

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ  
ESCUELA DE POSGRADO



Modelos espacio-temporales bayesianos para estudiar la incidencia de  
dengue en el Perú

Tesis para optar por el grado académico de Maestra en Estadística  
que presenta:

**Katia Alejandra Caro Ferreyra**

Asesora:

**Dra. Zaida Jesús Quiroz Cornejo**

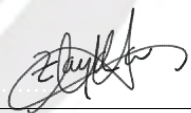
Lima, 2024

## Informe de Similitud

Yo Zaida Jesús Quiroz Cornejo docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesora de la tesis titulada *Modelos espacio-temporales bayesianos para estudiar la incidencia de dengue en el Perú*, de la autora Katia Alejandra Caro Ferreyra, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 15 %. Así lo consigna el reporte de similitud emitido por el software Turnitin el 07/08/2024.
- He revisado con detalle dicho reporte y confirmo que cada una de las coincidencias detectadas no constituyen plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lima, 7 de agosto de 2024

Apellidos y nombres de la asesora: Quiroz Cornejo Zaida Jesús	
DNI: 43704124	Firma: 
ORCID: <a href="https://orcid.org/0000-0003-3821-0815">https://orcid.org/0000-0003-3821-0815</a>	

# Dedicatoria

A Manuel, por ser el mejor compañero y mi apoyo en esta travesía. A mis padres y abuelos, por su invaluable inspiración. A mis amigos por estar siempre.



# Agradecimientos

Quiero agradecer a cada uno de los profesores de la Maestría, quienes motivaron mi deseo de seguir descubriendo. En especial, a la Dra. Zaida Quiroz, mi tutora, cuya orientación, apoyo y contención me impulsaron durante el desarrollo de la tesis. También, quiero agradecer a mi amigo Dante Baldeón, quien no solo compartió horas de estudio conmigo, sino además su gran conocimiento y apreciada amistad. Gracias a todos ellos, por lo que me enseñaron.



# Resumen

La prevención del dengue requiere un sistema para identificar las áreas con mayor riesgo, utilizando datos epidemiológicos con estructura espacial y temporal. Los enfoques bayesianos, que integran información previa y manejan estructuras jerárquicas, proporcionan un enfoque flexible y robusto, que permite estimaciones más precisas de la incertidumbre, además de captar la correlación espacial y espacio-temporal, registrando esta variabilidad en las estimaciones de riesgo de enfermedades. Estos enfoques jerárquicos bayesianos, a menudo requieren métodos numéricos sofisticados para proporcionar estimaciones de los parámetros. En este sentido, se pueden aplicar métodos como el Monte Carlo basado en cadenas de Markov (MCMC) o la Aproximación Anidada Integrada de Laplace (INLA), siendo ésta última una alternativa computacionalmente más eficiente para modelos gaussianos latentes (MGL), incluyendo modelos espaciales como el modelo jerárquico de Besag, York y Mollié (BYM), el cual puede extenderse a contextos espacio-temporales, que son de gran utilidad para evaluar el conteo de casos a lo largo del tiempo. En este marco, el presente trabajo evaluó tres modelos bayesianos, un modelo jerárquico de tendencia lineal paramétrica, un modelo jerárquico modelado dinámicamente usando un paseo aleatorio o *random walk* y un modelo de tendencia dinámica no paramétrica con interacción espacio-temporal. Para mostrar el aporte de esta propuesta, los tres modelos se ajustaron a datos reales que incluyeron tanto los casos de dengue como su incidencia. En el procedimiento de selección del modelo no solo se comparó la idoneidad de los modelos, sino también de distintas distribuciones de conteo añadiendo al análisis, covariables climáticas.

**Palabras-clave:** · INLA, MGL, dengue.

# Abstract

The prevention of dengue requires a system to identify areas at higher risk, using epidemiological data with spatial and temporal structure. Bayesian approaches, which integrate prior information and handle hierarchical structures, provide a flexible and robust method that allows for more accurate uncertainty estimates, as well as capturing spatial and spatiotemporal correlation, accounting for this variability in disease risk estimates. These hierarchical Bayesian approaches often require sophisticated numerical methods to provide parameter estimates. In this context, methods such as Markov Chain Monte Carlo (MCMC) or Integrated Nested Laplace Approximation (INLA) can be applied, the latter being a more computationally efficient alternative for latent Gaussian models (LGM), including spatial models such as the hierarchical Besag, York, and Mollié (BYM) model, which can be extended to spatiotemporal analyses, being very useful for evaluating the count of cases over time. In this framework, the present study evaluated three Bayesian models: a hierarchical model with a parametric linear trend, a hierarchical model dynamically modeled using a random walk, and a non-parametric dynamic trend model with spatiotemporal interaction. To demonstrate the contribution of this proposal, the three models were fitted to real data that included both dengue cases and their incidence. In the model selection procedure, not only was the suitability of the models compared, but also different count distributions were analyzed, adding climatic covariates to the analysis.

**Keywords:** INLA, LGM, dengue disease.

# Índice general

<b>Resumen</b>	<b>v</b>
<b>1. Introducción</b>	<b>1</b>
1.1. Consideraciones preliminares	1
1.1.1. Motivación del estudio	1
1.1.2. Descripción de datos	1
1.1.3. Justificación	2
1.1.4. Planteamiento del modelamiento estadístico	3
1.2. Objetivos	4
1.3. Organización del trabajo	4
<b>2. Conceptos y modelos</b>	<b>6</b>
2.1. Estadística espacial para datos de área	6
2.1.1. Matriz de vecindad o proximidad	7
2.1.2. Asociación espacial	7
2.1.3. Modelo condicional autorregresivo intrínseco (ICAR)	8
2.2. Modelos espaciales para datos de área de conteo	9
2.3. Modelos temporales	10
2.3.1. Paseo aleatorio de orden 1 (RW1)	11
2.4. Campo aleatorio gaussiano de Markov (GMRF)	11
2.5. Inferencia bayesiana	14
2.5.1. Introducción	14
2.5.2. Método Monte Carlo basado en Cadenas de Markov (MCMC)	15
2.5.3. Aproximación de Laplace	16
2.5.4. Modelo Gaussiano Latente (MGL)	17
2.5.5. Aproximación Anidada Integrada de Laplace (INLA)	19



<b>3. Modelos espacio-temporales</b>	<b>22</b>
3.1. Estructura espacial . . . . .	22
3.2. Estructura espacio-temporal . . . . .	24
3.2.1. Modelo 1: Modelo jerárquico de tendencia lineal paramétrica . . . . .	24
3.2.2. Modelo 2: Modelo jerárquico de tendencia dinámica no paramétrica a través de RW1 . . . . .	26
3.2.3. Modelo 3: Modelo jerárquico de tendencia dinámica no paramétrica e interacción espacio-temporal . . . . .	28
3.3. Inferencia bayesiana usando INLA . . . . .	30
3.4. Selección del modelo . . . . .	31
<b>4. Estudio de Simulación</b>	<b>32</b>
4.1. Modelo 1 . . . . .	33
4.2. Modelo 2 . . . . .	35
4.3. Modelo 3 . . . . .	37
<b>5. Aplicación</b>	<b>40</b>
5.1. Aplicación 1: Dengue por semanas epidemiológicas . . . . .	41
5.1.1. Análisis exploratorio de los datos . . . . .	41
5.1.2. Aplicación de modelos - casos de dengue . . . . .	44
5.1.3. Aplicación de modelo - incidencia de dengue semanal . . . . .	46
5.2. Aplicación 2: Dengue anual . . . . .	48
5.2.1. Análisis exploratorio de los datos . . . . .	48
5.2.2. Aplicación de modelos - casos de dengue anual . . . . .	50
5.2.3. Aplicación de modelos - incidencia de dengue anual . . . . .	51
5.2.4. Aplicación de Modelos con covariables climáticas - incidencia de dengue anual . . . . .	53
<b>6. Conclusiones</b>	<b>59</b>
<b>A. Análisis exploratorio de casos de dengue por semanas</b>	<b>61</b>
<b>B. Distribución de la variable respuesta</b>	<b>65</b>
<b>C. Resultados adicionales en la aplicación</b>	<b>66</b>
C.1. Aplicación 1: Casos de dengue por semana . . . . .	66
C.2. Aplicación 1: Incidencia de dengue por semana . . . . .	67



C.3. Aplicación 2: Casos de dengue por año, Periodo 2010 a 2023 . . . . .	70
C.4. Aplicación 2: Incidencia de dengue por año, Periodo 2010 a 2023 . . . . .	71
C.5. Aplicación 2: Incidencia de dengue por año considerando covariables climáticas, Periodo 2010 a 2023 . . . . .	72
<b>D. Modelos espacio-temporales adicionales</b>	<b>73</b>
D.1. Binomial Negativa (BN) . . . . .	73
D.2. Poisson Cero Inflacionada (ZIP) . . . . .	75
D.3. Binomial Negativa Cero Inflacionada (ZINB) . . . . .	76
<b>Bibliografía</b>	<b>79</b>



# Capítulo 1

## Introducción

### 1.1. Consideraciones preliminares

#### 1.1.1. Motivación del estudio

El dengue es una enfermedad viral transmitida por los vectores *Aedes aegypti* y *Aedes albopictus*, que en las últimas décadas ha incrementado su incidencia a nivel mundial (OMS, 2023). En ese contexto, en el 2023, la región de las Américas reportó 4.5 millones de casos de dengue (OMS, 2024), la mayor epidemia registrada hasta la fecha. En Perú, en los últimos años, se ha observado un incremento sostenido de los casos, los cuales, según los registros epidemiológicos del año 2022, han aumentado un 88.2 % en comparación con el año previo (CDC, 2022).

La presencia del vector juega un rol importante en la transmisión del dengue, informando la Dirección General de Salud Ambiental e Inocuidad Alimentaria (DIGESA, 2022), que el *Aedes aegypti* se ha dispersado en 22 departamentos, 94 provincias y 538 distritos a nivel nacional, diez distritos más que los infestados en julio de 2022.

#### 1.1.2. Descripción de datos

El virus del dengue es un arbovirus, perteneciente al género *Flavivirus* y la familia *Flaviviridae*; familia que agrupa virus de ácido ribonucleico (ARN), que se multiplican en células de vertebrados y de insectos vectores. El género *Flavivirus* reúne en su mayoría (55 %) a virus asociados con enfermedades humanas, entre ellos el dengue (Rice, 1996) que presenta cuatro serotipos (DENVs 1-4), con características antigénicas y serológicas diferentes, incluyendo variantes genéticas relacionadas con la virulencia y la procedencia geográfica dentro

de un mismo serotipo (Monath y Tsai, 1997). En el Perú, los primeros casos de dengue en forma epidémica fueron reportados en la Amazonía en 1990, aislándose el serotipo 1 (Mostorino et al., 2002). Desde ese año, el dengue se ha extendido en el país circulando los cuatro serotipos, más comúnmente el DENV-1 y el DENV-2 (Cabezas et al., 2015).

El dengue es endémico en algunas regiones del Perú, y es considerado una importante enfermedad reemergente (INS, 2018). Las condiciones ambientales, incluido el clima, cumplen un papel en la supervivencia, el comportamiento y la proliferación de *Aedes aegypti* y *Aedes albopictus* (Reinhold et al., 2018). El clima y la estacionalidad juegan un papel importante en la propagación del dengue (Ebi y Nealon, 2016; Morin et al., 2013), siendo conocido que la transmisión de esta enfermedad metaxénica está fuertemente correlacionada con las fluctuaciones de variables como la precipitación, la temperatura y la humedad relativa (Tsheten et al., 2020).

Además del clima, condiciones socioeconómicas como el crecimiento demográfico descontrolado, el hacinamiento, el deterioro de los sistemas de servicios de salud, la deficiente provisión del agua o la inadecuada disposición de residuos, contribuyen a la enfermedad (Stewart-Ibarra et al., 2014; Kouri et al., 2007). Asimismo, se tiene conocimiento, que favorecen la dispersión y transmisión del dengue, los cambios en el uso del suelo y la presencia de condiciones de habitabilidad no planificadas, en combinación con una mayor movilidad de las personas (Lana et al., 2017; Zellweger et al., 2017).

### 1.1.3. Justificación

La prevención del dengue, requiere de un sistema que permita la identificación de las regiones con mayor riesgo de la enfermedad. En ese sentido, un mapeo de la ocurrencia del dengue a una escala espacio-temporal fina es crucial para el sistema de alerta temprana e invaluable para los formuladores de las políticas de salud que requieren investigar sobre los factores de riesgo y planificar programas de control e intervención (Jaya y Folmer, 2020; Ugarte et al., 2014). Tal sistema requiere métodos estadísticos para generar pronósticos de riesgo espaciales y temporales (Jaya y Folmer, 2020; Ugarte et al., 2014).

