls \\ 88{:}2227 & 108{:}8841 & 129{:}5456 \end{array} \tag{55}$$

Al tener estos indicadores, se puede establecer un sistema de ecuaciones que determine los hiperparámetros de la distribución a priori de acuerdo con la familia conjugada seleccionada. Entonces para la distribución de Poisson, la familia conjugada es Gamma. Para determinar estos hiperparámetros, estos deben cumplir:

$$Sistema\ de\ ecuaciones \begin{cases} F_{Gamma}(\ j\ ;\ ) = 0{:}5 \\ F_{Gamma}(L_s j\ ;\ ) \quad F_{Gamma}(L_i j\ ;\ ) = 0{:}95 \end{cases} \tag{56}$$

Una manera de solucionar este sistema de ecuaciones, es por medio de métodos numéricos generando una única ecuación como lo es:

$$(F_{Gamma}(\ j\ ;\ ) \quad 0{:}5)^2 + (F_{Gamma}(L_s j\ ;\ ) \quad F_{Gamma}(L_i j\ ;\ ) \quad 0{:}95)^2 = 0 \tag{57}$$

Optimizando esta y resolviendo la ecuación, los valores asociados a los hiperpárametros y

la distribución a priori será:

$$g(\ ) \ / \ Gamma(\ = 107{:}121; \ = 0{:}9824992) \tag{58}$$

En el caso de un modelo binomial como se expresa en la ecuación (42), es de interés la incidencia de la enfermedad sobre el territorio ( ), es por eso que integrar la información a priori ya utilizada para el modelo Poisson es imperativo para el desarrollo del proyecto, así pues, retomando la expresión (54) y teniendo en cuenta el valor esperado de la distribución binomial, se puede obtener la incidencia de la enfermedad:

$$\begin{gathered} E(y_i) = \\ = - \end{gathered} \tag{59}$$

Donde es la población total y es la media de los casos esperados. Se obtiene entonces los cuartiles 50 % y 95 %.

$Media(\ )$	95 %
0:0064625	0:007688799

(60)

Los cuales tendrán que cumplir el sistema de ecuaciones:

$$Sistema\ de\ ecuaciones \begin{cases} G_{Beta}(\ j\ ;\ ) = 0{:}5 \\ G_{Beta}(95\,\%j\ ;\ ) = 0{:}95 \end{cases} \tag{61}$$

Estableciendo por métodos numéricos al igual que la ecuación (57), se obtiene la siguiente distribución a priori con sus respectivos hiperpárametros:

$$g(\ ) \ / \ Beta(\ = 78{:}27473; \ = 12030{:}50890) \tag{62}$$

La distribución *Binomial negativa* es una alteración a la distribución de Poisson la cual es de gran utilidad para datos de conteo. Teniendo en cuenta esto, la información a priori establecida para el modelo Poisson será de utilidad para encontrar los hiper parámetros de la función de distribución experimental PC.gamma establecida como a priori para los GMRF de la librería INLA. De acuerdo con los parámetros establecidos en la ecuación (56), se establece el sistema de ecuaciones para resolver por métodos numéricos.

$$Sistema\ de\ ecuaciones \begin{cases} F_{PC:Gamma}(Z_0 j\ ) = 0{:}5 \\ F_{PC:Gamma}(L_s j\ ) \quad F_{PC:Gamma}(L_i j\ ) = 0{:}95 \end{cases} \tag{63}$$

El valor asociado al hiper párametro que cumple con el anterior sistema de ecuaciones es:

$$g(\ ) \ / \ PC{:}Gamma(\ = 0{:}381679) \tag{64}$$

Al igual que la información a priori para la distribución Binomial, cuando se modela un ZIP, se necesita una a priori con distribución beta que de información no sobre la incidencia de la enfermedad si no de la cantidad de ceros que están dentro de la muestra. Para esto se utiliza la base de datos expuesta en la tabla (4.1), donde la variable Malaria esta en el periodo comprendido entre 2007-2015.

	Año	Mpios_reportados	Porcentaje de Ceros
1	2007	403	0.64
2	2008	474	0.58
3	2009	494	0.56
4	2010	530	0.53
5	2011	396	0.65
6	2012	373	0.67
7	2013	335	0.70
8	2014	311	0.72
9	2015	312	0.72

Cuadro 5.8: Reporte de municipios para Malaria en Colombia para el periodo 2007-2015

En la tabla (5.8), se observa la cantidad de municipios que reportaron por lo menos 1 caso de incidencia de Malaria sobre su territorio, así mismo, se realiza el cálculo de los que no reportaron casos dividiendo los municipios reportados entre 1122 (total de municipios de Colombia), así se obtiene el porcentaje reportado y al 100 % se le resta este valor.

$$\%Ceros = 1 \quad \frac{Mpios\_reportados}{1122}$$

Teniendo en cuenta le ecuación (53), se realiza el intervalo de confianza para la cantidad de ceros, así obteniendo los cuartiles 50 % y 95 % que den solución al sistema de ecuaciones (61) y hallando la distribución a priori para ZIP.

$$\begin{array}{cc} Media(p) & 95\,\% \\ 0{:}6407209 & 0{:}6991746 \end{array} \tag{65}$$

Estableciendo por métodos numéricos al igual que la ecuación (57), se obtiene la siguiente

distribución a priori con sus respectivos hiperpárametros:

$$g(\ ) \ / \ Beta(\ = 110{:}46824;\ \ = 62{:}09235) \tag{66}$$

Capítulo 6

Análisis confirmatorio

El correcto tratamiento de los efectos espaciales en el proceso de modelización espacial constituye el denominado análisis confirmatorio de datos. A partir de la información descrita en el capítulo anterior, se estimará el modelo óptimo para describir el riesgo relativo a la prevalencia malaria en Colombia para el año 2015. Para tal efecto se propone explorar el ajuste de modelos tipo Poisson, Binomial para enfrentar el 72 % de registros que son cero en la base de datos y los modelos NB y ZIP para contemplar la posible sobresdipersión en la parte de conteo. Para modelar la auto-correlación espacial se exploraran la especificación iCAR con la construcción de matriz W basada en K vecinos con 5 vecinos más cercanos y serán estimados por medio de la librería R-INLA (A. F. Zuur et al., 2017). Para efectos de la selección del mejor modelo no solo se compararan los valores DIC y WAIC, sino que se estudiará la consistencia teórica tanto de los riesgos relativos obtenidos para las covariables como los mapas de riesgo que se comparan con el estimado por medio de las tasas estandarizadas SIR.

6.1. Cálculo del radio o tasa de incidencia estandarizada para la malaria en Colombia 2015

Poisson modelizó las tasas estandarizadas de incidencia o mortalidad (SIR / SMR), es decir, método indirecto para calcular las tasas estandarizadas. SIR es una relación de casos observados y esperados. Los casos esperados se derivan multiplicando la tasa de población específica de los estratos con los años-persona correspondientes a la estratificación especificada.

En este caso la base de datos vista en la tabla (4.1), muestra que se pueden dar 2

estratificaciones para el cálculo del SIR, estas se pueden realizar por Grupo étnico, dado que se tienen los registros de los reportes de infectados con malaria de acuerdo a la etnicidad y también se puede realizar de acuerdo, al Sexo. Para esto se tuvo en cuenta la estratificación únicamente por sexo, se estimo un crecimiento porcentual lineal de acuerdo a la población de hombres y mujeres para el año 2005. Así pues, este crecimiento se tuvo en cuenta para el año 2015 y así se obtuvieron las poblaciones proyectadas estimadas de acuerdo al sexo para el año en cuestión.

Cod_dane	Nom_munici	Nom_depart	Pob	Casos reportados	Sexo
05001	MEDELLÍN	ANTIOQUIA	1242031	2	HOMBRE
05001	MEDELLÍN	ANTIOQUIA	1222291	2	MUJER
05002	ABEJORRAL	ANTIOQUIA	9722	0	HOMBRE
05002	ABEJORRAL	ANTIOQUIA	9568	0	MUJER
05004	ABRIAQUÍ	ANTIOQUIA	1073	0	HOMBRE
05004	ABRIAQUÍ	ANTIOQUIA	1055	0	MUJER
05021	ALEJANDRÍA	ANTIOQUIA	1747	0	HOMBRE
05021	ALEJANDRÍA	ANTIOQUIA	1719	0	MUJER
05030	AMAGÁ	ANTIOQUIA	14896	0	HOMBRE
05030	AMAGÁ	ANTIOQUIA	14659	0	MUJER

Cuadro 6.1: Primeros registros para el cálculo del valor esperado de la enfermedad malaria, software R

En la tabla (6.1), se observa la estratificación usada para cada uno de los municipios, allí se observan la cantidad de casos reportados y la población referida a cada estratificación en este caso es el sexo (Hombre, Mujer), por esto se deben tener la cantidad de registros para cada municipio como cada una de las clases de la estratificación seleccionada, en este caso será igual a 2.

Cod_dane	Nom_munici	Nom_depart	Sex_mascul	Sex_femeni	Pob_hombre	Pob_mujer	Esperado
05001	MEDELLÍN	ANTIOQUIA	2	2	1242031	1222291	1971.99
05002	ABEJORRAL	ANTIOQUIA	0	0	9722	9568	15.44
05004	ABRIAQUÍ	ANTIOQUIA	0	0	1073	1055	1.70
05021	ALEJANDRÍA	ANTIOQUIA	0	0	1747	1719	2.77
05030	AMAGÁ	ANTIOQUIA	0	0	14896	14659	23.65

Cuadro 6.2: Primeros registros del valor esperado de la enfermedad malaria, software R

De acuerdo con la ecuación (30), el valor esperado es una sumatoria entre la población y la estratificación seleccionada, sin embargo, esto indica que la enfermedad o el valor que se espera encontrar en una región seguirá la tendencia de crecimiento de acuerdo con la selección. Es por esto que en la tabla (6.2) aunque se hallan registrado cero casos de

acuerdo al reporte de la enfermedad, se esperaría un crecimiento igual al de la población, esto quiere decir que la población infectada se comporta igual a la población estatificada, por esta razón los valores esperados son muy superiores a los casos reportados de malaria.

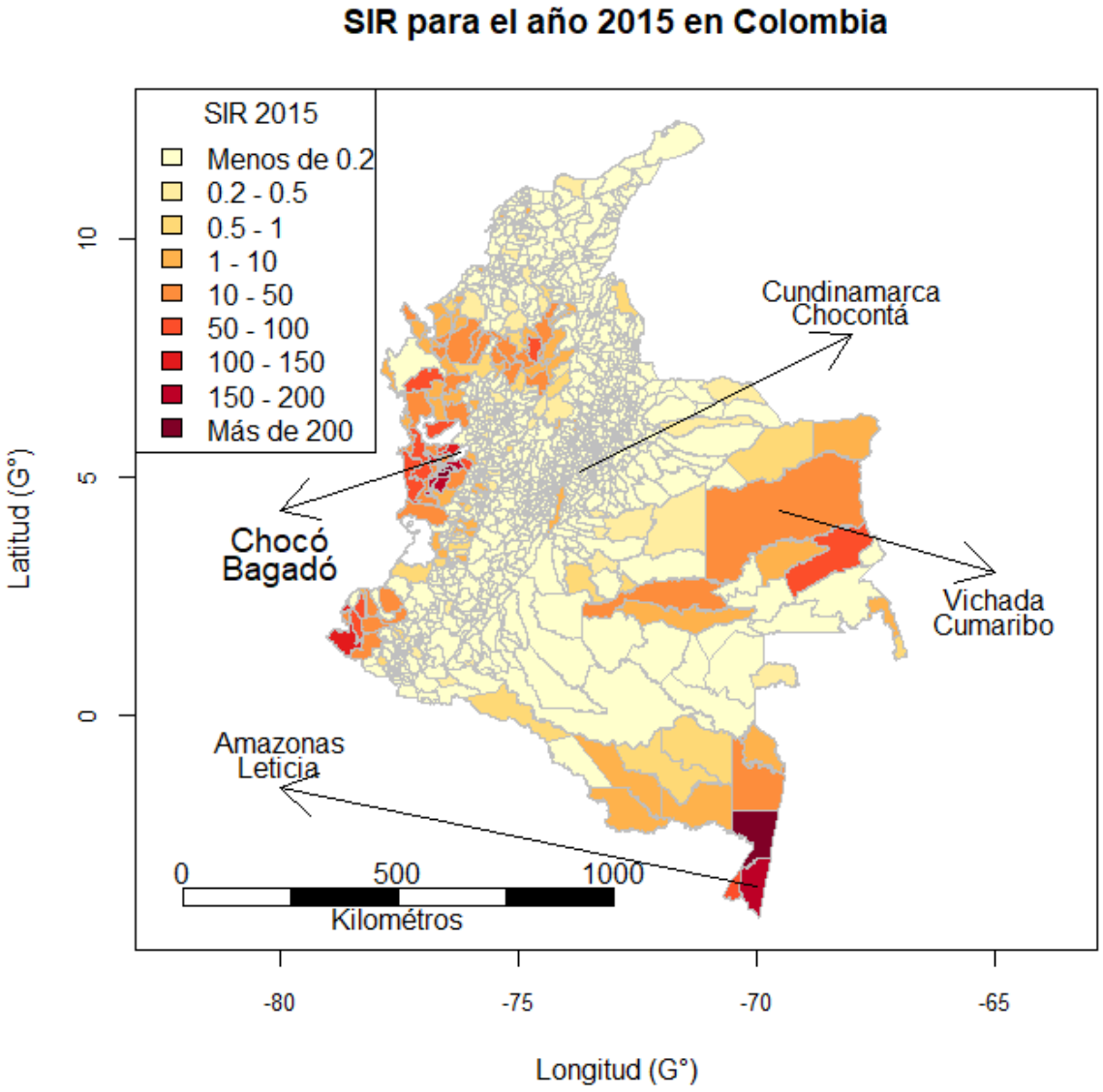


Figura 6.1: Tasa de incidencia estandarizada (SIR) para malaria en Colombia año 2015, software R