Modelar y pronosticar el riesgo de una enfermedad es desafiante, especialmente en países como Perú, donde la información puede estar incompleta o imprecisa, debido a los recursos limitados para la vigilancia (McMichael et al., 2013). Este tipo de desafíos, dadas las características del dengue, cuya incidencia puede ser similar en regiones vecinas (i.e. auto-

correlación espacial) y donde la cantidad de casos en una región particular, en un tiempo  $t$ , puede depender de la incidencia del dengue en la misma región en tiempos previos; implican plantear el uso de modelos espacio-temporales que incluyan tanto efectos aleatorios espaciales y temporales, así como sus interacciones (Wakefield, 2004).

En dicho contexto, Jaya y Folmer (2020) presentaron un modelo bayesiano espacio-temporal de efectos aleatorios del riesgo del dengue en Indonesia estimado mediante la aproximación anidada integrada de Laplace. Así también, Lowe et al. (2011, 2013, 2014, 2016) emplearon modelos espacio-temporales bayesianos para los sistemas de alerta temprana de dengue en Brasil, Tailandia y Ecuador; sin embargo, esos estudios no exploraron los modelos que incluyen efectos de interacción.

#### 1.1.4. Planteamiento del modelamiento estadístico

Teniendo en cuenta los antecedentes citados, para la presente tesis se propone aplicar modelos espacio-temporales para datos de áreas, con el fin de estudiar la distribución espacio-temporal de la incidencia de dengue en el Perú.

En particular, los modelos espacio-temporales, permitirán en relación a la incidencia de dengue, evaluar tanto los patrones de autocorrelación espacial entre locaciones (e.g. provincias), como su evolución temporal. Estos modelos son especialmente adecuados para capturar la dependencia espacial de los casos de dengue entre áreas cercanas y para analizar la dinámica temporal de la enfermedad. Los modelos espacio-temporales a usar en la tesis pertenecen a los llamados modelos gaussianos latentes (MGL), caracterizados por presentar un campo latente de gran dimensión que admite independencia condicional, y un número reducido de hiperparámetros (Rue et al., 2009). Estos modelos son flexibles dado que pueden modelar variables de respuesta que pueden ser gaussianas o no gaussianas.

En ese sentido, debido a la variable respuesta de conteo, y a la inclusión de los efectos aleatorios espaciales, temporales o espacio-temporales, se puede usar inferencia bayesiana para estimar los parámetros de estos modelos jerárquicos.

Los parámetros de los modelos espacio-temporales bayesianos se suelen estimar mediante el método de Monte Carlo basado en Cadenas de Markov (MCMC, por sus siglas en inglés, *Markov Chain Monte Carlo*) (Blangiardo et al., 2013), este método puede consumir una cantidad sustancial de tiempo de cálculo y presentar dificultades para llegar a la convergencia

de las cadenas (Blangiardo et al., 2013; Ugarte et al., 2014; Arab, 2015). Por ello, una alternativa al método MCMC, es la Aproximación Anidada Integrada de Laplace (INLA, por sus siglas en inglés, *Integrated Nested Laplace Approximation*) (Rue et al., 2009; Blangiardo et al., 2013; Jaya y Folmer, 2020). La metodología INLA reduce el tiempo de cálculo y produce estimaciones de parámetros confiables que son equivalentes a las estimaciones de MCMC (Bivand et al., 2015). Además, los métodos bayesianos como INLA, permiten incorporar en el pronóstico del riesgo de enfermedades infecciosas, las tendencias temporales y los patrones estacionales, así como sus interacciones; características que deben tenerse en cuenta para la prevención de la enfermedad (Jaya y Folmer, 2020).

## 1.2. Objetivos

El objetivo general de la tesis es estudiar propiedades, así como estimar y aplicar al conjunto de datos de dengue modelos espacio-temporales desde el punto de vista de la estadística bayesiana usando INLA. En particular, se propone modelar la incidencia de dengue a nivel departamental a través de las semanas epidemiológicas del año 2023 y modelar la incidencia del dengue a nivel provincial a través de los años 2010 al 2023.

De manera específica, para alcanzar los objetivos del estudio se consideran los siguientes procedimientos:

- Estudiar propiedades de la estimación de modelos espacio-temporales desde la perspectiva bayesiana.
- Realizar estudios de simulación de los modelos espacio-temporales usando métodos de inferencia bayesiana considerando la simulación INLA.
- Aplicar los modelos espacio-temporales a un conjunto de datos reales de dengue a nivel departamental y provincial, generando mapas de riesgo de enfermedad en diferentes escenarios temporales.

## 1.3. Organización del trabajo

El capítulo 2 se centra en una revisión de conceptos, exponiéndose las características principales de los modelos gaussianos latentes, modelos espaciales y temporales. El capítulo

3 presenta los detalles del enfoque INLA y de los modelos espacio-temporales ajustados de tendencia lineal paramétrica, de tendencia dinámica no paramétrica y de tendencia dinámica no paramétrica con interacción. El capítulo 4 muestra la inferencia bayesiana para modelos espacio temporales utilizando INLA a través de simulaciones. El capítulo 5 trata sobre las aplicaciones de los modelos ajustados. Finalmente, algunas conclusiones se comentan en el capítulo 6.





## Capítulo 2

# Conceptos y modelos

En la literatura estadística, existe una amplia variedad de modelos diseñados para describir o modelar tendencias y dependencias espacio-temporales. En ese sentido, este capítulo presenta algunos conceptos importantes en el desarrollo de las técnicas y herramientas que permiten comprender mejor las dinámicas de los procesos, tanto en el espacio, como en el tiempo.

Los métodos bayesianos que tratan datos espaciales y espacio-temporales se impulsaron en las últimas décadas, con el desarrollo de métodos de simulación, como el método MCMC (Casella y George, 1992; Gilks et al., 1996). El método MCMC es ampliamente utilizado, sin embargo, su mayor limitación es el tiempo de cómputo requerido cuando se disponen de estructuras espaciales y espacio-temporales; por lo que, para superar este problema, surge en la inferencia bayesiana, la metodología INLA que también es revisada en el presente capítulo.

### 2.1. Estadística espacial para datos de área

Los datos de área o de red surgen cuando un dominio fijo se divide en un número finito de subregiones en las que se agregan los resultados (Moraga, 2020). El componente espacial adquiere especial relevancia en ciertos eventos, por ejemplo, en la propagación de enfermedades, ya que la proximidad juega un papel fundamental en el aumento de la probabilidad de contagio, dado que es normal que elementos próximos posean características similares y presenten relaciones debido a su proximidad. En general, se espera observar similares incidencias de la enfermedad en áreas cercanas, en comparación con las áreas que están más



alejadas. De esta forma, la ubicación espacial puede actuar como un sustituto de covariables no observadas que inducen el patrón espacial.

### 2.1.1. Matriz de vecindad o proximidad

Un concepto principal que debe desarrollarse para datos de área, es el de la matriz de vecindad  $\mathbf{W}$ , la cual permite reflejar relaciones de proximidad entre áreas en el espacio. Esta matriz está conformada por entradas del tipo  $w_{ij}$  que representan la asociación entre las unidades  $i$  y  $j$ . En los casos más sencillos, estas relaciones se indican mediante: 0 (si las áreas  $i$  y  $j$  no tienen límites en común) y 1 (si las áreas  $i$  y  $j$  tienen algún límite en común). Por tanto,  $w_{ij} \neq 0$  si y solo si las áreas  $i$  y  $j$  son vecinas; en tanto  $w_{ij} = 0$  en cualquier otro caso (Blangiardo y Cameletti, 2015).

La matriz  $\mathbf{W}$  se define como:

$$\mathbf{W} = \begin{bmatrix} w_{11} & \dots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \dots & w_{nn} \end{bmatrix},$$

donde  $n$  representa el número de áreas.

### 2.1.2. Asociación espacial

Existen dos estadísticas usadas para medir la fuerza de una asociación espacial o autocorrelación espacial: el Índice de Morán y la C de Geary (Ripley, 1981).

**Índice de Morán (I):** Este Índice fue establecido por Morán (1950) y muestra, si un patrón espacial está agrupado, disperso o es aleatorio. Según Anselin (2010), el índice establece el tipo, la intensidad y el rango del patrón espacial, siendo análogo al coeficiente de correlación entre dos variables.

El índice de Morán se expresa como:

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

donde  $y_i$  representa la variable observada en la  $i$ -ésima unidad espacial;  $w_{ij}$  es un elemento

de la matriz  $\mathbf{W}$  y representa la asociación entre las unidades  $i$  y  $j$ ; en tanto,  $\bar{y}$  es el promedio de la variable analizada en todas las áreas espaciales, calculada como  $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$  para  $i = 1, 2, \dots, n$  áreas.

Los valores del Índice de Morán se encuentran en el intervalo  $[-1, 1]$ , donde:  $I = -1$  significa dispersión perfecta, es decir, que no se encuentran patrones espaciales claros;  $I = 0$  significa un patrón espacial aleatorio e  $I = 1$  significa autocorrelación espacial perfecta, lo que representa, una fuerte asociación espacial.

**C de Geary (C):** Este índice fue propuesto por Geary (1954), se basa en la disimilaridad y es una de las medidas de autocorrelación espacial más utilizadas. Los valores del Índice  $C$  se encuentran en el intervalo  $[0, 2]$ , donde  $0 < C \leq 1$  indica regiones similares (autocorrelación positiva).

El índice  $C$  se expresa como:

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

donde  $n$  es el total de unidades espaciales, y el resto de las variables se definen de forma similar a las descritas para el índice de Morán.

### 2.1.3. Modelo condicional autorregresivo intrínseco (ICAR)

Los modelos ICAR (por sus siglas en inglés, *Intrinsic conditional autoregressive*) son modelos utilizados en estadística espacial. Los modelos ICAR se consideran intrínsecos porque no tienen una varianza marginal finita por sí solos y generalmente son utilizados como distribuciones a priori en un campo aleatorio. Shaddick et al. (2024) indican que un enfoque común es asignar a los efectos aleatorios espaciales una distribución a priori ICAR; en esta especificación, se asume la siguiente distribución condicional:

$$y_i | y_j, \sim \mathcal{N} \left( \bar{y}_i, \frac{\tau^2}{m_i} \right), \quad j \in \delta_i,$$

donde  $\delta_i$  es el conjunto de vecinos del área  $i$ ;  $m_i$  es el número de vecinos;  $\bar{y}_i$  es la media de los efectos aleatorios espaciales de estos vecinos; y  $\tau^2$  es una varianza condicional (i.e. variabilidad condicionada a los valores vecinos) cuya magnitud determina la cantidad de variación espacial. Si  $\tau^2$  es "pequeño", entonces, aunque el residuo dependa fuertemente

del valor vecino, la contribución general al riesgo relativo residual es pequeña. En estos modelos, se asume que la condición de un área en particular está influenciada por sus vecinos directos. Este tipo de modelo es apropiado para situaciones con dependencia de primer orden o autocorrelación espacial relativamente local.

## 2.2. Modelos espaciales para datos de área de conteo

Según Wakefield (2004), cuando se desea estudiar tasas, en situaciones donde las probabilidades son pequeñas, el análisis debería centrarse en los modelos de Poisson. Esto permite evitar la llamada *falacia ecológica*, que tiende a exagerar ciertas correlaciones obtenidas a través de un ajuste agregado, correlaciones que no se obtendrían si los datos permitieran el ajuste de modelos basados en casos individuales de riesgo. Waller y Carlin (2010) señalan que el modelo de Poisson, en su forma más básica con efectos aleatorios espaciales, se define tal como se presenta a continuación:

$$y_i \sim \text{Poisson}(\lambda_i = E_i\theta_i),$$

donde  $y_i$  es el número de eventos de la enfermedad en el área  $A_i$ , para  $i = 1, \dots, n$ ,  $E_i$  es el número esperado de personas en riesgo en el área  $A_i$  y  $\theta_i$  es la tasa de incidencia. Waller y Carlin (2010) indican que estos datos de recuentos pueden modelarse como variables aleatorias con distribución Poisson, utilizando una función de enlace logarítmica sobre la tasa de incidencia, es decir  $\log(\theta_i)$ .

La principal característica del modelo Poisson es que asume que la media y la varianza de su distribución son iguales. Esta propiedad no se ajusta frecuentemente a la realidad, siendo una de las razones principales de introducción de sobredispersión (varianza > media) la heterogeneidad no observada, la cual puede ser debida a las características de los individuos.

El método más común de lidiar con la sobredispersión de los datos es utilizar la distribución Binomial Negativa (BN), la cual fue inicialmente estudiada por Anscombe (1949). Esta distribución se denota como:

$$y_i \sim \text{NegBin}(\lambda_i = E_i\theta_i, \Gamma),$$

donde se definen  $y_i$ ,  $E_i$  y  $\theta_i$  como fueran establecidos para el caso de la distribución

de Poisson. La distribución BN tiene una media  $E(y_i) = E_i\theta_i$  y una varianza  $\text{Var}(y_i) = E_i\theta_i + (E_i\theta_i)^2/\Gamma$  siendo  $\Gamma$  el parámetro de sobredispersión, es decir de variación adicional.

Aunque similar a la distribución de Poisson, la distribución BN se distingue por describir el número de éxitos antes de un fracaso y puede considerarse como una mezcla entre las distribuciones Gamma y Poisson, donde los datos observados siguen una distribución Poisson, y se presume la existencia de una heterogeneidad de los individuos (variabilidad) que sigue una distribución Gamma y que tiene influencia sobre la variable respuesta (Hilbe, 2007).

Una mayor incidencia de recuentos cero también puede causar sobredispersión. La inflación de ceros en los datos de conteo puede enfocarse mediante un modelo de ceros Inflacionados, el cual fue propuesto por Lambert (1992). En este modelo, se asumen dos clases de ceros: ceros aleatorios (i.e. ceros muestrales, que no representan una verdadera ausencia de la enfermedad) y ceros estructurales o verdaderos (i.e. ceros provenientes de aquellos que no tienen el atributo, como las personas sin la enfermedad). El modelo con distribución Poisson Cero Inflacionada (ZIP, por sus siglas en inglés, *Zero-Inflated Poisson*) combina dos partes: un modelo de conteo Poisson y un modelo de regresión Logística que contribuye al recuento de ceros en exceso. Mientras, el modelo con distribución Binomial Negativa (ZINB, por sus siglas en inglés, *Zero-Inflated Negative Binomial*) combina un modelo para el exceso de ceros que presenta una función de enlace logit y un modelo para datos discretos que siguen la distribución Binomial Negativa (modelo de conteo).

### 2.3. Modelos temporales

Incluir el tiempo en el análisis de un proceso espacial puede ayudar a entender la evolución del proceso. Si bien la dimensión espacial puede indicar qué áreas son las más riesgosas en términos de una enfermedad, es probable que este riesgo pueda disminuir o aumentar debido a las condiciones ambientales que por naturaleza cambian en el tiempo. En términos de control de enfermedades, la dimensión temporal puede ayudar a predecir posibles brotes a futuro.

Al igual que en los modelos espaciales, existen diferentes procesos temporales. En el presente estudio, el interés se basa en procesos de tiempo discreto, donde el proceso aleatorio  $y_j$  está indexado a tiempos fijos (e.g. semanas epidemiológicas o datos anuales). En estos procesos, se espera que dos variables  $y_j$  e  $y_{j+1}$  de dos tiempos consecutivos, sean más similares que aquellas separadas en el tiempo como  $y_j$  e  $y_{j+k}$ .

### 2.3.1. Paseo aleatorio de orden 1 (RW1)

Los paseos aleatorios (RW, por sus siglas en inglés, *Random Walk*) describen una curva en el tiempo o en el espacio. En particular, describimos el RW de orden 1 (RW1) en el tiempo. Se tiene que para los efectos aleatorios  $x_t$  indexados en el tiempo  $t$ , se asumen incrementos  $\Delta x_t = x_t - x_{t-1} \sim \mathcal{N}(0, \tau_1^{-1})$ , independientes, para  $t = 1, \dots, T-1$ . A partir de esta definición, se puede probar que

$$\pi(\mathbf{x}) \propto \tau_1^{T-1/2} \exp\left(-\frac{1}{2} \mathbf{x}^\top \mathbf{Q}_{RW1} \mathbf{x}\right),$$

donde  $\mathbf{Q} = \tau_1 \mathbf{R}$ ,

$$\mathbf{R} = \begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{pmatrix},$$

fuera de la banda diagonal, la matriz  $\mathbf{R}$  está llena de ceros.

En el RW1,  $x_t$  solo depende condicionalmente de los vecinos de primer orden y es independiente de todas las demás observaciones. Por lo tanto, los paseos aleatorios tienen la propiedad de Markov, lo que da lugar a una matriz de precisión  $\mathbf{Q}$  dispersa, lo que permite cálculos bayesianos rápidos.

## 2.4. Campo aleatorio gaussiano de Markov (GMRF)

El GMRF (por sus siglas en inglés, *Gaussian Markov Random Field*) es un campo aleatorio gaussiano cuyas variables tienen independencia condicional, es decir cumplen con la propiedad de Markov o Markoviana (Ross, 2014).

Los GMRFs se utilizan comúnmente en modelos jerárquicos (i.e. modelos estadísticos que estructuran datos en múltiples niveles de organización), donde se necesita modelar la dependencia entre los parámetros de manera estocástica, temporal, espacial, o espacio-temporal. Estos modelos permiten capturar la estructura jerárquica de los datos y la variabilidad entre



diferentes niveles de agrupación.

Existe gran variedad de GMRFs que han sido utilizados en diversos campos (Rue y Held, 2005). Estos GMRFs se caracterizan por tener una matriz de precisión dispersa, llena de ceros y una dimensión pequeña de parámetros ( $\dim(\theta) \leq 6$ ), lo que proporciona un beneficio computacional al realizar la inferencia bayesiana (Wang et al., 2018; Rue et al., 2009). Además de tener diferentes estructuras de independencia condicional que reflejan cómo las variables aleatorias para cada campo latente dependen localmente entre sí.

Para definir formalmente los GMRF, se establecen inicialmente dos conceptos importantes: los grafos no direccionados y la distribución gaussiana multivariada:

- Un grafo no direccionado  $G = (V, E)$  es una dupla donde  $V$  es el conjunto de nodos (o elementos) y  $E$  es el conjunto de aristas entre los nodos. Un grafo está completamente conectado si la arista entre los nodos  $\{i, j\} \subset E$  para todo  $\{i, j\} \in V$  con  $i \neq j$ .
- Un vector aleatorio  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^\top$  tiene una distribución gaussiana multivariada con media  $\boldsymbol{\mu}$  y matriz de covarianza  $\boldsymbol{\Sigma}$ . Luego,  $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  y tiene función de densidad conjunta dada por:

$$\pi(\boldsymbol{\eta}) = \frac{1}{(2\pi)^{n/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp \left( -\frac{1}{2} (\boldsymbol{\eta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\eta} - \boldsymbol{\mu}) \right), \quad \boldsymbol{\eta} \in \mathbb{R}^n.$$

De lo expuesto, podemos definir formalmente que un vector aleatorio  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^\top \in \mathbb{R}^n$  es un GMRF con respecto al grafo  $G=(V, E)$  con media  $\boldsymbol{\mu}$  y matriz de precisión  $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ , si y solo si, su densidad tiene la forma:

$$\pi(\boldsymbol{\eta}) = \frac{1}{(2\pi)^{n/2}} |\boldsymbol{Q}|^{1/2} \exp \left( -\frac{1}{2} (\boldsymbol{\eta} - \boldsymbol{\mu})^\top \boldsymbol{Q} (\boldsymbol{\eta} - \boldsymbol{\mu}) \right)$$

$$Q_{ij} \neq 0 \Leftrightarrow \{i, j\} \in E \quad \forall i \neq j.$$

Por lo tanto, un GMRF es una distribución gaussiana multivariada con propiedades de independencia condicional en sus variables y una matriz  $\boldsymbol{Q}$  dispersa. La razón por la que  $\boldsymbol{Q}$  tiene una estructura dispersa es que  $\eta_i$  solo depende de la variable precedente  $\eta_{i-1}$ , por lo que  $\eta_i$  y  $\eta_j$  son condicionalmente independientes para todo  $|i - j| > 1$ . Por ejemplo, para el vector  $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3, \eta_4)^\top$ ,  $\eta_2$  y  $\eta_4$  son condicionalmente independientes si:

$$\pi(\eta_2, \eta_4 \mid \eta_1, \eta_3) = \pi(\eta_2 \mid \eta_1) \pi(\eta_4 \mid \eta_1, \eta_2, \eta_3),$$

y esta fdp puede reescribirse como:

$$\pi(\eta_2, \eta_4 \mid \eta_1, \eta_3) = \pi(\eta_2 \mid \eta_1) \pi(\eta_4 \mid \eta_3).$$

Como ejemplo de un GMRF, se puede considerar un proceso autorregresivo de primer orden (AR1) con error gaussiano.

### Proceso Autorregresivo de Primer Orden (AR1)

Un proceso autorregresivo de primer orden (AR1) con error gaussiano, en el contexto del modelado de regresión bayesiana, se utiliza como un componente en la modelización de datos longitudinales o series temporales, y se define matemáticamente como:

$$\eta_t = \phi \eta_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, 1), \quad |\phi| < 1,$$

donde  $\eta_t$  es el valor actual en el tiempo  $t$  de la serie temporal y corresponde al parámetro de autorregresión, que representa la influencia del valor anterior  $\eta_{t-1}$  en el valor actual, y  $\epsilon_t$  es un término de error aleatorio en el tiempo  $t$ . Asumiendo que el valor actual se basa en el inmediatamente precedente, el modelo AR1 puede ser definido como:

$$\begin{cases} \eta_1 \sim N\left(0, \frac{1}{1-\phi^2}\right), \\ \eta_t \mid \eta_{t-1}, \dots, \eta_1 \sim N\left(\phi \eta_{t-1}, \sigma_\eta^2\right), \quad t = 2, \dots, n. \end{cases}$$

Dadas otras variables, cada  $\eta_t$  sigue una distribución normal y  $\phi$  es la correlación entre  $\eta_t$  y  $\eta_{t-1}$ . La distribución de  $\eta_1$  tiene en cuenta la varianza acumulada desde el inicio. Mientras que  $\eta_t$  dado los valores anteriores tiene una distribución donde cada término depende del previo.

La función de densidad conjunta de  $\eta$  se representa por:

$$\pi(\eta) = \pi(\eta_1) \pi(\eta_2 \mid \eta_1) \dots \pi(\eta_n \mid \eta_{n-1}) = \frac{1}{(2\pi)^{n/2}} |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2} \boldsymbol{\eta}^\top \mathbf{Q} \boldsymbol{\eta}\right),$$



donde la matriz de precisión  $\mathbf{Q}$  tiene la forma:

$$\mathbf{Q} = \begin{pmatrix} 1 & -\phi & 0 & \cdots & 0 \\ -\phi & 1 + \phi^2 & -\phi & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & -\phi & 1 + \phi^2 & -\phi \\ 0 & 0 & \cdots & -\phi & 1 \end{pmatrix}.$$

En un proceso AR1, los elementos fuera de la diagonal principal de la matriz tienen un valor de cero cuando las variables  $\eta_t$  y  $\eta_{t-1}$  tienen una distancia mayor que 1, es decir están condicionalmente no correlacionadas.

## 2.5. Inferencia bayesiana

En el enfoque bayesiano, las estimaciones, predicciones e inferencias se fundamentan en las distribuciones posteriores. El teorema de Bayes proporciona el marco matemático para calcular estas distribuciones posteriores. Según este teorema, la distribución posterior, que representa la probabilidad de un parámetro dado un conjunto de datos, se obtiene multiplicando la función de verosimilitud (i.e. la probabilidad de observar los datos dados los parámetros) por la distribución de probabilidad a priori para esos parámetros, y luego dividiendo el resultado por la probabilidad de observar los datos (Austin et al., 2002).

La regresión bayesiana es efectiva para describir datos epidemiológicos que presentan una estructura espacial y espacio-temporal (Dunson, 2001; Blangiardo et al., 2013). Constituyendo aspectos importantes de este enfoque: la inclusión de información previa a través de las distribuciones a priori; así como, la fácil especificación de una estructura jerárquica en los datos o parámetros, lo que beneficia la predicción de nuevas observaciones o la imputación de datos faltantes (Blangiardo y Cameletti, 2015).

### 2.5.1. Introducción

Para la inferencia frecuentista o clásica, los parámetros  $\boldsymbol{\theta}$  son fijos y desconocidos, siendo la forma más común de estimarlos, calcular los valores más probables al maximizar la función de verosimilitud  $L(\boldsymbol{\theta})$  mediante el método de estimación de máxima verosimilitud (MLE, por sus siglas en inglés, *Maximum Likelihood Estimate*). Mientras que, en la inferencia bayesiana,

los parámetros  $\theta$  son variables aleatorias asociadas a una distribución de probabilidad a la que se le denomina distribución a priori  $\pi(\theta)$ . El producto entre la información a priori y la verosimilitud  $\pi(\mathbf{y} | \theta)$  aportada por los datos, permitirá obtener la distribución a posteriori  $\pi(\theta | \mathbf{y})$ .