El índice de incidencia estandarizado (SIR) de acuerdo con la ecuación (31) es la razón entre los valores observados de la enfermedad y los cálculos relacionados con el valor esperado, esto determina que solamente los lugares en los cuales se presentaron

casos de la enfermedad tienen algún riesgo de contraerla nuevamente, puede ser más claro al comparar las figuras (6.1) y (5.2). Además, para los municipios seleccionados aleatoriamente, se observa que lo que determina el riesgo o incidencia de la enfermedad son los casos encontrados y no el valor esperado (validando nuevamente la ecuación (31))

	Cod_dane	Nom_munici	Mala_total	E	SIR
1	91001	LETICIA	1126	6.52	172.62
2	27073	BAGADÓ	1897	4.23	448.29
3	25183	CHOCONTÁ	3	27.14	0.11
4	99773	CUMARIBO	186	8.43	22.05

Cuadro 6.3: Comparativa entre el valor esperado y el SIR, software R

Se puede observar en la tabla (6.3) y la figura (6.1), los municipios que tienen valores de SIR igual a 1, lo que indica que los conteos observados son los mismos que los esperados, y qué municipios tienen SIR mayor (o menor) que 1, lo que indica que los conteos observados son mayores (o menores) de lo esperado. Los SIR pueden ser engañosos e insuficientemente confiables en municipios con poblaciones pequeñas. Por el contrario, los enfoques basados en modelos permiten incorporar covariables y tomar prestada información de los municipios vecinos para mejorar las estimaciones locales, lo que resulta en la suavización de valores extremos basados en tamaños de muestra pequeños. (Moraga, 2019)

Los valores de SIR más altos se encuentran en la costa pacifica presentando así el cluster mencionado en la enfermedad a modelar, además las regiones Orinoquía y Amazonia también tienen un alto riesgo de incidencia de la enfermedad.

6.2. Modelo Besag-York-Mollie (BYM)

La mayoría de las aplicaciones dependen del modelo BYM que se consideraba como el único modelo espacial completamente bayesiano que se usaba en las aplicaciones publicadas para el mapeo de enfermedades fuera de la literatura estadística. Estos muestran que el modelo BYM es uno de los modelos bayesianos jerárquicos más populares. (Samat y Mey, 2017)

Para los riesgos relativos, el modelo BYM incorpora efectos aleatorios debido a la heterogeneidad no estructurada (heterogeneidad correlacionada) y espacialmente estructurada (heterogeneidad no correlacionada) en el modelo log-lineal. La inclusión de estos efectos aleatorios permite suavizar los riesgos relativos a nivel global y local. (Besag, York, y Mollié, 1991)

Para la enfermedad de malaria, se supone que el número observado de casos de malaria (y_i) en el área i sigue una distribución de Poisson con media e_i :

$$y_i \quad Poisson(e_i)$$

donde e_i denota el número esperado de casos en la i ésima unidad geográfica, $_i$ es el riesgo relativo "verdadero"pero desconocido en el área i a estimar. En la siguiente etapa del modelo, la variabilidad del riesgo relativo $log(\ )$ se divide en tres componentes:

$$log(\ ) = \quad + u_i + v_i$$

donde es un nivel general del riesgo relativo, el efecto aleatorio espacial u_i refleja la heterogeneidad correlacionada y el efecto aleatorio v_i representa la heterogeneidad no correlacionada.

El modelado bayesiano necesita la especificación de distribuciones anteriores para efectos aleatorios. El modelo de distribución para la heterogeneidad correlacionada, u_i, depende de la ubicación geográfica y se supone que sigue una distribución normal con media cero y una varianza común (parámetro de precisión) T_u^2: (Wang et al., 2018)

$$u_i \quad N(0; T_u^2)$$

En un análisis completo del modelo bayesiano, se deben especificar distribuciones a priori para los parámetros de precisión T_u^2. Sin una estimación a priori de las precisiones de los efectos aleatorios, se recomiendan distribuciones con gran varianza. Estos siguen considerando las distribuciones $Gamma(0.5; 0.0005)$ que producen una probabilidad del 99 % para ambos, como lo sugieren Bernardinelli et al. (1995). Esta elección previa es menos informativa y permite que los datos de probabilidad dominen a la información a priori; por lo tanto, tendrá un efecto mínimo sobre la inferencia de riesgos relativos.

6.2.1. Poisson

A. Lawson et al. (1999) presenta el modelo de Poisson como una de las alternativas para realizar el mapeo de enfermedades, siendo este el más usado dada la naturaleza de conteo de los datos, sin embargo, esto se realiza cuando las enfermedades o mortandad de una enfermedad ocurre en cada uno de las áreas geográficas. La razón para presentar esta estimación es confirmar que la naturaleza de los datos y su inflación de ceros requiere un tratamiento especial puesto que los resultados no serían equiparables con el SIR inicial.

Se asume que los datos pueden describirse de la siguiente forma:

$$\begin{aligned} y_i | e_i; \ _i &\quad Poisson(e_i \ _i) \\ _i | \ ; &\quad Gamma(\ ; \ ) \end{aligned} \tag{64}$$

Donde los hiperparámetros y se establecen mediante información a priori informativa y no informativa descritas en el capítulo anterior, la función de enlace que se explica por las covariables es:

$$\begin{aligned} log(\ _i) = \ _0 + \ _1 ALCANT + \ _2 ACUE + \ _3 DENGUE + \ _4 CHICU + \\ \ _5 ZIKA + \ _6 BOSQUE + \ _7 PRECI + \ _8 ALT\_MED + \ _i \end{aligned} \tag{65}$$

Donde $_i = u_i + v_i$, aquí $_i \quad Normal(0; T_u^2)$ sigue la especificación iCAR del modelo BYM dejando que los datos dominen sobre la información a priori en el componente espacial.

	Media	Desv	0.025quant	0.5quant	0.975quant
$_0$	0.009521	0.001997	0.005600	0.009521	0.013440
ALCANT	-0.000112	0.000042	-0.000195	-0.000112	-0.000030
ACUE	0.000293	0.000032	0.000230	0.000293	0.000355
DENGUE	-0.000077	0.000003	-0.000083	-0.000077	-0.000070
CHICU	-0.000324	0.000048	-0.000419	-0.000324	-0.000229
ZIKA	-0.000416	0.000028	-0.000472	-0.000416	-0.000361
BOSQUE	0.098964	0.001925	0.095184	0.098964	0.102741
PRECI	0.117348	0.001928	0.113562	0.117348	0.121131
ALT_MED	-0.038739	0.001905	-0.042480	-0.038739	-0.035001

Cuadro 6.4: Parámetros estimados por R-INLA, distribución de Poisson con a priori NO informativas, software R

Como se observa en la tabla (6.4), los parámetros resultan ser estadísticamente significativos para el modelo Poisson inicial con información a priori NO informativa con un error de = 0:05.

Si se observa la figura (6.2) en donde se identifica el valor cero como una linea vertical, si la función de densidad dentro del intervalo de 0:975 no tiene esta linea intersectandola, el parámetro es significativo y concuerda con la información de la tabla (6.4).

Para observar el efecto positivo o negativo que realizan cada una de las covariables sobre la variable respuesta Malaria se debe realizar una transformación exponencial a la media de cada uno de los parámetros, esta transformación se da para quedar en escala original dada la ecuación de enlace establecida en la ecuación (65). También se puede

apreciar bajo la figura (6.2) donde si la función de densidad se encuentra a la derecha, el efecto será positivo o aumento del riesgo de registrar casos de Malaria y si se encuentra a la izquierda traduce a una reducción en la probabilidad de un recuento de Malaria.

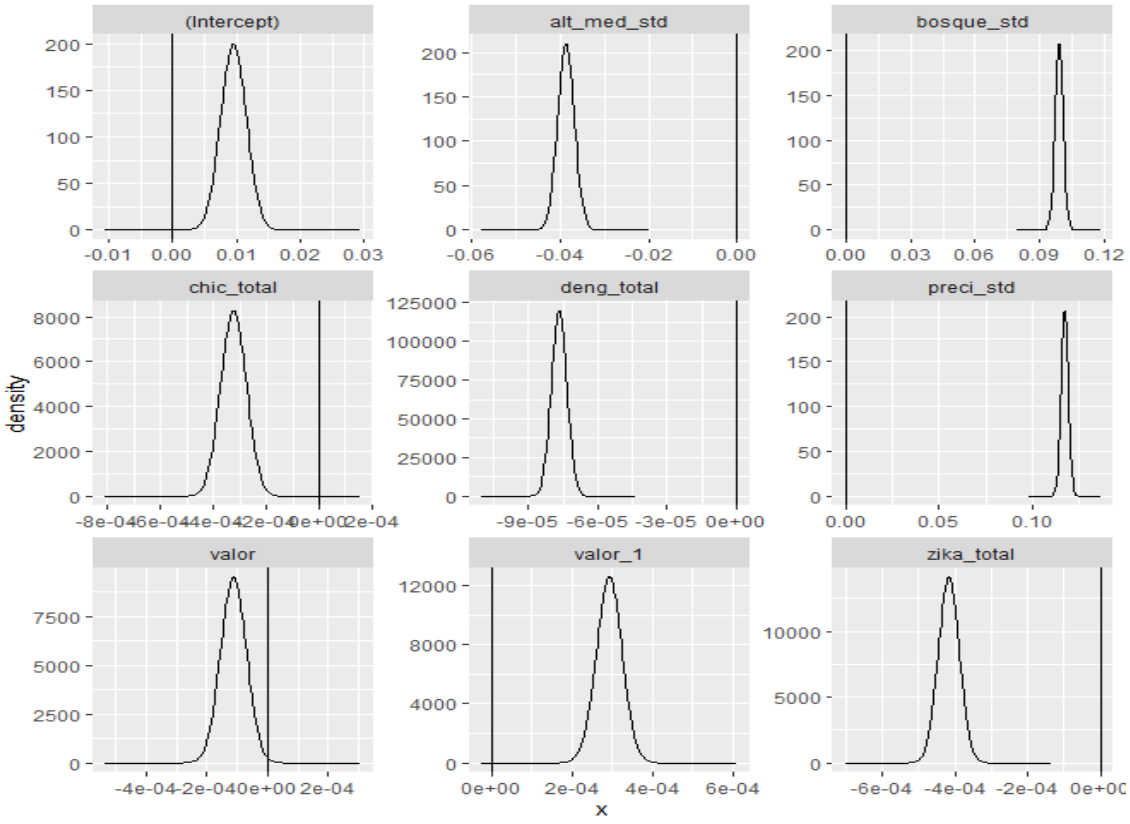


Figura 6.2: Densidad de los parámetros y significancia de acuerdo con el modelo Poisson NI, software R

Teniendo en cuenta la tabla (6.5), se encuentran los riesgos relativos medios asociados a cada una de las covariables significativas para el modelo Poisson, además se encontrar el riesgo relativo medio para todo Colombia.

Función	0	ALCANT	ACUE	DENGUE	CHICU
e	1.0095689	0.9998876	1.0002926	0.9999233	0.9996760
Función	ZIKA	BOSQUE	PRECI	ALT_MED	
e	0.9995839	1.1040291	1.1245127	0.9620032	

Cuadro 6.5: Escala natural de los parámetros para el modelo Poisson con a priori NO informativas

El riesgo relativo medio para el país esta determinado por $e^{0} = 0{:}95\,\%$ esto quiere

decir que cualquier municipio en Colombia tiene la probabilidad de 0.95 % de registrar un caso de Malaria.

Observando la figura (6.2) y la tabla (6.5), las variables ALCANT, DENGUE, CHICU, ZIKA y ALT_MED tienen un efecto negativo respecto a la presencia de casos de Malaria, esto quiere decir que un aumento en estas variables reduce la probabilidad de encontrar algún registro de Malaria. Sin embargo, sus aportes al riesgo relativo son menores al 1 %, a diferencia de la variable ambiental ALT_MED la cual si hay un aumento de esta, se reduce el riesgo en un 3.8 %.

Dentro de los efectos positivos del modelo se encuentran las variables ACUE, BOSQUE y PRECI donde lo más interesante son los aportes de la cobertura de bosque y el aumento de la precipitación media anual, dado que influyen en un 10.40 % y 12.45 % respectivamente, siendo las variables más influyentes en cuanto al registro de casos posibles de Malaria en el país.