De esta forma, el teorema de Bayes permite obtener la distribución a posteriori de los parámetros  $\theta$  condicionada al conjunto de datos observados  $\mathbf{y}$ , de acuerdo a la siguiente ecuación:

$$\pi(\theta | \mathbf{y}) = \frac{\pi(\mathbf{y} | \theta)\pi(\theta)}{\pi(\mathbf{y})} \propto \pi(\mathbf{y} | \theta)\pi(\theta). \quad (2.1)$$

La forma de la distribución a posteriori puede ser determinada de manera exacta solo en ciertos modelos, mientras que en la mayoría de los casos se requiere una estimación aproximada. Algunos modelos tienen la ventaja de contar con distribuciones a priori conjugadas, donde la forma de la distribución a priori coincide con la de la distribución a posteriori, lo que facilita significativamente la estimación. Sin embargo, en situaciones donde la distribución a posteriori no proviene de una distribución conocida, se hace necesario emplear métodos alternativos para su estimación o para obtener muestras de la misma. En tales casos, los métodos computacionales MCMC o INLA juegan un papel importante, ya que se encargan de calcular las integrales necesarias para la inferencia bayesiana.

### 2.5.2. Método Monte Carlo basado en Cadenas de Markov (MCMC)

A menudo, las densidades posteriores pueden ser difíciles o imposibles de integrar explícitamente, especialmente cuando tenemos un modelo complejo con muchos parámetros. Un enfoque alternativo, que evita los problemas de integración, es utilizar técnicas de simulación. En esta situación, se extraen numerosas muestras de  $\pi(\theta | \mathbf{y})$ , que se pueden utilizar para estimar cantidades de interés como la media posterior.

El Método MCMC proporciona un enfoque para obtener muestras aleatorias de la distribución posterior, y se basa en la premisa de que es posible construir cadenas de Markov de los parámetros:  $\theta_1, \theta_2, \theta_3, \dots, \theta_n$  cuya distribución estacionaria es la posterior conjunta  $\pi(\theta | \mathbf{y})$  de interés (Brooks et al., 2011). Dos métodos muy conocidos para obtener muestras de la distribución a posteriori son: i) el algoritmo de Metropolis-Hastings (Metropolis et al., 1953; Hasting, 1970) y ii) el muestreador de Gibbs (Smith y Roberts, 1993).

El algoritmo de muestreo de Gibbs, funciona si las distribuciones condicionales completas

son conocidas y establece que, para cualquier valor inicial, la cadena de Markov converge a la distribución a posteriori. Cuando no se pueden tomar muestras de todas las distribuciones condicionales completas, se puede utilizar el algoritmo Metropolis-Hasting.

### 2.5.3. Aproximación de Laplace

Una de las mayores dificultades de la estadística bayesiana es evaluar la integral en el denominador de la ecuación (2.1), también conocida como constante de normalización, definida específicamente en la siguiente ecuación:

$$\pi(\theta | y) = \frac{\pi(y | \theta)\pi(\theta)}{\pi(y)} = \frac{\pi(y | \theta)\pi(\theta)}{\int \pi(y | \theta)\pi(\theta) d\theta}. \quad (2.2)$$

Por simplicidad, si asumimos que  $\theta$  es de dimensión 1, la notación de la integral de la ecuación (2.2) puede simplificarse como:

$$\int_a^b g(\theta) d\theta.$$

Si  $h(\theta) = \log(g(\theta))$  se puede reescribir esta integral como:

$$\int_a^b g(\theta) d\theta = \int_a^b \exp(h(\theta)) d\theta.$$

Expandiendo  $h(\theta)$  usando la serie de Taylor de segundo orden alrededor de  $\theta_0$ , obtenemos:

$$\int_a^b \exp(h(\theta)) d\theta \approx \int_a^b \exp \left[ h(\theta_0) + h'(\theta_0)(\theta - \theta_0) + \frac{1}{2}h''(\theta_0)(\theta - \theta_0)^2 \right] d\theta.$$

Luego, al asumir un punto  $\theta_0$  donde la función  $g$  alcanza su máximo, es decir si  $\theta_0$  es la moda, entonces  $h'(\theta_0) = 0$  y se tiene que:

$$\begin{aligned} \int_a^b \exp(h(\theta)) d\theta &= \int_a^b \exp \left[ h(\theta_0) + \frac{1}{2}h''(\theta_0)(\theta - \theta_0)^2 \right] d\theta \\ &= \exp[h(\theta_0)] \int_a^b \exp \left[ \frac{1}{2}h''(\theta_0)(\theta - \theta_0)^2 \right] d\theta \\ &= \exp[h(\theta_0)] \int_a^b \exp \left[ -\frac{1}{2} \frac{(\theta - \theta_0)^2}{(-h''(\theta_0)^{-1})} \right] d\theta, \end{aligned}$$

integral que se conoce como la aproximación gaussiana, pues se reconoce que la función dentro de la integral es proporcional a una distribución a posteriori normal, con media  $\theta_0$  y

con una varianza  $-h''(\theta_0)^{-1}$ , por lo tanto,

$$g(\theta) \sim \mathcal{N}(\theta_0, -h''(\theta_0)^{-1}).$$

Luego, se puede aproximar la integral en términos de la función de distribución acumulada, obteniendo que:

$$\int_a^b g(\theta) d\theta = \exp[h(\theta_0)] \sqrt{\frac{2\pi}{-h''(\theta_0)}} [\Phi(b | \theta_0, -h''(\theta_0)^{-1}) - \Phi(a | \theta_0, -h''(\theta_0)^{-1})].$$

Si  $a = -\infty$  y  $b = \infty$ , el término  $\Phi(b | \theta_0, -h''(\theta_0)^{-1}) - \Phi(a | \theta_0, -h''(\theta_0)^{-1})$ , es igual a 1, por lo que se obtiene:

$$\int_{-\infty}^{\infty} g(\theta) d\theta = \exp[h(\theta_0)] \sqrt{\frac{2\pi}{-h''(\theta_0)}}.$$

La aproximación Laplace tiende a ser efectiva para distribuciones a posteriori con densidades suaves que están fuertemente concentradas alrededor de la moda. Esta aproximación tiende a ser eficiente cuando se puede computar la moda a posteriori y puede formarse eficientemente la Hessiana de la densidad log-posterior.

#### 2.5.4. Modelo Gaussiano Latente (MGL)

El modelo gaussiano latente (MGL) se utiliza para describir estructuras latentes (i.e. no observadas). Este modelo asume que hay una o más variables latentes que siguen una distribución gaussiana y que estas variables latentes influyen en las variables observadas. Los MGLs van desde modelos lineales generalizados hasta modelos lineales generalizados mixtos como los modelos espaciales y espacio-temporales.

Los MGLs son una clase de modelos bayesianos jerárquicos donde la variable respuesta pertenece a una familia unimodal, por ejemplo, la familia exponencial, y esta variable respuesta es condicionalmente independiente dado un campo latente (normalmente distribuido) y algunos hiperparámetros. Su formulación genérica puede escribirse en tres niveles:

- Un primer nivel formado por la función de verosimilitud, donde cada observación  $y_i$  está conectada con un elemento del campo latente  $\eta_i$ . En este nivel el conjunto de observaciones  $\mathbf{y} = (y_1, \dots, y_n)^\top$  presenta cierta distribución exponencial y son condicionalmente

independientes:

$$\pi(y \mid \boldsymbol{\eta}, \theta_1) = \prod_{i=1}^n \pi(y_i \mid \eta_i, \theta_1).$$

- Un segundo nivel donde la distribución de las variables del campo latente  $\boldsymbol{\eta}$ , condicionada a los hiperparámetros  $\theta_2$ , se asume como gaussiana multivariada:

$$\pi(\boldsymbol{\eta} \mid \theta_2) \propto |\mathbf{Q}(\theta_2)|^{1/2} \exp \left( -\frac{1}{2} \boldsymbol{\eta}^\top \mathbf{Q}(\theta_2) \boldsymbol{\eta} \right),$$

donde el componente  $\mathbf{Q}(\theta_2)$  se denomina matriz de precisión y describe la estructura de dependencia subyacente de los datos,  $\mathbf{Q}(\theta_2)$  se corresponde con una matriz definida semi-positiva que en la ecuación depende del hiperparámetro  $\theta_2$  y cuya inversa es la matriz de covarianza. La matriz de precisión se encuentra compuesta por una gran cantidad de valores cero, lo que genera una mayor eficiencia computacional.

- Un último nivel de la estructura jerárquica o tercer nivel, que designa la a priori  $\pi(\boldsymbol{\theta})$  para los hiperparámetros  $\boldsymbol{\theta} = [\theta_1, \theta_2]$ .

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}).$$

Combinando estos tres niveles, la distribución a posteriori se puede expresar como:

$$\begin{aligned} \pi(\boldsymbol{\eta}, \boldsymbol{\theta} \mid \mathbf{y}) &\propto \pi(\boldsymbol{\theta}) \pi(\boldsymbol{\eta} \mid \theta_2) \prod_{i=1}^n \pi(y_i \mid \eta_i, \theta_1) \\ &\propto \pi(\boldsymbol{\theta}) |\mathbf{Q}(\theta_2)|^{1/2} \exp \left( -\frac{1}{2} \boldsymbol{\eta}^\top \mathbf{Q}(\theta_2) \boldsymbol{\eta} \right) + \sum_{i=1}^n \log \pi(y_i \mid \eta_i, \theta_1). \end{aligned} \quad (2.3)$$

La ecuación (2.3) es utilizada para el cálculo de la distribución a posteriori  $\pi(\boldsymbol{\eta}, \boldsymbol{\theta} \mid \mathbf{y})$ , mediante el método MCMC o alternativamente mediante INLA. En ese sentido, debe tenerse en cuenta que no todos los MGLs, pueden ser ajustados eficientemente por el INLA, requiriéndose de acuerdo a Wang et al. (2018) cumplir las siguientes suposiciones para que el modelo trabaje eficientemente:

- El número de hiperparámetros  $\boldsymbol{\theta}$  debe ser pequeño, según Rue et al. (2009), menor o igual a 15, típicamente entre 2 a 5.
- Si  $n$  es grande ( $10^3$  o  $10^5$ ),  $\boldsymbol{\eta}$  debe ser un GMRF.
- Cada  $y_i$  solo depende de un componente de  $\boldsymbol{\eta}$ , por ejemplo  $\eta_i$ .



### 2.5.5. Aproximación Anidada Integrada de Laplace (INLA)

El enfoque INLA presenta tres componentes clave: el marco de trabajo que se restringe a los modelos gaussianos latentes, un campo aleatorio gaussiano de Markov y la aproximación anidada integrada de Laplace. Si se tiene un modelo que puede expresarse en términos de un MGL, entonces se pueden aprovechar los métodos basados en aproximaciones de Laplace para realizar la inferencia bayesiana. La computación requerida para realizar la inferencia, estará determinada en gran medida por las características de la matriz de covarianza, que a menudo es densa, es decir, tiene muchas entradas que no son cero, lo que conlleva una carga computacional alta al realizar las inversiones de matriz requeridas.

En ese sentido, si se tiene un conjunto de hiperparámetros  $\boldsymbol{\theta}$  y  $\boldsymbol{\eta} \mid \boldsymbol{\theta}$  puede ser expresado en términos de un GMRF, es posible aprovechar métodos para reducir el esfuerzo computacional durante el análisis bayesiano (Rue y Held, 2005). El uso de un GMRF significa que típicamente la inversa de la matriz de covarianza, o sea la matriz de precisión ( $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ ), será dispersa (i.e. tendrá más entradas cero) debido a la independencia condicional (Rue y Held, 2005).

Los MGLs presentan un GMRF de alta dimensión y un vector de parámetros ( $\boldsymbol{\theta}$ ) de baja dimensión. La alta dimensionalidad en el GMRF y la fuerte dependencia entre los componentes del campo latente  $\boldsymbol{\eta}$ , y entre los componentes del campo latente  $\boldsymbol{\eta}$  y  $\boldsymbol{\theta}$ , crean problemas en la convergencia del algoritmo MCMC. Para superar este problema, se ha desarrollado INLA, siendo su principal ventaja la computacional.

INLA consta de tres actividades propuestas por Rue y Martino (2007). La primera actividad consiste en proponer una aproximación  $\tilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ , a la conjunta a posteriori de  $\pi(\boldsymbol{\theta} \mid \mathbf{y})$  usando la aproximación de Laplace. La segunda actividad implica proponer una aproximación  $\tilde{\pi}(\eta_i \mid \boldsymbol{\theta}, \mathbf{y})$  a las marginales de la distribución condicional de  $\eta_i$ , considerando la data  $\mathbf{y}$ , así como los hiperparámetros  $\boldsymbol{\theta}$ . La tercera actividad combina las dos actividades anteriores utilizando integración numérica.

El detalle de estas actividades, tomando como base las referencias bibliográficas Rue y Martino (2007); Rue et al. (2009) y Wang et al. (2018), se indica a continuación:

- **Primera actividad:** La aproximación de Laplace de la función de densidad (fdp) conjunta a posteriori de los hiperparámetros  $\boldsymbol{\theta}$  puede escribirse como:

$$\tilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \frac{\pi(\boldsymbol{\eta}, \boldsymbol{\theta} \mid \mathbf{y})}{\tilde{\pi}(\boldsymbol{\eta} \mid \boldsymbol{\theta}, \mathbf{y})} \bigg|_{\boldsymbol{\eta}=\boldsymbol{\eta}^*(\boldsymbol{\theta})}, \quad (2.4)$$

donde  $\tilde{\pi}(\boldsymbol{\eta} \mid \boldsymbol{\theta}, \mathbf{y})$  es una aproximación gaussiana de la condicional completa de  $\boldsymbol{\eta}$ , obtenida mediante el ajuste de la configuración modal y la curvatura en la moda. En tanto,  $\boldsymbol{\eta}^*(\boldsymbol{\theta})$  es la moda de la condicional completa para  $\boldsymbol{\eta}$  dado un valor de  $\boldsymbol{\theta}$ . Esta aproximación de Laplace será exacta si  $\pi(\boldsymbol{\eta}, \boldsymbol{\theta} \mid \mathbf{y})$  es gaussiana. La aproximación de la ecuación (2.4) puede mejorar al usar transformaciones estabilizadoras de la varianza de  $\boldsymbol{\theta}$ , que tiendan a disminuir las colas largas y reducir la asimetría, lo que resulta en densidades posteriores menos complejas.

- **Segunda actividad:** Para la aproximación de  $\pi(\boldsymbol{\eta} \mid \boldsymbol{\theta}, \mathbf{y})$ , existen tres opciones, una de ellas es la opción de aproximación gaussiana  $\tilde{\pi}(\boldsymbol{\eta} \mid \boldsymbol{\theta}, \mathbf{y})$ , la cual si bien es más rápida de obtener, en algunas situaciones produce errores en la ubicación y no logra capturar el comportamiento de asimetría (Rue y Martino, 2007). Siendo por ello preferible realizar una aproximación de Laplace del tipo:

$$\tilde{\pi}_{LA}(\eta_i \mid \boldsymbol{\theta}, \mathbf{y}) \propto \frac{\pi(\boldsymbol{\eta}, \boldsymbol{\theta} \mid \mathbf{y})}{\tilde{\pi}(\boldsymbol{\eta}_{-i} \mid \eta_i, \boldsymbol{\theta}, \mathbf{y})} \Big|_{\boldsymbol{\eta}_{-i} = \boldsymbol{\eta}_{*-i}(\eta_i, \boldsymbol{\theta})}, \quad (2.5)$$

donde  $i = 1, \dots, n$ ,  $\tilde{\pi}(\boldsymbol{\eta}_{-i} \mid \eta_i, \boldsymbol{\theta}, \mathbf{y})$  es una aproximación gaussiana con la configuración modal  $\boldsymbol{\eta}_{*-i}(\eta_i, \boldsymbol{\theta})$ . Debido a que la aproximación de Laplace puede ser muy costosa en términos computacionales, una tercera opción predeterminada en el paquete R-INLA es la aproximación simplificada de Laplace, esta aproximación tiene un menor costo en tiempo y soluciona satisfactoriamente las inexactitudes de ubicación y sesgo de la aproximación gaussiana. Para más detalles sobre estas aproximaciones ver Rue et al. (2009).

- **Tercera actividad:** Rue et al. (2009) proponen varios esquemas de exploración dependiendo del número de hiperparámetros. Todos estos esquemas requieren una reparametrización del espacio  $\boldsymbol{\theta}$  para hacer que la densidad sea más regular. Sin pérdida de generalidad, se asume que  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m) \in \mathbb{R}^m$  y se procede de la siguiente manera: se encuentra la moda  $\boldsymbol{\theta}^*$  de  $\tilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ , luego se calcula la matriz Hessiana  $\mathbf{H}$  a partir de  $\boldsymbol{\Sigma} = \mathbf{H}^{-1} = \mathbf{V} \mathbf{A} \mathbf{V}^\top$ , la matriz de covarianza de  $\boldsymbol{\theta}$  si la densidad fuese gaussiana. Luego se estandariza  $\boldsymbol{\theta}$  para obtener una nueva variable:  $\mathbf{z} = (\mathbf{V} \mathbf{A}^{1/2})^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ . Este proceso es útil para construir una malla cubriendo la mayor parte de la masa de probabilidad  $\tilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$  y mejorar la precisión de las aproximaciones.

Finalmente, como resultado de estas actividades, las fdp marginales a posteriori aproxi-



madras son calculadas numéricamente:

$$\begin{aligned}\tilde{\pi}(\boldsymbol{\eta}_i | \mathbf{y}) &= \sum_k \tilde{\pi}(\boldsymbol{\eta}_i | \boldsymbol{\theta}^{(k)}, \mathbf{y}) \tilde{\pi}(\boldsymbol{\theta}^{(k)} | \mathbf{y}) \Delta \boldsymbol{\theta}^{(k)}, \\ \tilde{\pi}(\theta_j | \mathbf{y}) &= \sum_k \tilde{\pi}(\boldsymbol{\theta}^{(k)} | \mathbf{y}) \Delta \boldsymbol{\theta}_{-j}^{(k)},\end{aligned}$$

donde  $\boldsymbol{\theta}_{-j}$  denota el vector de  $\boldsymbol{\theta}$  con sus  $j$ -ésimos elementos excluidos.

La clase de modelos que pueden expresarse en forma de un MGL, y por lo tanto sus parámetros pueden ser estimados a través INLA, es muy amplia e incluye, entre otros, los modelos espacio-temporales. La inferencia bayesiana utilizando INLA puede implementarse mediante el software R-INLA (Rue et al., 2012). La sintaxis básica para ejecutar modelos en R-INLA es muy similar en apariencia a la de los modelos lineales generalizados y sigue la forma general de: fórmula, data y familia que corresponde a la especificación de la distribución de la data, es decir, la sintaxis por ejemplo tiene la forma `inla(formula=y ~ x + f(.), data, family="poisson")`. En la fórmula la especificación de los efectos aleatorios se realiza a través de  $f(\cdot)$ . Para este último componente, algunos ejemplos incluyen:  $f(i, \text{model}="iid")$  que define un modelo gaussiano independiente;  $f(i, \text{model}="bym")$  que refiere a un modelo Besag, York y Mollié (BYM);  $f(i, \text{model}="rw1")$  que utiliza un modelo de paseo aleatorio RW1.

## Capítulo 3

# Modelos espacio-temporales

En entornos donde los conteos de enfermos se observan a lo largo del tiempo, se pueden utilizar modelos espacio-temporales que tengan en cuenta no solo la estructura espacial sino también las correlaciones temporales e interacciones espacio-temporales (Martínez-Beneito et al., 2008; Ugarte et al., 2014). De esta forma, siendo  $y_{it}$  la incidencia de la enfermedad en el área  $A_i$  para  $i = 1, \dots, n$  en el tiempo  $t$  para  $t = 1, \dots, T$ , se puede asumir que  $y_{it}$  sigue una distribución Poisson:

$$\begin{aligned} y_{it} &\sim \text{Poisson}(\lambda_{it} = E_{it}\theta_{it}) \\ \log(\theta_{it}) &= \alpha + \xi_i + f_t, \end{aligned} \tag{3.1}$$

donde  $E_{it}$  representa la población en riesgo en el área  $A_i$  en el tiempo  $t$ ,  $\theta_{it}$  es la tasa de incidencia en el área  $A_i$  en el tiempo  $t$ ,  $\alpha$  representa el intercepto,  $\xi_i$  está asociado a los efectos espaciales y  $f_t$  se puede especificar como una estructura paramétrica (tendencia temporal) o no paramétrica (paseo aleatorio o RW).

### 3.1. Estructura espacial

Para el mapeo de enfermedades, el modelo jerárquico bayesiano más común es el BYM propuesto por Besag, York y Mollié (Besag, 1974). Este modelo tiene en cuenta que los datos pueden estar correlacionados espacialmente, y las observaciones en áreas vecinas pueden tener características más similares a las observaciones en áreas lejanas; por lo que, incorpora efectos espaciales aleatorios no estructurados para explicar la variabilidad espacial que no puede ser explicada por patrones claros o predecibles; e incorpora efectos espaciales aleatorios

estructurados para explicar la asociación espacial entre vecinos.

Así, al usar la estructura del modelo BYM, se puede especificar  $\xi_i = u_i + \nu_i$ , donde  $u_i$  representa los efectos aleatorios estructurados cuya distribución sigue el modelo ICAR y  $\nu_i$  representa los efectos aleatorios no estructurados, tal que  $\nu_i \sim \mathcal{N}(0, \sigma_\nu^2)$ , siendo  $\tau_\nu = 1/\sigma_\nu^2$  la precisión. La inclusión de efectos aleatorios, tanto espacialmente estructurados como no estructurados, contribuyen a suavizar las tasas de enfermedad tanto a nivel global como local (Bell y Broemeling, 2000). El parámetro de variabilidad no estructurada,  $\nu_i$  es modelado usando una a priori normal; este parámetro proporciona información sobre la variabilidad global (*global smoothing*) no espacial de las tasas de enfermedad, es decir proporciona flexibilidad para modelar variaciones en las tasas de enfermedad que no pueden ser explicadas por relaciones espaciales. Mientras que el componente  $u_i$ , describe la variabilidad de los riesgos de enfermedad en relación con las áreas vecinas (*local smoothing*), este último componente sigue comúnmente una distribución ICAR o CAR, que permite capturar la autocorrelación entre áreas adyacentes.

A través del modelo ICAR, el cual será utilizado en el presente trabajo, la distribución de  $\mathbf{u} = (u_1, \dots, u_n)^\top$  puede típicamente estar especificada via las distribuciones condicionales:

$$u_i \mid u_{j \neq i} \sim \mathcal{N}(m_i, s_i^2)$$

$$m_i = \frac{\sum_{j \in \mathcal{N}(i)} u_j}{\#\mathcal{N}(i)} \quad \text{y} \quad s_i^2 = \frac{\sigma_u^2}{\#\mathcal{N}(i)},$$

donde  $\mathcal{N}(i)$  es el conjunto de áreas vecinas del área  $A_i$ ,  $\#\mathcal{N}(i)$  es el número de áreas que comparten límite geográfico con el área  $A_i$ , en tanto,  $\sigma_u^2$  es una varianza espacial, a partir de la cual se define el parámetro de precisión  $\tau_u = (\sigma_u^2)^{-1}$ .

Cabe resaltar que el componente espacialmente estructurado  $\mathbf{u}$  se modela como un GMRF (Rue y Held, 2005), pues la distribución de  $\mathbf{u}$  es dada por:

$$\pi(\mathbf{u} \mid \tau_u) \propto \exp\left(-\frac{\tau_u}{2} \mathbf{u}^\top \mathbf{Q}_u \mathbf{u}\right), \quad (3.2)$$

donde la matriz de precisión puede ser reescrita como  $\mathbf{Q}_u = \tau_u \mathbf{R}_u$  donde  $\mathbf{R}_u$  es la matriz  $(\mathbf{D}_W - \mathbf{W})$  cuyas entradas diagonales son iguales a  $n_i$  (i.e. el número de vecinos de la región

$i$ ) y los elementos fuera de la diagonal son iguales a -1 si  $i$  es vecino de  $j$  y 0 en caso contrario:

$$\mathbf{R}_u = \begin{cases} n_i, & \text{si } i = j \\ -1, & \text{si } i \sim j \\ 0, & \text{otro caso.} \end{cases} \quad (3.3)$$

Así la distribución del componente espacial estructurado  $\mathbf{u}$  en el modelo BYM es una Normal de la forma  $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, [\tau_u (\mathbf{D}_W - \mathbf{W})]^{-1})$ .

Por su parte, para el componente espacial no estructurado  $\boldsymbol{\nu}$  la a priori, corresponde a un GMRF cuya matriz de estructura es la matriz identidad, distribuyéndose normalmente con  $\boldsymbol{\nu} = (\nu_1, \dots, \nu_n)^\top \sim \mathcal{N}(\mathbf{0}, \tau_\nu^{-1} \mathbf{I}_n)$ . Específicamente,  $\mathbf{I}_n$  es una matriz identidad y  $\tau_\nu$  el parámetro de precisión.