La ecuación (65) estima un modelo con distribución Poisson teniendo en cuenta información a priori NO informativa (basándose en la regla de Je□rey's), sin embargo, se debe tener en cuenta también esta información extra muestral que data la experiencia del evento en cuestión, en este caso el número de casos reportados de Malaria en Colombia. De acuerdo con la misma especificación de la ecuación (65), pero incluyendo información a priori informativa se observa la significancia de cada una de las covariables a continuación:

	Media	Desv	0.025quant	0.5quant	0.975quant
$_0$	0.009352	0.002002	0.005421	0.009352	0.013279
ALCANT	-0.000110	0.000042	-0.000193	-0.000110	-0.000028
ACUE	0.000287	0.000032	0.000225	0.000287	0.000350
DENGUE	-0.000076	0.000003	-0.000083	-0.000076	-0.000070
CHICU	-0.000319	0.000048	-0.000414	-0.000319	-0.000224
ZIKA	-0.000410	0.000028	-0.000465	-0.000410	-0.000354
BOSQUE	0.097563	0.001929	0.093776	0.097563	0.101347
PRECI	0.115803	0.001932	0.112009	0.115803	0.119593
ALT_MED	-0.038069	0.001910	-0.041820	-0.038069	-0.034322

Cuadro 6.6: Parámetros estimados por R-INLA, distribución de Poisson con a priori informativas, software R

Se puede sacar algunas conclusiones sobre la fuerza de estos efectos simplemente mirando los resúmenes numéricos, pero es mejor verificar las densidades posteriores como se ve a continuación:

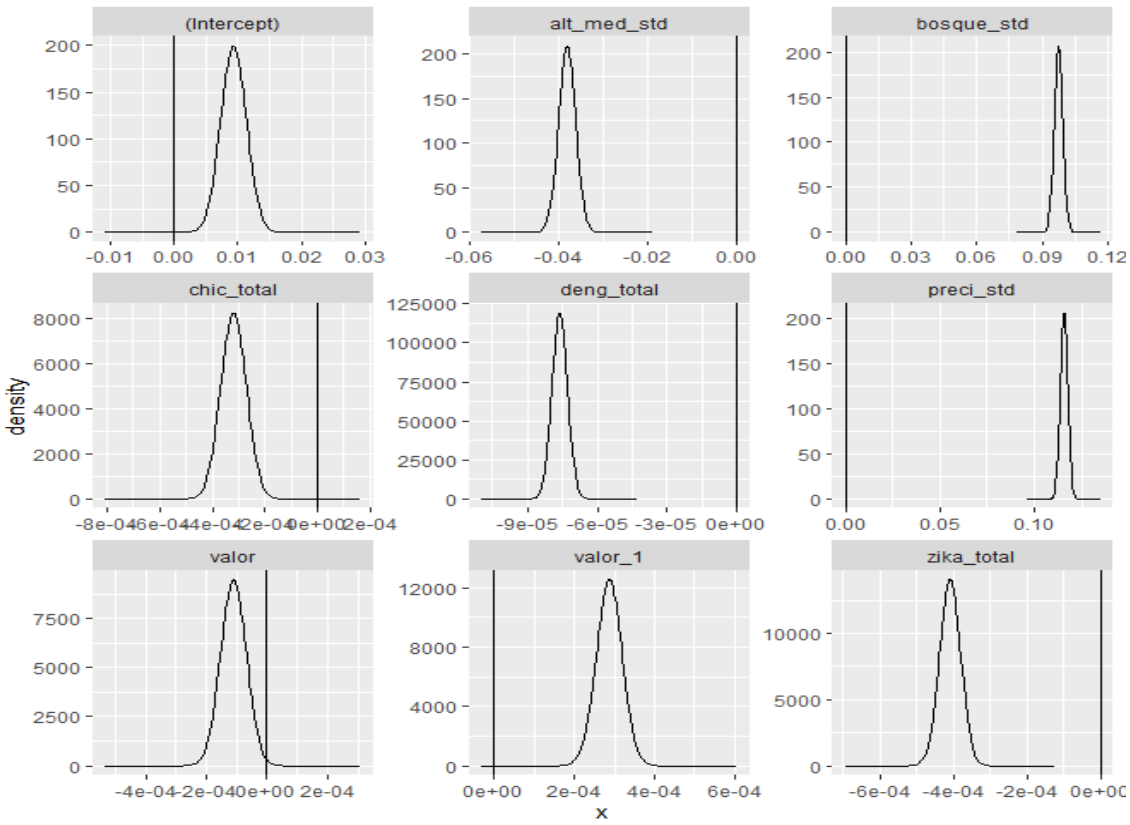


Figura 6.3: Densidad de los parámetros y significancia de acuerdo con el modelo Poisson I, software R

Como se pude apreciar en la tabla (6.6) y la figura (6.3) todos los parámetros resultan ser significativos para el modelo (con un = 0:05). Se realiza la transformación exponencial a la media de cada uno de los parámetros, para obtener así el aporte de cada covariable en términos porcentuales. Los resultados terminan siendo muy parecidos a los referidos al primer modelo especificados en la tabla (6.5).

Función	$_0$	ALCANT	ACUE	DENGUE	CHICU
e	1.0093977	0.9998900	1.0002874	0.9999239	0.9996808
Función	ZIKA	BOSQUE	PRECI	ALT_MED	
e	0.9995904	1.1024833	1.1227765	0.9626479	

Cuadro 6.7: Escala natural de los parámetros para el modelo Poisson con a priori informativas

El intercepto es el riesgo relativo medio para cada municipio en Colombia con un 0:9 % y varia dentro de un intervalo de (0:2 % 1:3 %), como se observa es positivo y se encuentra a la derecha de cero en la figura (6.3). Esto quiere decir que existe un riesgo positivo en cada municipio del país para tener registros de casos de Malaria. La cobertura

de alcantarillado (variable valor en la figura (6.3)) disminuye el riesgo relativo medio para cada municipio en 0:01 % esto quiere decir que por cada porcentaje que aumente la cobertura de alcantarillado en un municipio, su riesgo de registrar algún caso de Malaria disminuye en esa cantidad. A diferencia de la cobertura de alcantarillado, la cobertura de acueducto (variable valor_1 en la figura (6.3)) aumenta el riesgo relativo medio de registrar casos de la enfermedad, aunque sigue siendo un aumento mínimo del 0:02 %, así que, si aumenta la cobertura de acueducto el riesgo aumenta para la enfermedad de Malaria.

En cuanto a las enfermedades especificadas en el modelo, como lo son Dengue, Chicungunya y Zika, estas tienen riesgos negativos respecto a la variable respuesta, esto quiere decir que en cuanto más se registran casos de las otras tres enfermedades, menor probabilidad o riesgo hay de que exista algún caso de contagio para Malaria. Respecto a las variables ambientales, son las de mayor significancia e incidencia en el modelo, aquí, las variables Bosque y Precipitación aumentan el riesgo relativo medio para cada municipio en un 10:2 % y 12:3 % respectivamente, mientras que el aumento en Altura disminuye el riesgo e un 3:8 % lo que resulta ser cierto dado que a mayor altura, la población no se encuentra en riesgo de registrar casos de Malaria.

En la figura (6.4), se observa el riesgo relativo estimado por un modelo BYM de acuerdo a una distribución de Poisson con a priori informativas (la elección de este mapa se encuentra resumida en la tabla (6.21)). Estos riesgos son comparables con la estimación del SIR que se encuentra en la figura (6.1), se puede observar que los efectos espaciales y la sobredispersión de los datos afectaron en gran medida los valores del riesgo relativo estimado en el modelo, puesto que las zonas de mayor afectación por el SIR son la parte del Amazonia, pacífico colombiano y Orinoquía. Sin embargo, en el modelo BYM no se determina esos clusters, la parte del Amazonia que cuenta con un gran radio de incidencia, su riesgo relativo es bajo, en la parte del pacífico siendo la de mayor radio de incidencia desaparece totalmente esta puesto que dadas sus relaciones afectan la parte de los andes colombianos. El único en mantenerse constante es la parte del Orinoco dado que las relaciones espaciales del Amazonas pueden llegar a afectar esta región. No solo se da este resultado por las relaciones espaciales y la sobredispersión de los datos, también juegan un papel importante las covariables en donde las de mayor incidencia son las variables ambientales como se menciono anteriormente.

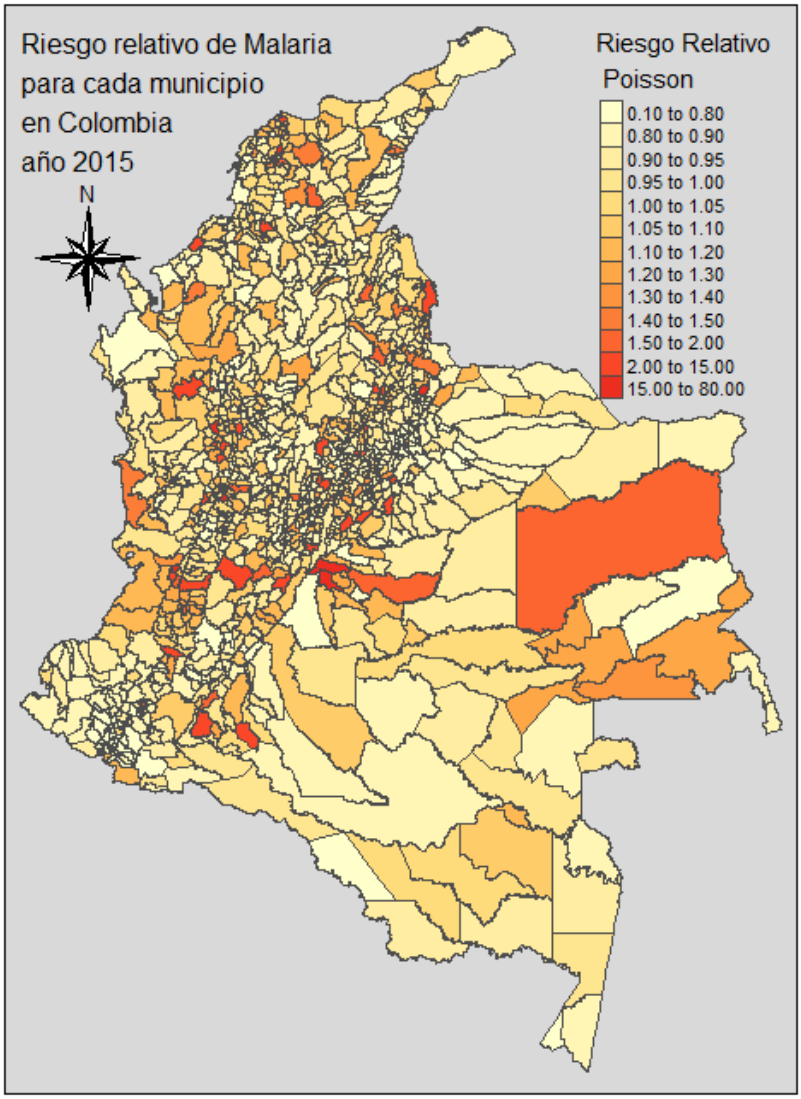


Figura 6.4: Riesgo relativo para malaria en Colombia año 2015 de acuerdo con el modelo Poisson con a priori Informativa, software R

6.2.2. Binomial

En el caso donde se examinan áreas pequeñas arbitrarias (como distritos censales, condados, zonas postales, municipios, distritos de salud), generalmente se observa un recuento de enfermedades dentro de cada unidad espacial. Dos aplicaciones que se adaptan bien a este enfoque son el análisis de las proporciones de nacimientos por sexo y el análisis de los resultados de los nacimientos (por ejemplo, anormalidades de nacimiento) en comparación con los nacimientos totales. A. Lawson (2013) indica que el modelo binomial

para los datos de conteo son útiles a la hora de mapear una enfermedad rara o cuando se exige un nivel de detalle muy grande en cuanto a unidades espaciales. El modelo binomial arroja como resultado la prevalencia de una enfermedad en cada área de acuerdo a los casos registrados y el total de la población. Se asume que los datos se describen por:

$$\begin{aligned} y_i | \ _i; \ _i \quad & Binomial(\ _i; \ _i) \\ \ _i | \ ; \quad & Beta(\ ; \ ) \end{aligned} \tag{66}$$

Donde $_i$ es el total de la población en la i ésima área y los efectos aleatorios espaciales iCAR se especifican en la transformación logística de junto con las covariables:

$$\begin{aligned} logit(\ _i) = \ _0 + \ _1 ALCANT + \ _2 ACUE + \ _3 DENGUE + \ _4 CHICU + \\ \ _5 ZIKA + \ _6 BOSQUE + \ _7 PRECI + \ _8 ALT\_MED + u_i + v_i \end{aligned} \tag{67}$$

Luego, la transformación de logit inversa de $_i$ devolverá la prevalencia específica posteriori del área, mientras que $exp(u_i + v_i)$ proporciona el "odds ratio" residual para cada área. Para verificar la significancia de cada una de las covariables, se realiza una prueba de hipótesis en donde la hipótesis nula $H_0 = \ _j \neq 0$ con un intervalo al 95% de confianza, esto puede verse reflejado en la tabla (6.8).