## 3.2. Estructura espacio-temporal

El modelo BYM puede extenderse al contexto espacio-temporal. La estructura a priori CAR o ICAR puede definir estructuras de vecindad a través del espacio y el tiempo, de modo que el conjunto de vecinos de una región incluye no solo a los vecinos espaciales, sino también su propio valor en los periodos de tiempo previos y posteriores (Carlin y Xia, 1999). En este contexto, un amplio rango de modelos espacio-temporales para el mapeo de enfermedades han sido propuestos en la literatura (Moraga, 2020). En términos del presente trabajo, se definen los modelos detallados a continuación:

### 3.2.1. Modelo 1: Modelo jerárquico de tendencia lineal paramétrica

La clásica formulación paramétrica fue introducida por Bernardinelli et al. (1995). Este modelo se corresponde con un modelo bayesiano paramétrico con una tendencia lineal en el tiempo y una tendencia temporal diferencial para cada área. El modelo paramétrico asume que el predictor lineal puede ser escrito como:

$$\log(\theta_{it}) = \alpha + \xi_i + (\beta + \delta_i) \times t, \quad (3.4)$$

donde  $\beta$  representa el efecto de tendencia lineal temporal global y  $\delta_i$  es un efecto aleatorio que se distribuye normalmente, específicamente  $\delta_i \stackrel{ind}{\sim} \mathcal{N}(0, \tau_\delta^{-1})$ . Este efecto aleatorio  $\delta_i$  es denominado tendencia diferencial de la  $i$ -ésima área, y denota la modificación de la tendencia lineal para cada área  $i$  en el tiempo  $t$ , por lo que representa la cantidad en la que la tendencia temporal del área específica  $i$  difiere de la tendencia temporal global  $\beta$ . En ese sentido, si  $\delta_i < 0$  la tendencia de un área específica tiene una pendiente menos pronunciada que la tendencia global. En resumen en este modelo, cada una de las áreas tiene su propio perfil de riesgo a lo largo del tiempo, cuya intersección está dada por  $\alpha + u_i + \nu_i$  y cuya tendencia (pendiente) está dada por  $(\beta + \delta_i) \times t$ .

La aplicación de la INLA en el modelo jerárquico de tendencia lineal paramétrica se puede definir como un MGL:

1. Primer nivel: Las variables respuesta son independientes condicionalmente dado el campo gaussiano latente e hiperparámetros:

$$y_{it} \mid \boldsymbol{\eta}, \boldsymbol{\theta} \stackrel{ind}{\sim} \text{Poisson}(\lambda_{it}),$$

donde  $\lambda_{it} = E_{it}\theta_{it}$  y

$$\log(\lambda_{it}) = \log(E_{it}) + \log(\theta_{it})$$

$$\log(\theta_{it}) = \alpha + u_i + \nu_i + (\beta + \delta_i) \times t.$$

La función de verosimilitud está dada por:

$$\begin{aligned} L(\mathbf{y} \mid \boldsymbol{\eta}, \boldsymbol{\theta}) &= \prod_{t=1}^T \prod_{i=1}^n \pi(y_{it} \mid \boldsymbol{\eta}, \boldsymbol{\theta}) \\ &= \left[ \prod_{t=1}^T \prod_{i=1}^n \frac{\lambda_{it}^{y_{it}} e^{-\lambda_{it}}}{y_{it}!} \right], \end{aligned}$$

donde  $\pi(y_{it} \mid \boldsymbol{\eta}, \boldsymbol{\theta})$  es la fdp de los datos.

2. Segundo nivel: El campo aleatorio gaussiano de Markov, es definido por

$\boldsymbol{\eta} = (\alpha, \beta, u_1, u_2, \dots, u_n, \nu_1, \nu_2, \dots, \nu_n, \delta_1, \delta_2, \dots, \delta_n) = (\alpha, \beta, \mathbf{u}, \boldsymbol{\nu}, \boldsymbol{\delta})$ , donde

$$\begin{aligned}\alpha &\sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2), \\ \beta &\sim \mathcal{N}(\mu_\beta, \sigma_\beta^2), \\ \mathbf{u} &\sim \mathcal{N}(0, [\tau_u(\mathbf{D}_W - \mathbf{W})]^{-1}), \\ \boldsymbol{\nu} &\sim \mathcal{N}(0, \tau_\nu^{-1} \mathbf{I}_n), \\ \boldsymbol{\delta} &\sim \mathcal{N}(0, \tau_\delta^{-1} \mathbf{I}_n).\end{aligned}$$

En particular el GMRF captura la variabilidad espacial a través del componente espacial estructurado ( $\mathbf{u}$ ). Esto se logra con una matriz de precisión  $\mathbf{Q}_u$  que refleja la dependencia entre áreas vecinas, como se explicó previamente.

3. Tercer nivel: El vector de hiperparámetros es definido por  $\boldsymbol{\theta} = (\tau_u, \tau_\nu, \tau_\delta)$ . Se asignó una distribución Gamma(1, 0.0005) a todos los parámetros de precisión definidos en  $\boldsymbol{\theta}$ .

### 3.2.2. Modelo 2: Modelo jerárquico de tendencia dinámica no paramétrica a través de RW1

El Modelo 1, si bien es flexible al permitir que cada unidad de área presente su propia tendencia, es a su vez restrictivo al requerir que esas tendencias sean lineales. Este supuesto de linealidad en  $\delta_i$ , puede ser no adecuado en muchas aplicaciones (Knorr-Held, 2000), limitación que puede ser evitada mediante el Modelo 2, un modelo dinámico para el predictor lineal. Este modelo no asume linealidad e incluye efectos aleatorios espaciales y temporales. El modelo combina efectos aleatorios temporales y espaciales de manera aditiva, tal como se observa en la siguiente fórmula:

$$\log(\theta_{it}) = \alpha + u_i + \nu_i + \gamma_t + \phi_t, \quad (3.5)$$

donde los componentes  $\alpha, u_i$  and  $\nu_i$  tienen la misma parametrización que el Modelo 1. Mientras que  $\phi_t$  denota un efecto temporal no estructurado, que es modelado con una distribución normal, es decir,  $\phi_t \stackrel{ind}{\sim} \mathcal{N}(0, \tau_\phi)$ , y finalmente el término  $\gamma_t$  representa el efecto temporal estructurado, modelado dinámicamente usando un paseo aleatorio de orden 1 (RW1) a través



de una estructura de la siguiente forma:

$$\begin{aligned} \gamma_t \mid \gamma_{-t} &\sim \mathcal{N}(\gamma_{t+1}, \tau_\gamma) && \text{para } t = 1, \\ \gamma_t \mid \gamma_{-t} &\sim \mathcal{N}\left(\frac{\gamma_{t-1} + \gamma_{t+1}}{2}, \frac{\tau_\gamma}{2}\right) && \text{para } t = 2, \dots, T-1, \\ \gamma_t \mid \gamma_{-t} &\sim \mathcal{N}(\gamma_{t-1}, \tau_\gamma) && \text{para } t = T. \end{aligned}$$

Luego cada uno de los vectores aleatorios  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_T)^\top$ ,  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_T)^\top$ ,  $\mathbf{u} = (u_1, \dots, u_n)^\top$ ,  $\boldsymbol{\nu} = (\nu_1, \dots, \nu_n)^\top$  presenta una distribución multivariada gaussiana. Así, para los componentes  $\mathbf{u}$  y  $\boldsymbol{\nu}$ , su distribución toma la forma especificada en el Modelo 1. Los efectos temporales estructurados  $\boldsymbol{\gamma}$  son modelados con una distribución normal con media cero y matriz de precisión  $\tau_\gamma \mathbf{K}_\gamma$ , donde

$$\mathbf{K}_\gamma = \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \vdots & \vdots & \vdots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{bmatrix}.$$

Para el componente temporal no estructurado  $\boldsymbol{\phi}$  se asume una distribución normal con media cero y matriz de precisión  $\tau_\phi \mathbf{K}_\phi$ , donde la matriz  $\mathbf{K}_\phi$  es la matriz identidad ( $\mathbf{K}_\phi = \mathbf{I}$ ).

El modelo jerárquico de tendencia dinámica no paramétrica con RW1, se puede definir como un modelo gaussiano latente:

1. **Primer nivel:** Asumiendo independencia condicional:

$$y_{it} \mid \boldsymbol{\eta}, \boldsymbol{\theta} \sim \text{Poisson}(\lambda_{it}),$$

donde  $\lambda_{it} = E_{it}\theta_{it}$ , y

$$\log(\lambda_{it}) = \log(E_{it}) + \log(\theta_{it}),$$

$$\log(\theta_{it}) = \alpha + u_i + \nu_i + \gamma_t + \phi_t.$$



2. **Segundo nivel:** El campo aleatorio gaussiano de Markov es dado por

$$\boldsymbol{\eta} = (\alpha, u_1, u_2, \dots, u_n, \nu_1, \nu_2, \dots, \nu_n, \gamma_1, \gamma_2, \dots, \gamma_T, \phi_1, \phi_2, \dots, \phi_T),$$

y se asumen las siguientes fdp a priori:

$$\begin{aligned}\alpha &\sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2), \\ \mathbf{u} &\sim \mathcal{N}(0, [\tau_u(\mathbf{D}_W - \mathbf{W})]^{-1}), \\ \boldsymbol{\nu} &\sim \mathcal{N}(0, \tau_\nu^{-1} \mathbf{I}_n), \\ \boldsymbol{\gamma} &\sim \mathcal{N}(0, [\tau_\gamma K_\gamma]^{-1}), \\ \boldsymbol{\phi} &\sim \mathcal{N}(0, [\tau_\phi K_\phi]^{-1}).\end{aligned}$$

En resumen en este modelo, el GMRF se extiende para capturar tanto la estructura espacial como temporal, mediante el componente  $\mathbf{u}$  que establece la dependencia espacial entre regiones vecinas, y el componente  $\boldsymbol{\gamma}$  que captura la dependencia temporal a través de un RW1.

**Tercer nivel:** El vector de hiperparámetros es:

$\boldsymbol{\theta} = (\tau_u, \tau_\nu, \tau_\delta, \tau_\phi)$ . Se asignó una distribución Gamma(1, 0.0005) a todos los parámetros de precisión definidos en  $\boldsymbol{\theta}$ .

### 3.2.3. Modelo 3: Modelo jerárquico de tendencia dinámica no paramétrica e interacción espacio-temporal

Se puede expandir el Modelo 2 para permitir una interacción entre el espacio y el tiempo, lo que explicaría las diferencias en la tendencia temporal para diferentes áreas. Para ello, se incorpora el parámetro de interacción  $\delta_{it}$ , para  $i = 1, \dots, n$  y  $t = 1, \dots, T$ , usando la siguiente especificación:

$$\log(\theta_{it}) = \alpha + u_i + \nu_i + \gamma_t + \phi_t + \delta_{it}, \quad (3.6)$$

donde todos los parámetros y efectos aleatorios se definen como en el modelo 2, y  $\delta_{it}$  es una interacción entre el espacio y el tiempo. En ese sentido, en el presente trabajo se asume que los dos efectos no estructurados  $\nu_i$  y  $\phi_t$  interactúan (Knorr-Held, 2000), es decir se asume una interacción entre  $\nu_i$  y  $\phi_t$ , tal que el término  $\delta_{it}$  sigue una distribución gaussiana con media cero y presenta una matriz de precisión  $\tau_\delta \mathbf{K}_\delta$ , donde  $\tau_\delta$  es un escalar desconocido y  $\mathbf{K}_\delta$  es un

producto Kronecker de matrices, tal que:  $\mathbf{K}_\delta = \mathbf{K}_\phi \otimes \mathbf{K}_\nu = \mathbf{I} \otimes \mathbf{I} = \mathbf{I}$ . El producto Kronecker resultante se corresponde con una matriz identidad (Blangiardo y Cameletti, 2015).

La fdp a priori para el parámetro de interacción  $\delta_{it}$  de acuerdo con Knorr-Held (2000), puede escribirse como:

$$\pi(\boldsymbol{\delta} \mid \tau_\delta) \propto \exp \left( -\frac{\tau_\delta}{2} \sum_{i=1}^n \sum_{t=1}^T (\delta_{it})^2 \right),$$

es decir  $\delta_{it} \stackrel{ind}{\sim} \mathcal{N}(0, \tau_\delta^{-1})$ .

En tanto, el resto de componentes  $(u_i, \nu_i, \gamma_t, \phi_t)$  mantienen la misma distribución, matriz de precisión y distribución a priori detalladas en el Modelo 2.