	Media	Desv	0.025quant	0.5quant	0.975quant
$_0$	-4.800666	0.051005	-4.898827	-4.801383	-4.698107
ALCANT	-0.030950	0.001137	-0.033201	-0.030945	-0.028733
ACUE	-0.031382	0.000890	-0.033141	-0.031378	-0.029646
DENGUE	-0.000457	0.000075	-0.000605	-0.000457	-0.000309
CHICU	-0.003907	0.001038	-0.005948	-0.003907	-0.001872
ZIKA	-0.002701	0.000646	-0.003973	-0.002700	-0.001438
BOSQUE	0.296839	0.045459	0.207552	0.296851	0.385977
PRECI	0.487591	0.046417	0.396417	0.487603	0.578605
ALT_MED	-0.134794	0.042287	-0.217824	-0.134793	-0.051844

Cuadro 6.8: Parámetros estimados por R-INLA, distribución Binomial con a priori NO informativas, software R

Todas las variables resultan ser significativas para el modelo y de acuerdo con su densidad, se puede verificar que dentro del intervalo de confianza no se encuentra el valor de la hipótesis nula reflejado en la figura (6.5).

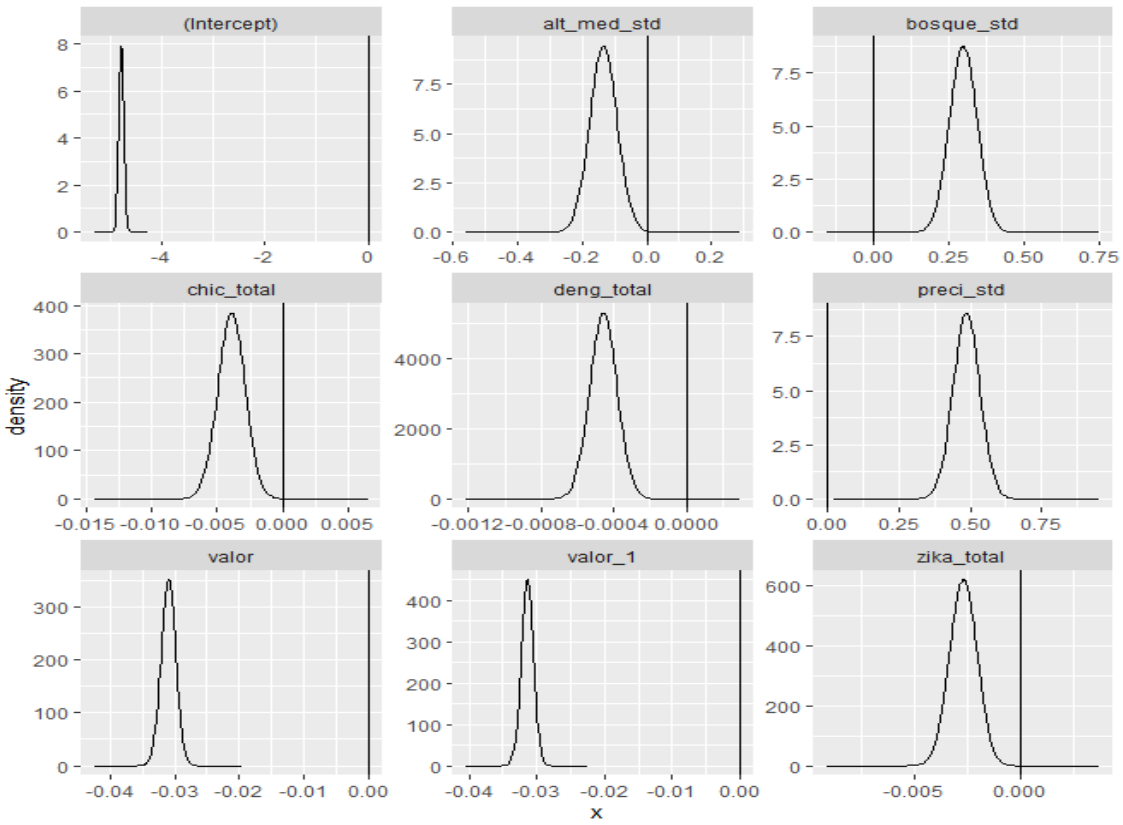


Figura 6.5: Densidad de los parámetros y significancia de acuerdo con el modelo Binomial NI, software R

Para obtener la probabilidad promedio de que cualquier municipio pueda tener algún registro de la enfermedad en cuestión (tasa de prevalencia media), se debe aplicar la función *logit* ^{1} en la función de densidad posteriori del intercepto ($_0$).

	Media	Desv	0.025quant	0.5quant	0.975quant
$\frac{e^{\ 0}}{1+\ _0}$	0.00816738	0.000411713	0.00788227	0.00815069	0.00902463

Cuadro 6.9: Escala natural de $_0$ para el modelo Binomial con a priori NO informativas

Como se observa en la tabla (6.9), la tasa de prevalencia media de la enfermedad para cualquier municipio en el país es bastante baja con un 0:0082 %, o un 99:28 % de disminución en cuanto a la prevalencia de la Malaria en el territorio colombiano.

Blangiardo y Cameletti (2015) establece que la probabilidad de que se de un recuento dada una covariable, se estima mediante la diferencia en el aumento de esta covariable

cuando se tiene el conteo y su estado inicial:

$$P(y_i = 1 | x_i + 1) = logit^{-1}(\beta_0 + \beta_1(x_i + 1))$$

$$P(y_i = 1 | x_i) = logit^{-1}(\beta_0 + \beta_1(x_i))$$

Entonces la diferencia entre ambas da la probabilidad de interés, sin embargo, existen las medidas de riesgo (odds ratio) que sirven para dar interpretación a los parámetros β. Usando la definición de logit, y considerando nuevamente el caso simple de un predictor, se puede definir el logaritmo del odds ratio:

$$log\_odds(x + 1) : log(\frac{p(y = 1 | x_i + 1)}{p(y = 0 | x_i + 1)}) = \beta_0 + \beta_1(x_i + 1)$$

$$log\_odds(x) : log(\frac{p(y = 1 | x_i)}{p(y = 0 | x_i)}) = \beta_0 + \beta_1(x_i)$$

Entonces la diferencia entre los dos logaritmos de odds es β_1 y, exponiéndola, se obtienen las medidas de asociación (OR):

$$exp(\frac{log\_odds(x + 1)}{log\_odds(x)}) = exp(\beta_1)$$

entonces el exponencial de β_1 representa el cambio en la razón de posibilidades cuando x_i aumenta en una unidad. Este segundo tipo de interpretación es el más utilizado en estudios epidemiológicos.

Función	ALCANT	ACUE	DENGUE	CHICU
e^{β}	0.9695244	0.9691058	0.9995432	0.9961007
Función	ZIKA	BOSQUE	PRECI	ALT_MED
e^{β}	0.9973028	1.3469895	1.6301420	0.8746776

Cuadro 6.10: Escala natural de los parámetros β (medida de asociación) para el modelo Binomial con a priori NO informativa

De acuerdo con la tabla (6.10) las variables de cobertura de acueducto y alcantarillado disminuyen la probabilidad de registrar casos de Malaria en el territorio colombiano en un 3:5 % y 3:09 % respectivamente, esto quiere decir que si se aumentan esas coberturas el riesgo de Malaria sobre el municipio decrece. En cuanto a las variables referentes a las enfermedades todas tienen un impacto inversamente proporcional, sin embargo, este no es de gran escala puesto que Chicungunya disminuye el riesgo en un 0:38 % siendo la

de mayor impacto. Las variables ambientales tienden a ser las de mayor significancia o aporte a la disminución o aumento del riesgo de la enfermedad, se apoya la hipótesis de a mayor altura disminuye el riesgo en un 12:53 % y las variable de Bosque y Precipitación, aumentan el riesgo en 34:7 % y 63:01 % respectivamente. Como se observa en la tabla (6.9), la tasa de prevalencia de la enfermedad en Colombia es muy baja por ende los resultados no superan el 25 % de prevalencia de la enfermedad. No se observa que la enfermedad tenga clusters, sin embargo, es curiosos que los valores más altos se hallen dentro de los andes colombianos puesto que la variable de altura media disminuye el riesgo en un 12 % aproximadamente.

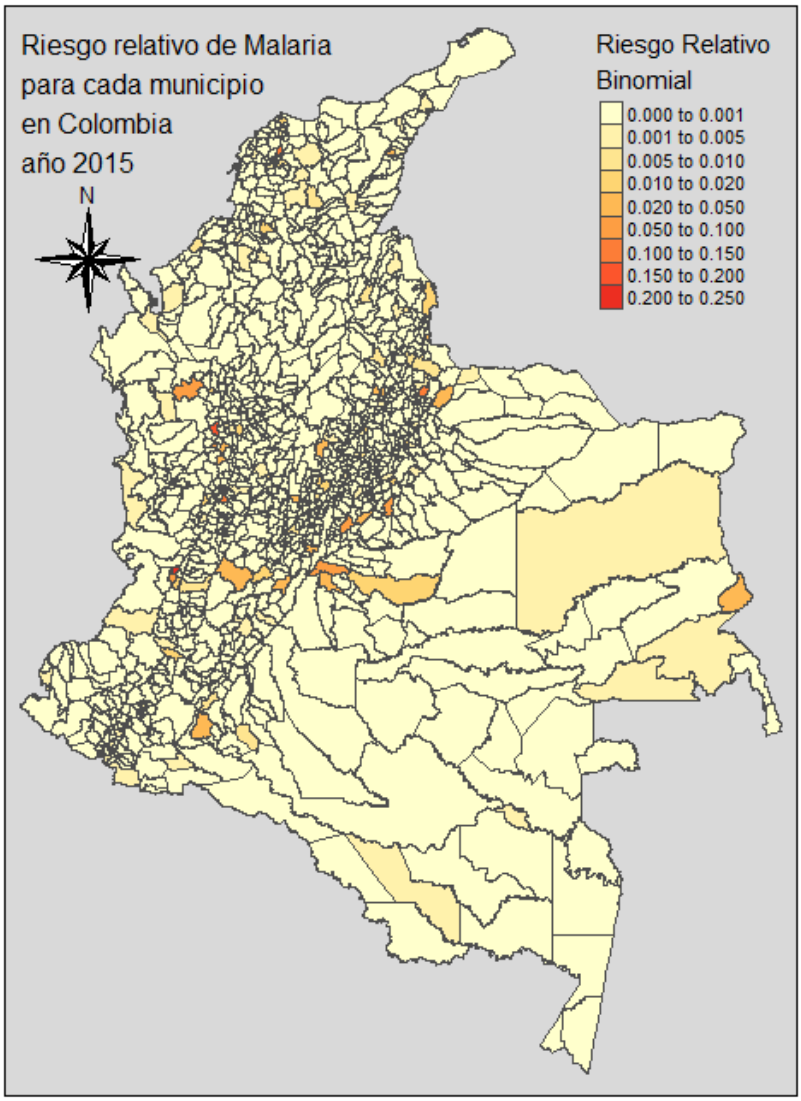


Figura 6.6: Riesgo relativo para malaria en Colombia año 2015 de acuerdo con el modelo Binomial con a priori No Informativa, software R

6.2.3. Binomial Negativo (NB)

El modelo de regresión binomial negativa es un modelo estadístico verdaderamente inusual. Típicamente, aquellos en la comunidad estadística se refieren al binomio negativo como un modelo único, como al referirse a la regresión de Poisson, la regresión logística o la regresión probit. El modelo binomial negativo, como modelo de mezcla de Poisson-gamma, es apropiado para usar cuando se cree que la sobredispersión en un modelo de Poisson de otro modo toma la forma de una distribución o forma gamma. (Hilbe, 2011) La estimación de este modelo tendrá como finalidad poder dar explicación a la sobredispersión observada en el capítulo anterior. Así pues, se asume que los datos son descritos por:

$$\begin{aligned} y_i &\quad NB(\ _i; k) \\ _i &\quad PC{:}Gamma(\ ^1; \ ^1) \end{aligned} \tag{68}$$

Donde $_i$ es el valor esperado (media) de y_i que tiene distribución experimental a priori PC.Gamma. k se le denomina parámetro de dispersión. (A. Zuur et al., 2009)

La función de enlace que se explica por el conjunto de covariables es:

$$\begin{aligned} log(\ _i) = \ _0 + \ _1 ALCANT + \ _2 ACUE + \ _3 DENGUE + \ _4 CHICU + \\ _5 ZIKA + \ _6 BOSQUE + \ _7 PRECI + \ _8 ALT\_MED + u_i + v_i \end{aligned} \tag{69}$$

De acuerdo con la información a priori no informativa (= 7) establecida en el capítulo anterior se ajusta el modelo NB por medio de la metodología INLA y da como resultado:

	Media	Desv	0.025quant	0.5quant	0.975quant
$_0$	-0.007015	0.001972	-0.010886	-0.007015	-0.003147
ALCANT	-0.000092	0.000041	-0.000173	-0.000092	-0.000012
ACUE	-0.000099	0.000031	-0.000160	-0.000099	-0.000038
DENGUE	-0.000045	0.000018	-0.000081	-0.000045	-0.000010
CHICU	-0.000320	0.000360	-0.001027	-0.000320	0.000387
ZIKA	-0.000123	0.000148	-0.000415	-0.000123	0.000168
BOSQUE	-0.007461	0.001845	-0.011084	-0.007461	-0.003841
PRECI	-0.007562	0.001861	-0.011215	-0.007562	-0.003912
ALT_MED	0.006395	0.001820	0.002821	0.006395	0.009965

Cuadro 6.11: Parámetros estimados por R-INLA, distribución de NB con a priori NO informativas, software R

Como se observa en la tabla (6.11), las variables CHICU y ZIKA resultan ser no significativas para el modelo, sin embargo, las estimaciones resultan ser diferentes a lo

encontrado en los modelos previos.