La aplicación de la INLA en el modelo jerárquico de tendencia dinámica no paramétrica con RW1 y término de interacción, requiere definir el modelo como un MGL:

1. **Primer nivel:** Debido a la independencia condicional de  $y_{it}$  :

$$y_{it} \mid \boldsymbol{\eta}, \boldsymbol{\theta} \sim \text{Poisson}(\lambda_{it}),$$

donde  $\lambda_{it} = E_{it}\theta_{it}$  y usando una función de enlace logarítmica:

$$\log(\lambda_{it}) = \log(E_{it}) + \log(\theta_{it}),$$

$$\log(\rho_{it}) = \alpha + v_i + \nu_i + \gamma_t + \phi_t + \delta_{it}.$$

2. **Segundo nivel:** El campo aleatorio gaussiano de Markov es dado por

$$\boldsymbol{\eta} = (\alpha, u_1, u_2, \dots, u_n, \nu_1, \nu_2, \dots, \nu_n, \gamma_1, \gamma_2, \dots, \gamma_T, \phi_1, \phi_2, \dots, \phi_T).$$

$$\begin{aligned}
\alpha &\sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2), \\
\mathbf{u} &\sim \mathcal{N}(0, [\tau_u(\mathbf{D}_W - \mathbf{W})]^{-1}), \\
\boldsymbol{\nu} &\sim \mathcal{N}(0, \tau_\nu^{-1} \mathbf{I}_n), \\
\gamma &\sim \mathcal{N}(0, [\tau_\gamma K_\gamma]^{-1}), \\
\phi &\sim \mathcal{N}(0, [\tau_\phi K_\phi]^{-1}), \\
\delta &\sim \mathcal{N}(0, [\tau_\delta \mathbf{I}_{nT}]^{-1}).
\end{aligned}$$

3. **Tercer nivel:** El vector de de hiperparámetros es:  $\boldsymbol{\theta} = (\tau_u, \tau_\nu, \tau_\delta, \tau_\phi, \tau_\delta)$ . Se asignó una distribución Gamma(1, 0.0005) a todos los parámetros de precisión definidos en  $\boldsymbol{\theta}$ .

Estos modelos espacio-temporales son extendidos de forma relativamente directa usando la distribución Binomial Negativa, y modelos cero inflacionados usando la distribución Poisson y Binomial Negativa. Detalles sobre estos modelos se brindan en el Apéndice D.

### 3.3. Inferencia bayesiana usando INLA

En todos los modelos espacio-temporales combinando los tres niveles, la distribución conjunta a posteriori se puede expresar como:

$$\begin{aligned}
\pi(\boldsymbol{\eta}, \boldsymbol{\theta} \mid \mathbf{y}) &\propto \pi(\boldsymbol{\theta}) \pi(\boldsymbol{\eta} \mid \boldsymbol{\theta}) \prod_{t=1}^T \prod_{i=1}^n \pi(y_{it} \mid \boldsymbol{\eta}, \boldsymbol{\theta}) \\
&\propto \pi(\boldsymbol{\theta}) |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2} \boldsymbol{\eta}^\top \mathbf{Q} \boldsymbol{\eta}\right) + \sum_{t=1}^T \sum_{i=1}^n \log \pi(y_{it} \mid \boldsymbol{\eta}, \boldsymbol{\theta}).
\end{aligned}$$

Como esta fdp no tiene forma conocida, se procede a estimar los parámetros usando INLA. El INLA, tal como fue detallado en la sección 2.5.5, consta de tres actividades propuestas por Rue y Martino (2007): i) La primera actividad consiste en proponer una aproximación  $\tilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ , a la conjunta a posteriori de  $\pi(\boldsymbol{\theta} \mid \mathbf{y})$  usando la aproximación de Laplace; ii) la segunda actividad consiste en proponer una aproximación  $\tilde{\pi}(\eta_j \mid \boldsymbol{\theta}, \mathbf{y})$ , es decir de las marginales de la distribución condicional de  $\eta_j$ , considerando la data  $\mathbf{y}$ , así como los hiperparámetros  $\boldsymbol{\theta}$ , se puede realizar otra aproximación de Laplace; iii) la tercera actividad consiste combina las dos actividades anteriores utilizando integración numérica.

### 3.4. Selección del modelo

Para la selección de modelos, Watanabe (2010) propuso el criterio de información de Watanabe-Akaike o criterio de información Ampliamente Aplicable (WAIC, por sus siglas en inglés, *Widely Applicable Information Criterion*). El WAIC se obtiene usando la distribución a posteriori de  $\boldsymbol{\eta}, \boldsymbol{\theta}$ . Desde una perspectiva bayesiana, el basarse en la densidad predictiva posterior, es su principal ventaja sobre otros criterios similares. Gelman et al. (2014) afirmaron que el WAIC es particularmente útil para modelos con estructuras jerárquicas y de mezcla, y propusieron un cambio leve en la versión original del WAIC de Watanabe, como sigue:

$$\text{WAIC} = -2 \sum_{t=1}^T \sum_{i=1}^n \left[ \log \left( \frac{1}{M} \sum_{m=1}^M \pi(y_{it} | \boldsymbol{\eta}^{(m)}, \boldsymbol{\theta}^{(m)}) \right) - V_{m=1}^M \log \left( \pi(y_{it} | \boldsymbol{\eta}^{(m)}, \boldsymbol{\theta}^{(m)}) \right) \right],$$

donde  $\boldsymbol{\eta}^{(m)}, \boldsymbol{\theta}^{(m)}$  son muestras de  $\pi(\boldsymbol{\eta}, \boldsymbol{\theta} | \mathbf{y})$  y  $V_{m=1}^M(\cdot)$  es la varianza muestral. Cuanto menor sea el valor de WAIC, mejor será el modelo.

## Capítulo 4

# Estudio de Simulación

En este capítulo, se realizó un estudio de recuperación de parámetros con el objetivo de mostrar la idoneidad de los modelos y métodos de estimación propuestos. Para lo cual, se implementó la inferencia bayesiana en los modelos planteados, utilizando el enfoque INLA, a través del paquete R-INLA.

Para realizar la simulación de datos se utilizó un mapa de Perú, el cual se muestra en la Figura 4.1. El análisis se centró en los 24 departamentos y una provincia constitucional, que constituyeron las 25 localidades evaluadas durante el periodo 2010 a 2023. Los valores de los parámetros se definen al azar.



Figura 4.1: Mapa del Perú según departamentos.

## 4.1. Modelo 1

En esta sección se describe la simulación de los efectos espacio-temporales, mediante el modelo paramétrico de tendencia lineal descrito en el Capítulo 3. Los valores de los parámetros utilizados fueron:  $\alpha = 3.6$ ,  $\beta = 0.01$ ,  $\tau_u = 2500$ ,  $\tau_\nu = 0.8$  y  $\tau_\delta = 100$ . Los términos  $\alpha$  y  $\beta$  describen los efectos fijos: intercepto y tendencia lineal temporal global. Los términos  $\tau_u$ ,  $\tau_\nu$  y  $\tau_\delta$  se corresponden con los hiperparámetros de los componentes espaciales  $u$  y  $\nu$ , y de la tendencia lineal temporal de un área específica ( $\delta_i$ ). Para el Modelo 1 se utilizó el predictor lineal que fuera descrito en la ecuación (3.4), específicamente:

$$\log(\lambda_{it}) = \alpha + \xi_i + (\beta + \delta_i) \times t,$$

donde  $i = 1, 2, \dots, 25$  áreas y  $t = 1, \dots, 14$  años. Para cada área y tiempo se simuló una variable aleatoria  $y_{it}$ , con distribución Poisson,  $y_{it} \sim \text{Poisson}(\lambda_{it})$ , que representa el número de casos en cada  $i$ -ésima área ( $i = 1, 2, \dots, 25$ ) por año ( $t = 1, \dots, 14$ ). Nótese que fue asumido  $\lambda_{it} = \theta_{it}$ , es decir  $E_{it} = 1$  para todo  $i$  y  $t$ . La Figura 4.2 muestra año a año la variable respuesta simulada en cada localidad, donde el color más fuerte (rojo) indica una mayor cantidad de casos.

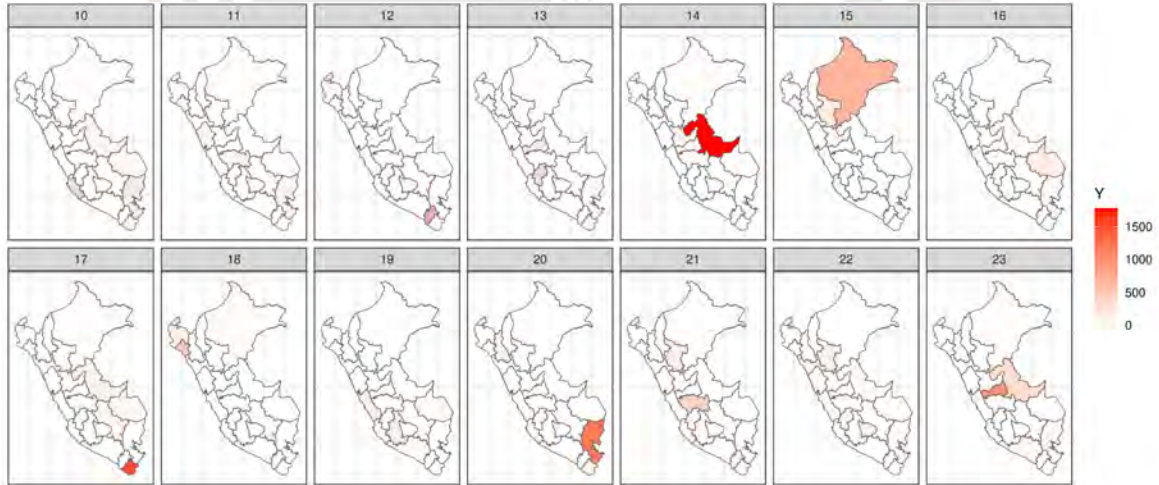


Figura 4.2: Casos simulados según el Modelo 1 entre 2010 a 2023.

Al estimar los parámetros del Modelo 1 usando INLA se obtuvieron las estimaciones a posteriori resumidas en el Cuadro 4.1. El Cuadro incluye las columnas correspondientes al valor original (valor del parámetro asignado en la simulación), la media a posteriori, la desviación estándar a posteriori (sd) y los intervalos de credibilidad al 95 %, ( $Q_{0.025}$  ,  $Q_{0.975}$ ).



Los resultados indican que, en todos los casos, los intervalos de credibilidad contienen el verdadero valor del parámetro. Los parámetros de precisión (inversas de las varianzas) asociados a los efectos aleatorios espaciales estructurados ( $\tau_u$ ) y al efecto de tendencia temporal para las áreas ( $\tau_\delta$ ) son los que presentaron la mayor incertidumbre, la cual se ve reflejada en la amplitud de los intervalos de credibilidad.

Cuadro 4.1: Tabla: Modelo 1 - Resumen de estimaciones a posteriori.

Modelo 1					
Términos	original	media	sd	$Q_{0.025}$	$Q_{0.975}$
$\alpha$	3.60	4.002	0.209	3.587	4.415
$\beta$	0.01	0.006	0.024	-0.041	0.054
$\tau_u$	2500.00	2180.335	2150.886	203.860	7890.727
$\tau_v$	0.80	0.992	0.279	0.546	1.636
$\tau_\delta$	100	75.079	21.202	41.424	124.183

En la Figura 4.3 se observan las fdp marginales a posteriori de los efectos fijos (coeficientes de regresión) e hiperparámetros del modelo.