Todas las variables con excepción de ALT_MED resultan tener un impacto negativo en cuanto a la presencia o riesgo de la enfermedad. Esto quiere decir que un aumento de cualquier variable traduce a una reducción porcentual de encontrar un recuento de la enfermedad en cuestión. Pese a esto, de acuerdo con el IPA las consideraciones para establecer ese índice es que las poblaciones en riesgo son aquellas que se encuentran a menos de 1600msnm, pero el modelo dice que al aumentar la altura, aumentara la probabilidad de un recuento de caso. Esto puede ser por la inclusión de las variables no significativas para el modelo, así que se procede a estimar un nuevo modelo de acuerdo con:

$$\begin{aligned} log(\ _i) = \ _0 + \ _1 ALCANT + \ _2 ACUE + \ _3 DENGUE + \\ \ _4 BOSQUE + \ _5 PRECI + \ _6 ALT\_MED + u_i + v_i \end{aligned} \tag{70}$$

La significancia de las covariables de acuerdo con la especificación a priori no informativa se presenta a continuación:

	Media	Desv	0.025quant	0.5quant	0.975quant
$_0$	-0.007355	0.001954	-0.011191	-0.007355	-0.003522
ALCANT	-0.000097	0.000041	-0.000177	-0.000097	-0.000017
ACUE	-0.000102	0.000031	-0.000164	-0.000102	-0.000041
DENGUE	-0.000048	0.000018	-0.000084	-0.000048	-0.000013
BOSQUE	-0.007355	0.001844	-0.010976	-0.007355	-0.003738
PRECI	-0.007417	0.001858	-0.011064	-0.007417	-0.003772
ALT_MED	0.006647	0.001808	0.003097	0.006647	0.010194

Cuadro 6.12: Parámetros estimados por R-INLA, distribución de NB con a priori NO informativas segundo modelo, software R

Posterior a estimar el segundo modelo sin incluir las variables menos significativas, se observa según la tabla (6.12) que los efectos directa e inversamente proporcional no han variado, el riesgo de que cualquier municipio en el país tenga un recuento de Malaria es negativo, además de que las 5 primeras covariables también disminuyan el riesgo, siendo así la ALT_MED la única variable directamente proporcional, sin embargo, esto no es factible de acuerdo a las zonas endémicas localizadas para la enfermedad.

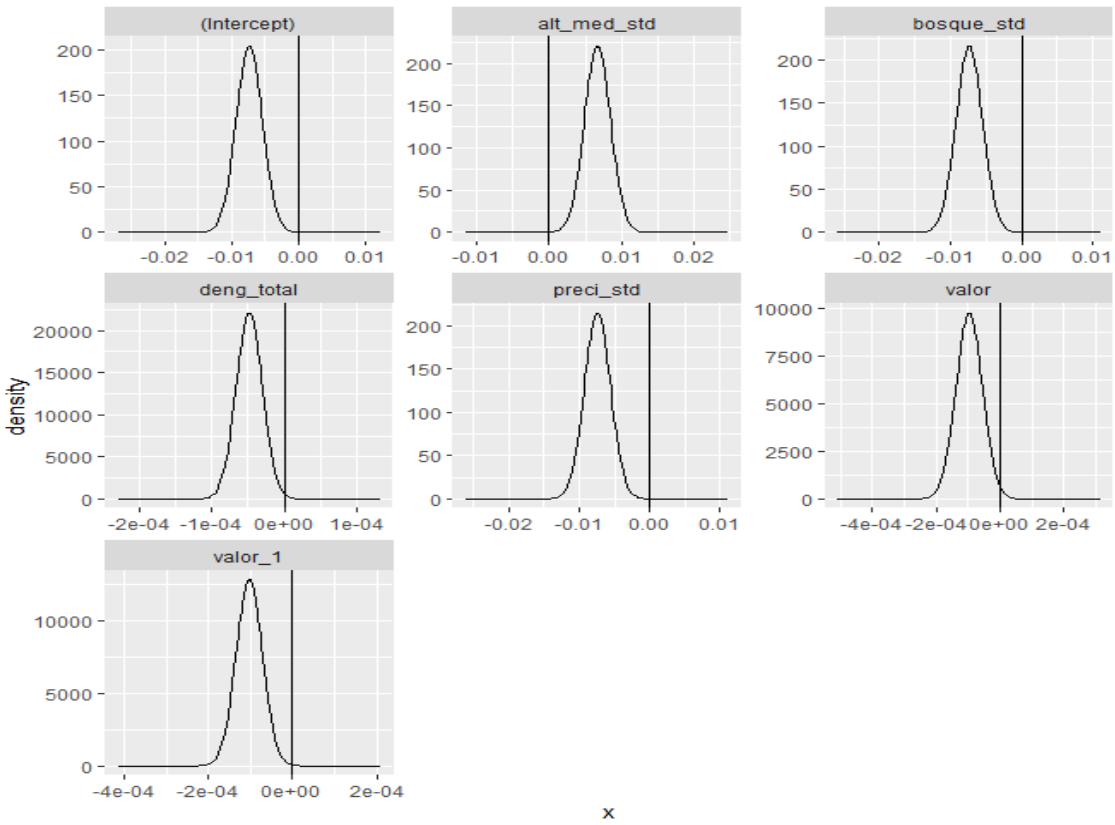


Figura 6.7: Densidad de los parámetros y significancia de acuerdo con el modelo NB NI, software R

Para poder interpretar el aporte de cada covariable como respuesta de la Malaria, se debe realizar una exponencial de la media marginal de cada una de estas para volver a escala original (de acuerdo con la ecuación (69)).

Función	$_0$	ALCANT	ACUE	DENGUE
e	0.9926743	0.9999031	0.9998976	0.9999519
Función	BOSQUE	PRECI	ALT_MED	
e	0.9926733	0.9926125	1.0066706	

Cuadro 6.13: Escala natural de los parámetros para el modelo NB con a priori NO informativas

Se aprecian los efectos de cada uno de los betas en escala normal, teniendo en cuenta esto se observa que para cualquier municipio de Colombia se tiene un riesgo relativo medio negativo de 0:73 % que es casi nulo así que la enfermedad no debería aparecer como recuento (se debería tratar como una enfermedad rara según este modelo). El aporte de cada covariable también resulta ser casi que significativo a la respuesta de Malaria. Las coberturas de Acueducto y Alcantarillado disminuyen la presencia de Malaria en un

0:0097 % y 0:010 % respectivamente. La enfermedad Dengue resulta ser la única variable de este tipo en ser significativa para el modelo, sin embargo, como variable respuesta su contribución al recuento de casos de Malaria es el más bajo con un riesgo negativo de 0:005 %. Las variables ambientales son las de mayor contribución, pero en realidad estas tienen un aporte distinto a lo que el comportamiento de estas refleja en la vida real. La variable bosque y precipitación tienen una contribución negativa al conteo de Malaria, esto quiere decir que por un aumento en la cobertura de bosque o en la precipitación media sobre un municipio, la probabilidad de que aparezca un recuento de Malaria disminuye en 0:73 % y 0:74 % respectivamente y si aumenta la altura media también aumentará el riesgo en un 0:67 %.

Como se puede apreciar en la figura (6.7), las variables resultan ser significativas con un = 5 %. Para determinar que tan influyente es la información a priori, se estima un modelo similar al de la ecuación (69) pero cambiando el hiper parámetro de $_i$ en la ecuación (68). Se puede observar la significancia de las covariables a continuación:

	Media	Desv	0.025quant	0.5quant	0.975quant
$_0$	-0.059893	0.006120	-0.071908	-0.059893	-0.047887
ALCANT	-0.000731	0.000126	-0.000978	-0.000731	-0.000484
ACUE	-0.000829	0.000097	-0.001019	-0.000829	-0.000640
DENGUE	-0.000366	0.000055	-0.000474	-0.000366	-0.000258
CHICU	-0.002127	0.001062	-0.004212	-0.002127	-0.000044
ZIKA	-0.000956	0.000440	-0.001820	-0.000956	-0.000092
BOSQUE	-0.058952	0.005396	-0.069547	-0.058952	-0.048367
PRECI	-0.059247	0.005436	-0.069921	-0.059247	-0.048584
ALT_MED	0.059495	0.005397	0.048898	0.059495	0.070083

Cuadro 6.14: Parámetros estimados por R-INLA, distribución de NB con a priori informativa, software R

A diferencia de la tabla (6.11) todas las variables resultan ser significativas, aunque sigue existiendo la particularidad de que todas las variables con excepción de ALT_MED reducen el riesgo de que exista el registro de la enfermedad Malaria.

Como se puede apreciar en la figura (6.8), la única curva de densidad a la derecha del cero es la de ALT_MED que corrobora que sigue siendo la variable que aumenta el riesgo de registrar un caso de Malaria sobre un determinado municipio. Las variables CHICU y ZIKA resultan significativas, sin embargo a continuación se puede observar que pueden ser prescindibles por estar tan cercanas al cero.

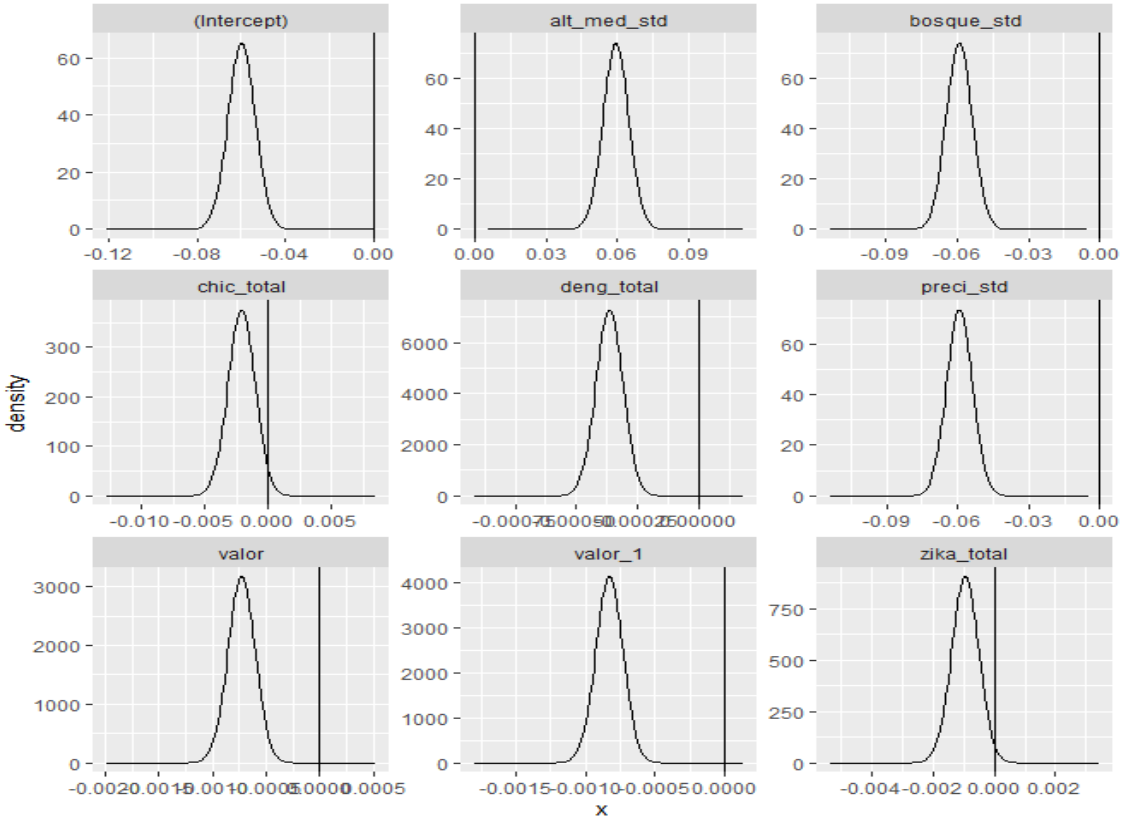


Figura 6.8: Densidad de los parámetros y significancia de acuerdo con el modelo NB I, software R

Se realiza la transformación exponencial a la media de cada uno de los parámetros, para obtener así el aporte de cada covariable en términos porcentuales. Los resultados terminan siendo muy parecidos para cada una de las covariables referidas en el primero modelo especificados en la tabla (6.13):

Función	$_0$	ALCANT	ACUE	DENGUE	CHICU
e	0.9418833	0.9992695	0.9991713	0.9996340	0.9978759
Función	ZIKA	BOSQUE	PRECI	ALT_MED	
e	0.9990446	0.9427653	0.9424876	1.0613157	

Cuadro 6.15: Escala natural de los parámetros para el modelo NB con a priori informativa

De acuerdo con la tabla (6.15), el riesgo relativo medio para Colombia es negativo con un 5:8 %, esto quiere decir que en Colombia no debería presentarse recuento de esta enfermedad o visto como una enfermedad “rara”. Al igual que el modelo sin información a priori (a priori no informativa), las variables que más aportan al cambio del riesgo son las variables ambientales en donde la cobertura de bosque y la precipitación media reducen el

riesgo de registrar un caso de Malaria en 5:7 % y 5:8 % siendo esto mayor que el modelo de la tabla (6.13), y la altura media aumenta este riesgo en 6:1 % lo que en realidad no se observa puesto que los municipios con mayor altitud no cuentan con casos registrados. En la figura (6.9), se observa que el riesgo relativo para Malaria no es superior al 60 % esto no es un indicador alentador puesto que existen municipios donde el recuento supera los 2000 casos. También se puede apreciar que la mayoría de las variables disminuyen el riesgo de Malaria, sin embargo, a mayor altura aumenta lo que contradice el IPA dado que el contagio no se da a más de 1600 msnm.

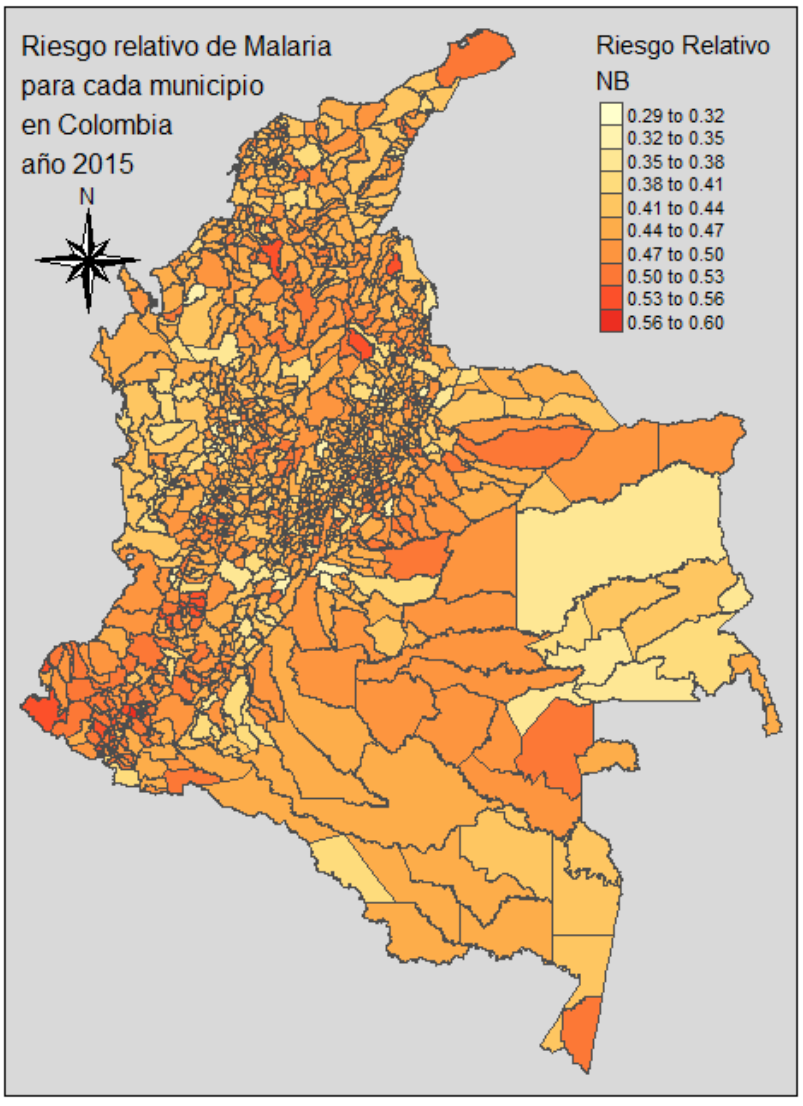


Figura 6.9: Riesgo relativo para malaria en Colombia año 2015 de acuerdo con el modelo Binomial Negativo con a priori Informativa, software R

6.2.4. Poisson inflado con ceros (ZIP)

Como se puede apreciar en la tabla (5.1) y en el mapa de caja (5.2), los casos de malaria en el territorio colombiano para el año 2015 se encuentran con un número excesivo de ceros en su reporte. La implicación de esto es que se observan pocos casos dentro del área de estudio o, para áreas pequeñas, son comunes los recuentos cero. En este caso, la distribución espacial de los casos a menudo formará grupos aislados. La principal pregunta que plantea esta situación es si ¿los modelos estándar para el mapeo de enfermedades se mantienen cuándo surge tal escasez de datos?. En algunas aplicaciones, la especificación de los modelos de Poisson o Binomial para los datos de conteo puede ser inapropiada debido al exceso de ceros en los datos en comparación con lo que se espera del modelo (sobredispersión). Para superar este problema, se pueden especificar los llamados modelos de cero inflado. Estos son una mezcla de dos componentes: una masa puntual en cero y una distribución de conteo. De esta manera, tales modelos distinguen entre ceros estructurales, para unidades donde cero es el único valor observable, y ceros de muestra, para unidades en las que observamos un cero, pero otros valores también podrían haberse registrado. (Blangiardo y Cameletti, 2015)

Para que la metodología sea equiparable al SIR, se debe realizar un tratamiento especifico a la sobre dispersión del parámetro. Siendo así suponga que cada observación se distribuye de la siguiente forma:

$$\begin{aligned} y_i &\sim ZIP(\ _i; p) \\ \ _i &= logit(p) \\ p &\sim Beta(\ ;\ ) \end{aligned} \tag{71}$$

$$y_i | e_i;\ _i \sim pPois(0) + (1 - p)Pois(e_i\ _i)$$

El parámetro es igual a $e_i\ _i$ y este se determina mediante $log(\frac{p}{1-p}) = logit(p)$. p es la proporción de ceros que se tienen y se distribuye mediante una función Beta. Luego, condicional a que y_i no sea un cero estructural, la transformación logarítmica de $_i$ se modela como:

$$\begin{aligned} log(\ _i) = \ _0 + \ _1 ALCANT + \ _2 ACUE + \ _3 DENGUE + \ _4 CHICU + \\ \ _5 ZIKA + \ _6 BOSQUE + \ _7 PRECI + \ _8 ALT\_MED + u_i + v_i + log(e_i) \end{aligned} \tag{72}$$

El resultado de ajustar un modelo como el de la ecuación (72) se encuentra a continuación:

	Media	Desv	0.025quant	0.5quant	0.975quant
$_0$	-1.642043	0.122407	-1.881866	-1.642220	-1.401454
ALCANT	-0.008065	0.002394	-0.012771	-0.008063	-0.003374
ACUE	0.000824	0.002195	-0.003481	0.000823	0.005134
DENGUE	-0.000121	0.000078	-0.000274	-0.000121	0.000032
CHICU	-0.000834	0.001199	-0.003202	-0.000829	0.001504
ZIKA	-0.003981	0.001007	-0.006038	-0.003953	-0.002084
BOSQUE	0.877327	0.076966	0.726095	0.877368	1.028190
PRECI	0.598195	0.079523	0.442198	0.598146	0.754324
ALT_MED	-0.601790	0.088470	-0.775513	-0.601783	-0.428269

Cuadro 6.16: Parámetros estimados por R-INLA, distribución de ZIP con a priori NO informativas, software R

Como se observa en la tabla (6.16), todas las variables son significativas con un = 5 %, sin embargo, las variables ACUE, CHICU y DENGUE resultan no serlo, esto se puede evidenciar de forma gráfica de acuerdo a su función de densidad de probabilidad.

De acuerdo con la tabla (6.16) y la gráfica (6.10), solo las variables BOSQUE y PRECI resultan ser directamente proporcional a la enfermedad, esto quiere decir que una mayor cobertura de bosque y un aumento en la precipitación sobre un determinado territorio podrían influir en la aparición de casos registrados de la enfermedad Malaria. Esto podría explicar el por qué de los altos registros en el Chocó y el Amazonia, además que al aumentar la cobertura de alcantarillado y acueducto disminuya la probabilidad de riesgo en los municipios, también se observa como a mayor altura menos riesgo de registrar el conteo de la enfermedad se obtiene.

Para evidenciar los efectos positivos y negativos, en la gráfica (6.10), se establece que función de densidad se encuentra a la derecha de cero (efecto positivo o directamente proporcional) y la que se encuentra a la izquierda (efecto negativo o inversamente proporcional).

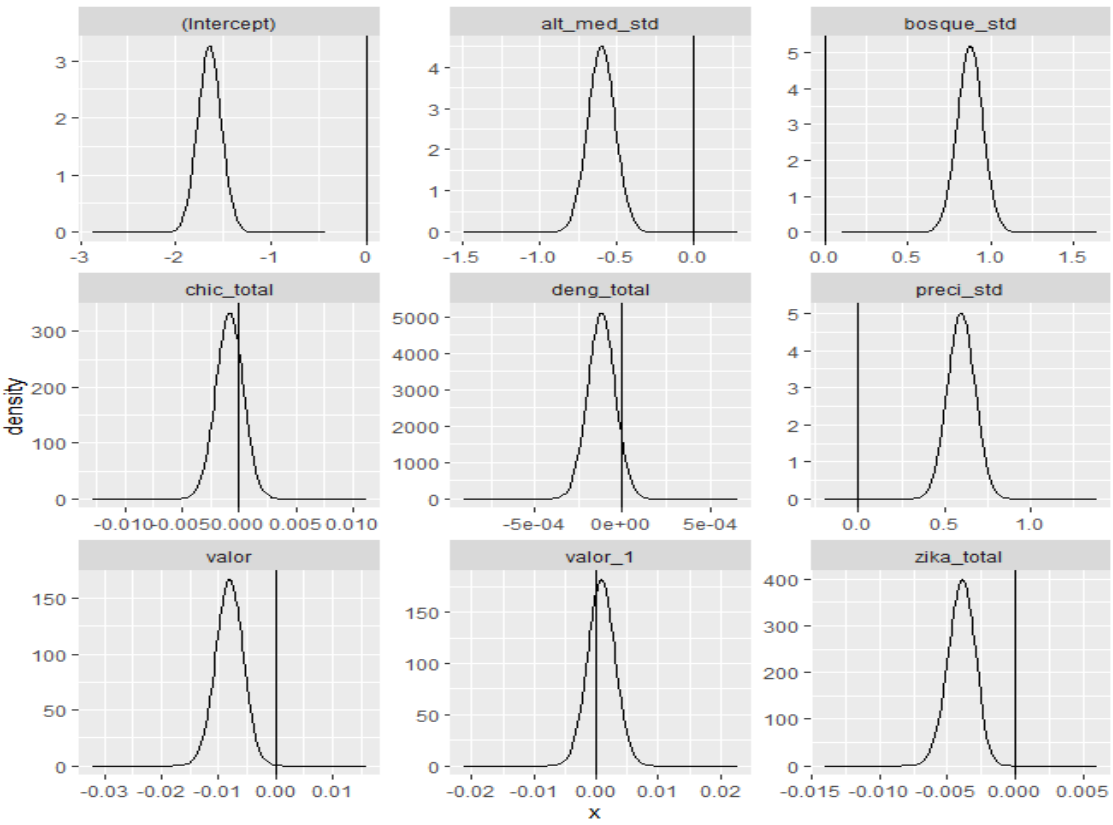


Figura 6.10: Densidad de los parámetros y significancia de acuerdo con el modelo ZIP NI, software R

Dado que existen tres variables no explicativas para el modelo, se descartan como variables que puedan influir en la presencia de Malaria, siendo así, se especifica un nuevo modelo de la forma:

$$\begin{aligned} log(\ _i) = \ _0 + \ _1 ALCANT + \ _2 ZIKA + \ _3 BOSQUE \\ + \ _4 PRECI + \ _5 ALT\_MED + u_i + v_i + log(e_i) \end{aligned} \tag{73}$$

Ajustando el modelo mediante INLA se obtiene la siguiente significancia de las covariables:

	Media	Desv	0.025quant	0.5quant	0.975quant
0	-1.625486	0.105181	-1.831616	-1.625619	-1.418806
ALCANT	-0.007930	0.001852	-0.011568	-0.007929	-0.004297
ZIKA	-0.004272	0.000972	-0.006250	-0.004247	-0.002434
BOSQUE	0.877782	0.076537	0.727405	0.877818	1.027816
PRECI	0.606177	0.078929	0.451347	0.606128	0.761142
ALT_MED	-0.598724	0.087828	-0.771168	-0.598722	-0.426447

Cuadro 6.17: Parámetros estimados por R-INLA, distribución de ZIP con a priori NO informativas (segundo modelo estimado), software R

Como se había observado en la tabla (6.16) los únicas variables en ser directamente proporcionales en la presencia de la enfermedad son las variables BOSQUE y PRECI, al igual que corroborar que un aumento en altitud genera que el riesgo disminuya.

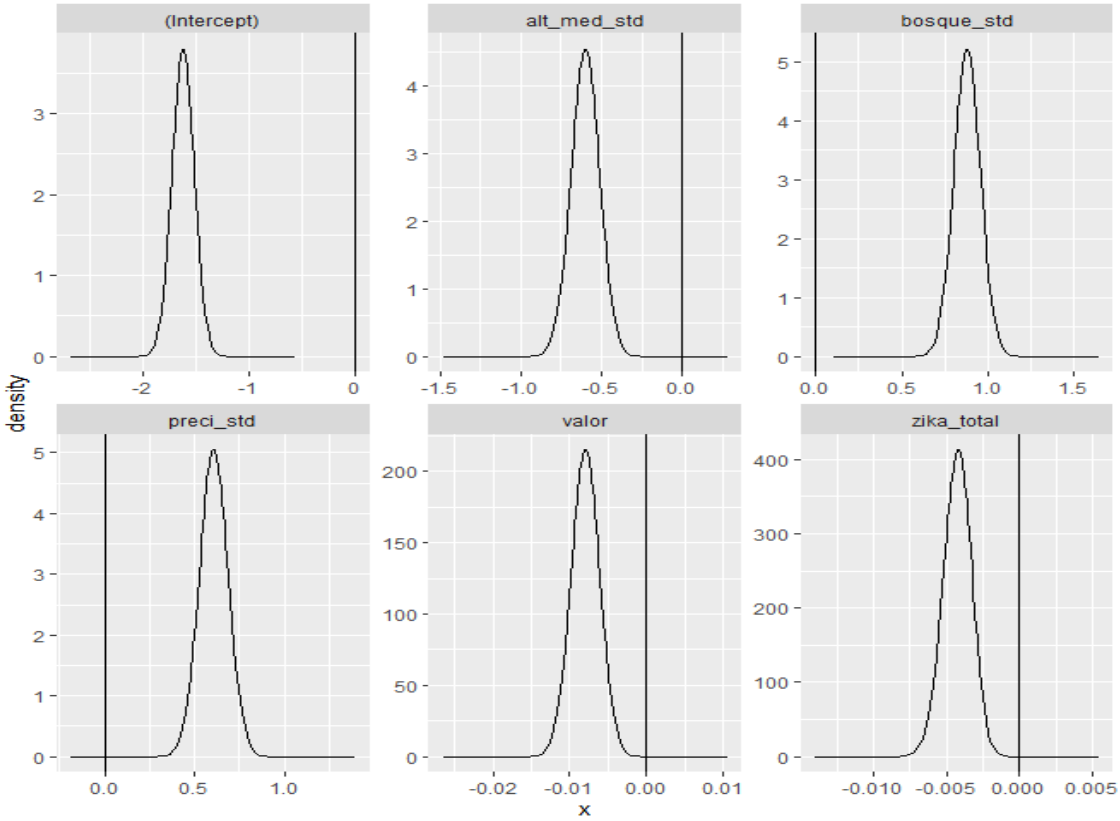


Figura 6.11: Densidad de los parámetros y significancia de acuerdo con el segundo modelo ZIP NI, software R

Para interpretar el aporte que realiza cada variable al posible recuento de Malaria se realiza la transformación exponencial a cada uno de los betas, los resultados se presentan a continuación:

Función	$_0$	ALCANT	ZIKA
e	0.1979079	0.9921033	0.9957378
Función	BOSQUE	PRECI	ALT_MED
e	2.4126115	1.8391258	0.5516342

Cuadro 6.18: Escala natural de los parámetros para el modelo ZIP con a priori NO informativa

Se observa en la tabla (6.18) que el riesgo relativo para Colombia es de 80:2 % negativo, esto quiere decir que su ocurrencia dentro de cualquier municipio del país es poco común lo que sería tema de discusión puesto que si no hay casos registrados sobre todo el territorio si existen conteos muy elevados para un municipio determinado. La cobertura de alcantarillado disminuye el recuento de la enfermedad sobre el territorio, esto quiere decir que un aumento en la cobertura hace que baje la probabilidad de riesgo sobre un departamento en 0:79 % lo que resulta de mucha utilidad para reducir la cantidad de casos. Zika resulta ser también inversamente proporcional a el conteo de Malaria con un 0:43 % de disminución lo que en realidad dice que un paciente que tiene Zika no puede tener Malaria. Las variables ambientales resultan ser las más interesantes puesto que son las que aportan mayor impacto a la presencia de los casos de infección por Malaria donde la cobertura de bosque aumenta la probabilidad en un 141:3 %, lo que conlleva a pensar el por qué en la Amazonia se encuentran casos tan elevados. La precipitación juega un papel importante en donde esta aumenta la probabilidad en un 83:9 % lo que daría una explicación a los departamentos del Chocó, y el aumento en altura evidencia una disminución del riesgo relativo con un 44:8 % donde a mayor altura como en las cordilleras disminuye la probabilidad del evento. Otro dato interesante que arroja este modelo es la probabilidad del parámetro cero que ajusta de acuerdo a las covariables, siendo así, $p = 0{:}6671$, lo que quiere decir que la probabilidad de ceros muéstrales para el modelo ronda el 66:7 %.

Como se aprecia en la ecuación (73), este sería el modelo estimado para una distribución Poisson truncada en cero con a priori NO informativas, sin embargo, ¿qué pasaría si se incluye la información extra muestral de la información a priori?, para esto se estima un modelo que incluya esta información mediante la misma metodología.

Partiendo de la ecuación (71) y la información extra muestral en los hiperparámetros y , se establece un modelo mediante la transformación logarítmica de $_i$ como:

$$\begin{aligned} log(\ _i) = \ _0 + \ _1 ALCANT + \ _2 ACUE + \ _3 DENGUE + \ _4 CHICU + \\ \ _5 ZIKA + \ _6 BOSQUE + \ _7 PRECI + \ _8 ALT\_MED + u_i + v_i + log(e_i) \end{aligned} \tag{74}$$

Se observa la significancia de cada una de las variables a continuación:

	Media	Desv	0.025quant	0.5quant	0.975quant
$_0$	0.023410	0.002018	0.019447	0.023410	0.027369
ALCANT	0.000151	0.000042	0.000068	0.000151	0.000234
ACUE	0.000453	0.000032	0.000390	0.000453	0.000516
DENGUE	-0.000060	0.000003	-0.000066	-0.000060	-0.000053
CHICU	-0.000284	0.000049	-0.000380	-0.000284	-0.000188
ZIKA	-0.000276	0.000029	-0.000334	-0.000276	-0.000219
BOSQUE	0.089489	0.001941	0.085678	0.089489	0.093297
PRECI	0.106628	0.001945	0.102810	0.106628	0.110443
ALT_MED	-0.033422	0.001924	-0.037200	-0.033422	-0.029646

Cuadro 6.19: Parámetros estimados por R-INLA, distribución ZIP con a priori informativas, software R

A diferencia de lo observado en la tabla (6.16) la información a priori tuvo incidencia en la significancia de las covariables, en donde absolutamente todas las variables son significativas al 95 % de confianza.

Aunque todas las variables fuesen significativas, sus efectos positivos o negativos son ligeramente distintos en donde se establece según la tabla (6.19) que para todo Colombia existe un riesgo relativo positivo, al igual que aumentar la cobertura de Alcantarillado y Acueducto generaría un plausible aumento del riesgo de hallar casos de Malaria.

Función	$_0$	ALCANT	ACUE	DENGUE	CHICU
e	1.0236880	1.0001514	1.0004533	0.9999404	0.9997164
Función	ZIKA	BOSQUE	PRECI	ALT_MED	
e	0.9997236	1.0936177	1.1125224	0.9671326	

Cuadro 6.20: Escala natural de los parámetros para el modelo ZIP con a priori informativa

A diferencia con las a priori no informativas, el riesgo relativo medio para Colombia de acuerdo con un modelo ZIP es de 2:37 % siendo este positivo. Las variables de cobertura de acueducto y alcantarillado tienen valores positivos poco significativos puesto que un aumento en la cobertura de estas sugiere un crecimiento de 0:015 % y 0:045 % respectivamente. Las demás enfermedades tienen un riesgo negativo esto quiere decir que la presencia de estas disminuye el riesgo de contraer Malaria. Las variables ambientales siguen siendo de mayor aporte al riesgo de la enfermedad en donde la cobertura de bosque y precipitación aumentan el riesgo de Malaria en 9:36 % y 11:25 % respectivamente y la variabilidad de altitud disminuye el riesgo de Malaria en 3:28 %. La probabilidad del

parámetro cero que ajusta de acuerdo a las covariables es igual a $p = 0{:}715$, lo que quiere decir que la probabilidad de ceros muéstrales para el modelo ronda el 71:5 % lo que significa que la información a priori tuvo mayor peso puesto que los hiper parámetros se establecieron con un $p = 0{:}72$.

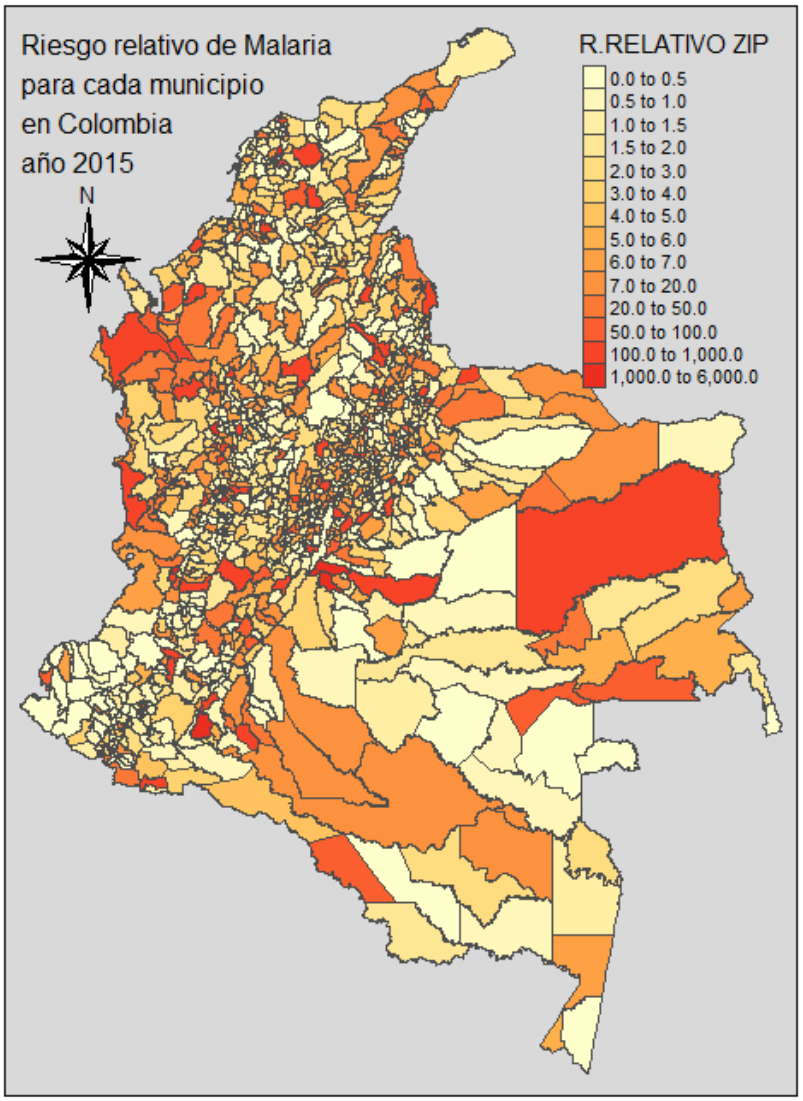


Figura 6.12: Riesgo relativo para malaria en Colombia año 2015 de acuerdo con el modelo ZIP con a priori No Informativa, software R

6.3. Selección del mejor modelo

En el análisis de regresión, a menudo se quiere encontrar un modelo reducido con el mejor subconjunto de las variables del modelo completo. La selección del modelo en el análisis frecuentista se basa comúnmente en el criterio de información de Akaike (AIC), un criterio basado en MLE. (Sakamoto, Ishiguro, y Kitagawa, 1986)

El modelo bayesiano estimado depende de las especificaciones de la estructura de densidad de muestreo y la distribución previa de los parámetros del modelo, y por lo tanto, los problemas cruciales con el modelo estadístico bayesiano son la evaluación del modelo. Se han propuesto muchos enfoques a lo largo de los años para abordar esta cuestión clave en el modelo estadístico bayesiano.(Ando, 2010)

Cuando el interés radica en la comparación entre diferentes modelos en términos de rendimiento, se puede utilizar su Deviance (generalización del análisis de la varianza para los GLM). Dados los datos y con probabilidad $P(yj\ )$ la desviación del modelo se define como:

$$D(\ ) = \quad 2log(P(yj\ ))$$

La desviación del modelo mide la variabilidad vinculada a la verosimilitud, que es la estructura probabilística utilizada para la observación (condicional a los parámetros). Esta cantidad es una variable aleatoria en el marco bayesiano, por lo que es posible sintetizarla a través de varios índices (media, mediana, etc.). Típicamente, la desviación media posterior $\overline{D} = E_{\ jy}(D(\ ))$ se usa como una medida de ajuste, ya que es muy robusta. (Blangiardo y Cameletti, 2015)

6.3.1. Criterio de información de desviación (DIC)

La medida más utilizada del ajuste del modelo basada en la desviación para los modelos bayesianos es el criterio de información de desviación (DIC), propuesto por Spiegelhalter et al. (2002), desarrollado especialmente para la comparación del modelo bayesiano y es la suma de dos componentes, uno para cuantificar el ajuste del modelo y el otro para evaluar la complejidad del modelo. El primer componente se mide a través de la expectativa posterior de la desviación $D(\ ) = \quad 2log(P(yj\ ))$ mientras que la complejidad del modelo se mide a través del número efectivo de parámetros:

$$P_D = E_{\ jy}(D(\ )) \quad D(E_{\ jy}(\ )) = \overline{D} \quad D(\bar{\ })$$

Entonces el DIC será:

$$DIC = \overline{D} + p_D \tag{75}$$

En el análisis bayesiano, DIC, una generalización de AIC, es una de las medidas más populares para la comparación de modelos bayesianos, que se define como la suma de una medida de bondad de ajuste más una medida de complejidad del modelo. El modelo con el DIC más bajo proporciona el mejor equilibrio entre ajuste y complejidad del modelo. Desafortunadamente, no hay una función disponible para la selección de modelo por pasos "Stepwise" por DIC en la biblioteca INLA. (Wang et al., 2018)

6.3.2. Criterio de información Watanabe-Akaike (WAIC)

El término de desviación utilizado en el cálculo de WAIC requiere el logaritmo de la densidad predictiva puntual (LPPD), que se calcula como:

$$LPPD = \sum log \int p(y_i|\ )P_{post}(\ )d$$

Como el cálculo de LPPD usa toda la distribución posterior $P_{post}(\ )$, LPPD puede verse como un análogo completamente bayesiano de $logp(y|\hat{\ }_{mle})$ en el cálculo de AIC y BIC y $logp(y|\hat{\ }_{EAP})$ en el cálculo de DIC. Similar a LPPD, el término de penalización de WAIC es completamente bayesiano y se puede expresar como:

$$P_{WAIC} = \sum var_{post}(logp(y_i|\ ))$$

donde el término de penalización es la varianza de términos individuales en el registro de densidad predictiva sumada sobre los n puntos de datos. WAIC se calcula como:

$$WAIC = 2LPPD + 2P_{WAIC} \tag{76}$$

Al igual que con DIC, se define WAIC como -2 veces la expresión (76) para estar en la escala de desviación. En la definición original de Watanabe, WAIC es el negativo de la densidad predictiva de registro puntual promedio (suponiendo la predicción de un único punto de datos nuevo) y, por lo tanto, se divide por n y no tiene el factor 2; aquí se escala para que sea comparable con AIC, DIC y otras medidas de desviación. (Gelman, Hwang, y Vehtari, 2014)

Como se puede apreciar en la tabla (6.21) el mejor modelo de acuerdo a los dos anteriores criterios es el ZIP con a priori no informativas el cual se puede observar

en la figura (6.12). Dado que en la muestra de Malaria para el año 2015 se observaba una sobredispersión con mas del 72 % sin reporte en los municipios, el ajuste de este modelo es más que satisfactorio dado que al igual que el modelo Binomial negativo son determinantes para el tratamiento de este tipo de datos.

Modelo	A priori	Tipo(I/NI)	Hiper parámetros	DIC	WAIC
Poisson	Gamma	Informativa	= 107:121, = 0:982	91724:64	134040:9
		No informativa	= 0:5, = 0	92396:9	134598:5
Binomial	Beta	Informativa	= 78:274, = 2030:508		
		No informativa	= 0:5, = 0:5	7009:356	9109:246
NB	PC.gamma	Informativa	= 0:3816794	14151:28	14475:66
		No informativa	= 0:1428571	16963	17162:13
ZIP	logit Beta	Informativa	= 110:468, = 62:092	68236:14	109227:4
		No informativa	= 0:5, = 0:5	2860:71	2912:744

Cuadro 6.21: Resumen de criterios para los modelos ajustados bajo la metodología INLA

Capítulo 7

Conclusiones

En el presente trabajo se modeló el comportamiento de la enfermedad MALARIA para el año 2015 implementando modelos como Poisson y binomial, como estos no dan respuesta exacta dada la naturaleza de los datos (muchos ceros dentro de la muestra), se optó por realizar un modelamiento más complejo con dos modelos como lo fueron el Binomial negativo y el Poisson con ceros inflados. Estos modelos son espacializados por medio de la especificación intrínseca condicional autorregresiva estimados por la aproximación empírica a la inferencia Bayesiana conocida como INLA.

Dado que INLA es una aproximación Bayesiana, tiene dos partes dentro del estudio que son de gran importancia, la primera la definición de las distribuciones a priori, las cuales tienen una gran impacto en la predictividad del modelo y ajuste del mismo. Aquí se encontró en varios modelos que sin importar las especificación a priori, los criterios de validación (DIC y WAIC) fueron muy cercanos tanto si fuese información informativa o no informativa (para el modelo Poisson y Binomial negativo).

Una limitación que se identificó fue en la validación para los modelos Bayesianos, lo que direcciono a elegir los modelos con mayor cantidad de variables significativas y no seguir la metodología frecuentista donde se establece el mejor modelo de acuerdo al ajuste de los datos, esto se puede dar dada la gran carga computacional del algoritmo.

La comparación por criterios de bondad de ajuste como DIC y WAIC muestra no ser totalmente efectiva puesto que en la comparativa entre los modelos Poisson y Binomial, en donde el modelo Binomial muestra valores más bajos para ambos criterios en comparación con el Poisson este no describe el fenómeno en ningún aspecto. Aunque el modelo Poisson tampoco lo explique tiene mayor similitud al calculo del SIR.

De acuerdo con los resultados que constan en el trabajo el modelo mas adecuado para modelar la malaria en Colombia para 2015 es aquel que sigue una distribución Poisson

para ceros inflados con una componente que describe la auto correlación espacial con especificación Besag. Este modelo muestra un ajuste adecuado para bases de datos con un altísimo porcentaje de observaciones cero, que presentan además sobredispersión la componente de conteo del modelo. La inclusión de estas observaciones en el proceso de modelamiento brinda información sobre la estructura de correlación espacial que, de ser omitidas, se perdería. Sin embargo, existen unos atípicos que se pueden justificar por la influencia de las covariables dado que las variables ambientales como precipitación y cobertura de bosque tienen la mayor influencia sobre el modelo. La región andina donde la variable altura media predomina, esta solo tiene una disminución de un 44% aproximadamente, mientras que la precipitación tiene un impacto de aumento en un 88% y una cobertura de bosque aunque mínima (menor al 40% en la región) tiene un impacto del 141% lo que genera un aumento exponencial en algunos municipios de la región.

En comparativa la región Amazonia teniendo valores altos en cobertura de bosque, y precipitación y estando a menos de 200 msnm tiene un riesgo relativo bajo en comparativa con el país. Se aprecia el cluster en la región del Pacifico y la región Orinoquía expresado en el SIR. La región Caribe tiene altas probabilidades, sin embargo predominan las demás enfermedades así que el riesgo en esta zona depende de la disminución de las demás enfermedades.

La posible desaparición de los clusters puede estar determinada por establecer la matriz de pesos que aumente la autocorrelación, así mismo, si se estableciera una matriz más sencilla esta podría determinar un resultado equiparable con el SIR.

Como las variables socio económicas, cobertura de acueducto y alcantarillado fueron de bajo impacto (menor al 1%) en cualquiera de los modelos, no se puede determinar si el aumentar estas coberturas pueda disminuir a ciencia cierta el riesgo de la enfermedad. Aquí se podría establecer que aumentar esta cobertura es mucho más costoso que el tratar un enfermo, entonces para trabajos posteriores se deben contemplar otro tipo de variables.

Los mapas y resultados se pueden visualizar en el siguiente aplicativo "Malaria Story Map"

Referencias

Ancot, L., Paelinck, J., Klaassen, L., Molle, W., Albegov, M., Andersson, A., y Snickars, F. (1982). Topics in regional development modelling. *M. Albegov, Å. Andersson and F. Snickars (eds, pp. 341-359), Regional Development Modelling in Theory and Practice. Amsterdam: North Holland*.

Ando, T. (2010). *Bayesian model selection and statistical modeling*. Chapman and Hall/CRC.

Anselin, L. (1998). *Spatial econometrics: methods and models* (Vol. 1). Springer Science.

Arbia, G. (2014). *A primer for spatial econometrics: with applications in R*. Springer.

Barndor□-Nielsen, O., y Cox, D. (1989). *Asymptotic techniques for use in statistics*. Taylor & Francis.

Bauman, D., Drouet, T., Fortin, M.-J., y Dray, S. (2018). Optimizing the choice of a spatial weighting matrix in eigenvector-based methods. *Ecology*, *99*(10), 2159–2166.

Bayes, T. (1763). Lii. An essay towards solving a problem in the doctrine of chances. By the late Rev. mr. Bayes, FRS communicated by mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*(53), 370–418.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., y Songini, M. (1995). Bayesian analysis of space—time variation in disease risk. *Statistics in medicine*, *14*(21-22), 2433–2443.

Besag, J., York, J., y Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, *43*(1), 1–20.

Blangiardo, M., y Cameletti, M. (2015). *Spatial and spatio-temporal bayesian models with r-inla*. John Wiley & Sons.

Box, G. E., y Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Wiley Interscience.

Correa Morales, J. C. (2013). *Introducción a la estadística bayesiana con R*. Universidad Nacional de Colombia sede Medellín.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, *4*(5), 613–617.

Dray, S., Legendre, P., y Peres-Neto, P. R. (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *ecological modelling*, *196*(3-4), 483–493.

Gelman, A., Hwang, J., y Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, *24*(6), 997–1016.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Holden, E. (1880). *Mortality and sanitary record of newark, nj from 1859 to 1879.* A report presented to the president and directors of the Mutual Benefit Life ....

Je□reys, H. (1998). *The theory of probability*. OUP Oxford.

Kandhasamy, C., y Ghosh, K. (2017). Relative risk for HIV in india–an estimate using conditional auto-regressive models with bayesian approach. *Spatial and spatio-temporal epidemiology*, *20*, 27–34.

Laplace, P. S. (1820). *Théorie analytique des probabilités*. Courcier.

Lawson, A. (2013). *Bayesian disease mapping: Hierarchical modeling in spatial epidemiology, second edition*. Taylor & Francis.

Lawson, A., Biggeri, A., Böhning, D., Lesa□re, E., Viel, J. F., y Bertollini, R. (1999). *Disease mapping and risk assessment for public health*. Wiley New York.

Lawson, A. B., Banerjee, S., Haining, R. P., y Ugarte, M. D. (2016). *Handbook of spatial epidemiology*. CRC Press.

Lawson, A. B., Williams, F. L., y Williams, F. (2001). *An introductory guide to disease mapping*. Wiley Online Library.

Lesa□re, E., y Lawson, A. B. (2012). *Bayesian biostatistics*. John Wiley & Sons.

Liu, H., y Powers, D. A. (2012). Bayesian inference for zero-inflated poisson regression models. *Journal of Statistics: Advances in Theory and Applications*, *7*(2), 155–188.

Mersad, M., Ganjali, M., y Rivaz, F. (2015). Some extensions of zero-inflated models and bayesian tests for them. *Journal of Statistical Computation and Simulation*, *85*(18), 3792–3810.

Moraga, P. (2019). *Geospatial health data: Modeling and visualization with r-inla and shiny*. Chapman and Hall/CRC.

Naeeim, N. S. A., y Rahman, N. A. (2017). Estimating relative risk for dengue disease in peninsular malaysia using inla. *Malaysian Journal of Fundamental and Applied Sciences*, *13*(4), 721–727.

Pfei□er, D., Robinson, T. P., Stevenson, M., Stevens, K. B., Rogers, D. J., y Clements, A. C. (2008). *Spatial analysis in epidemiology* (Vol. 142) (n.º 10.1093). Oxford

University Press Oxford.

Press, S. J. (1989). *Bayesian statistics: principles, models, and applications*. Wiley New York.

Ross, S. M. (2013). *Simulation*. Academic Press.

Rue, H., Martino, S., y Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, *71*(2), 319–392.

Ruiz, A. E., y Peña, E. G. (2007). *Monografía de estadística bayesiana*.

Sakamoto, Y., Ishiguro, M., y Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, *81*.

Samat, N. A., y Mey, L. W. (2017). Malaria disease mapping in malaysia based on Besag-York-Mollie (BYM) model. En *Journal of physics: Conference series* (Vol. 890, p. 012167).

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*(318), 626–633.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., y Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, *32*(1), 1–28.

Sloan, C. D., Nordsborg, R. B., Jacquez, G. M., Raaschou-Nielsen, O., y Meliker, J. R. (2015). Space-time analysis of testicular cancer clusters using residential histories: A case-control study in denmark. *PloS one*, *10*(3), e0120285.

Snow, J. (1857). Cholera, and the water supply in the south districts of london. *British Medical Journal*, *1*(42), 864.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., y Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, *64*(4), 583–639.

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

Wang, X., Ryan, Y. Y., y Faraway, J. J. (2018). *Bayesian regression modeling with INLA*. Chapman and Hall CRC.

Yrigoyen, C. C. (2003). *Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales*. Dirección General de Economía y Planificación.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., y Smith, G. M. (2009). *Mixed e ects models and extensions in ecology with R*. Springer Science & Business Media.

Zuur, A. F., Ieno, E. N., y Saveliev, A. A. (2017). *Beginner's guide to spatial, temporal, and spatial-temporal ecological data analysis with R-INLA*.