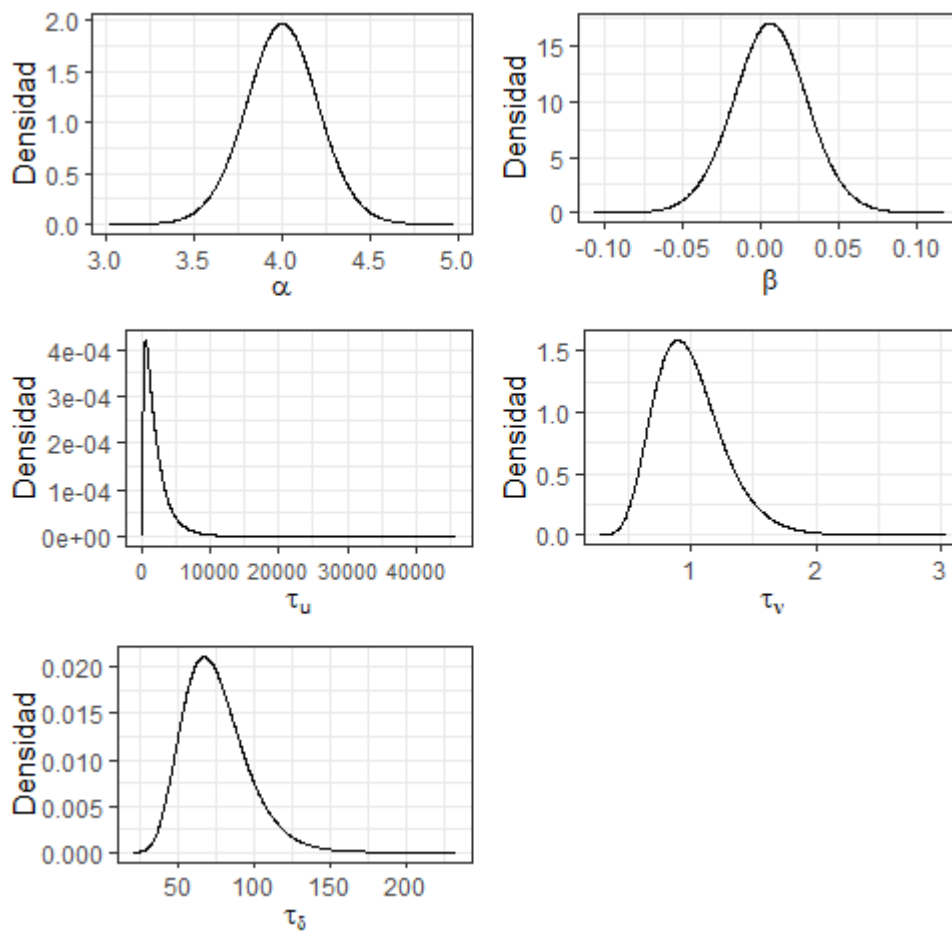


Figura 4.3: Distribuciones marginales a posteriori de los efectos fijos e hiperparámetros según el Modelo 1.

## 4.2. Modelo 2

En esta sección se describe la simulación de los efectos espacio-temporales, mediante el modelo jerárquico de tendencia dinámica no paramétrica con RW1 detallado en el Capítulo 3. Para el Modelo 2 se simuló el predictor lineal descrito en la ecuación (3.5), los valores de los parámetros fueron:  $\alpha = 3$ ,  $\tau_u = 1000$ ,  $\tau_\nu = 2500$ ,  $\tau_\gamma = 10000$  y  $\tau_\phi = 10000$ . Específicamente,

$$\log(\lambda_{it}) = \alpha + u_i + \nu_i + \gamma_t + \phi_t,$$

donde  $i = 1, 2, \dots, 25$  áreas y  $t = 1, \dots, 14$  años. Para cada área y tiempo se simuló una variable aleatoria  $y_{it}$ , con distribución Poisson,  $y_{it} \sim \text{Poisson}(\lambda_{it})$ . La Figura 4.4 muestra año a año los valores simulados por el modelo. La mayor cantidad de casos corresponde a las localidades donde el color rojo es más intenso.

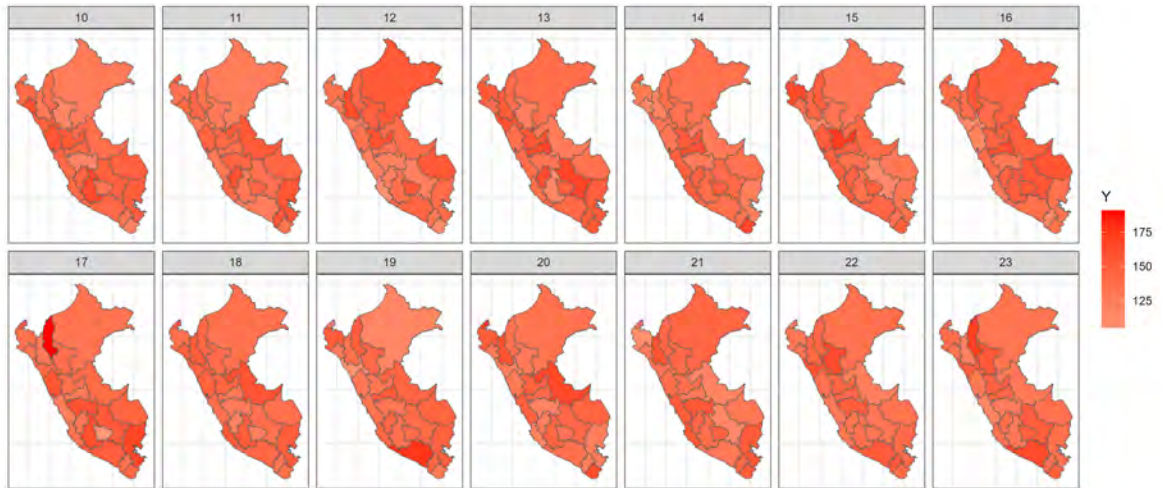


Figura 4.4: Casos simulados según el Modelo 2, del 2010 al 2023.

En el resumen de las estimaciones a posteriori del modelo ajustado (Cuadro 4.2), se observa que los intervalos de credibilidad contienen el verdadero valor del parámetro. Las estimaciones de la media a posteriori tanto para las precisiones de los efectos aleatorios estructurados como no estructurados presentaron una gran incertidumbre, la cual se ve reflejada en la amplitud de los intervalos registrados.

En la Figura 4.5 se observan las fdp marginales a posteriori. Al respecto, el efecto fijo  $\alpha$  tiene una distribución simétrica. Mientras los componentes de precisión espaciales y temporales mostraron mayor incertidumbre según la magnitud de su variabilidad.

Cuadro 4.2: Tabla: Modelo 2 - Resumen de estimaciones a posteriori

Modelo 2					
Términos	original	media	sd	$Q_{0.025}$	$Q_{0.975}$
$\alpha$	3.00	2.979	0.013	2.953	3.005
$\tau_u$	1000	3362.008	2071.732	965.386	8769.631
$\tau_v$	2500	3709.771	2720.984	724.448	10829.325
$\tau_\gamma$	10000	35446.198	27143.061	6200.392	106609.203
$\tau_\phi$	10000	28018.561	26601.262	3344.443	98519.036

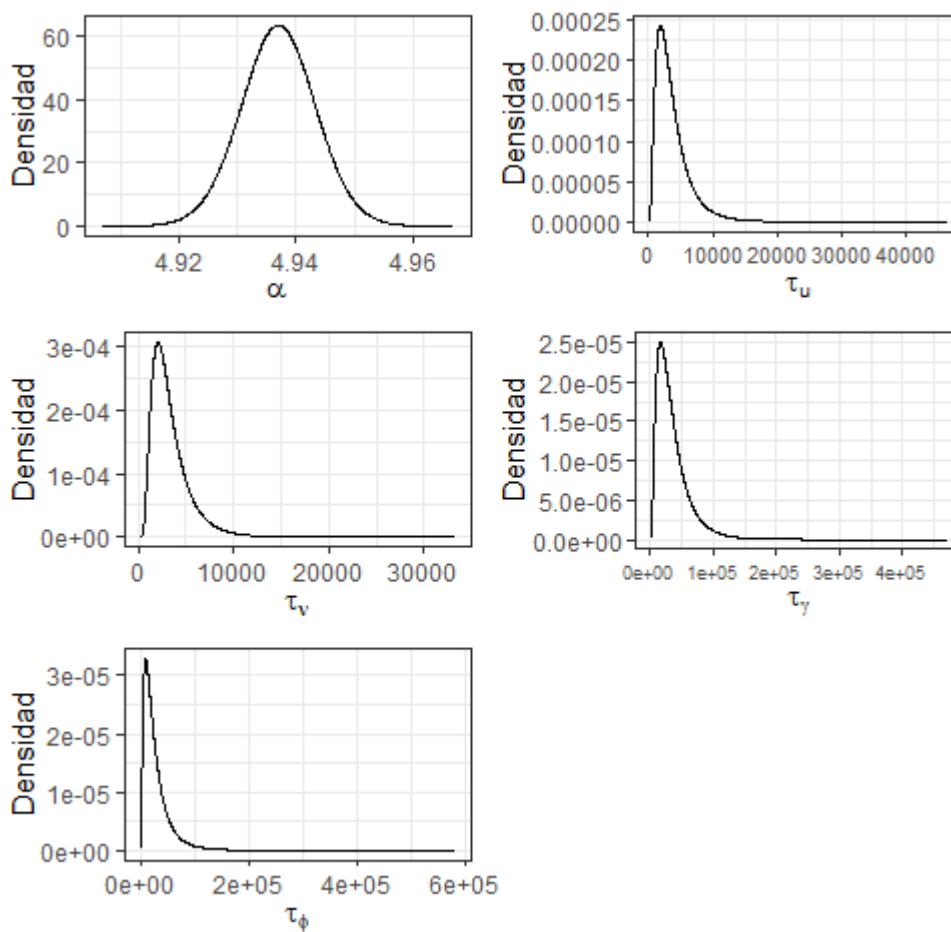


Figura 4.5: Modelo 2: Distribuciones marginales a posteriori.

En el Modelo 2 el efecto aleatorio temporal fue modelado como un RW1. La Figura 4.6 muestra la media a posteriori estimada de la estructura del RW1, indicando cómo el efecto aleatorio evoluciona temporalmente. En el gráfico se observan fluctuaciones que reflejan cómo el efecto aleatorio puede variar considerablemente a lo largo del tiempo, mostrando la influencia del componente temporal en el modelo.

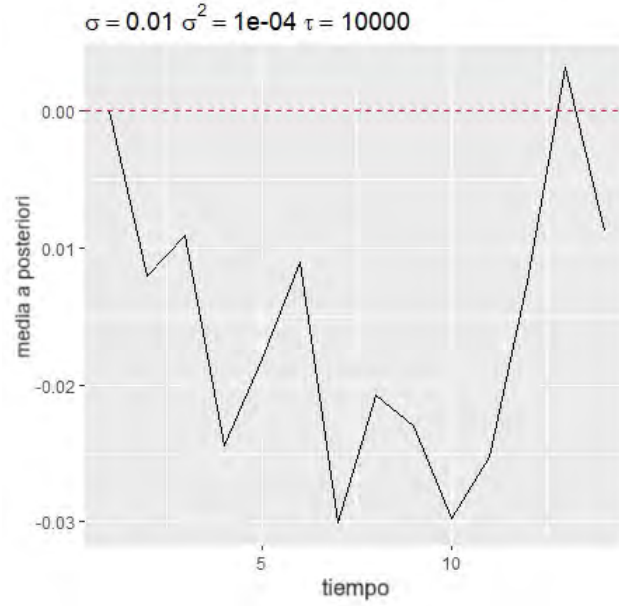


Figura 4.6: Estimación de la media a posteriori de  $\gamma_t$  según el Modelo 2.

### 4.3. Modelo 3

En esta sección se describe la simulación de los efectos espacio-temporales, mediante el modelo jerárquico de tendencia dinámica no paramétrica con interacción espacio-temporal. Para el Modelo 3 se simuló el predictor lineal descrito en la ecuación (3.6), los valores de los parámetros fueron:  $\alpha = 3$ ,  $\tau_u = 2600$ ,  $\tau_\nu = 25$ ,  $\tau_\gamma = 2000$ ,  $\tau_\phi = 10000$  y  $\tau_\delta = 18$ . Específicamente:

$$\log(\lambda_{it}) = \alpha + u_i + \nu_i + \gamma_t + \phi_t + \delta_{it},$$

donde  $i = 1, 2, \dots, 25$  áreas y  $t = 1, \dots, 14$  años. Para cada área y tiempo se simuló una variable aleatoria  $y_{it}$ , con distribución Poisson,  $y_{it} \sim \text{Poisson}(\lambda_{it})$ . Los valores simulados por el Modelo 3 se muestran año a año en la Figura 4.7. En este modelo se agrega el término de interacción entre los efectos temporales y espaciales no estructurados.



Figura 4.7: Casos simulados según el Modelo 3 del año 2010 al 2023.

En el resumen de las estimaciones a posteriori del modelo ajustado se muestran en el Cuadro 4.3. Si bien se observa que los intervalos de credibilidad contienen el verdadero valor del parámetro, las estimaciones de los hiperparámetros presentaron una gran incertidumbre, la cual se ve reflejada en la desviación estándar y en la amplitud de los intervalos de credibilidad. El parámetro que presentó mejor ajuste fue el intercepto  $\alpha$ .

Cuadro 4.3: Tabla: Modelo 3 - Resumen de estimaciones a posteriori.

Modelo 3					
Términos	original	media	sd	$Q_{0.025}$	$Q_{0.975}$
$\alpha$	3.00	2.985	0.034	2.918	3.051
$\tau_u$	2600.00	2870.763	3288.122	248.551	11505.058
$\tau_v$	25.00	53.533	20.647	23.862	103.841
$\tau_\gamma$	2000.00	21659.837	26629.294	1151.508	91841.099
$\tau_\phi$	10000.00	27166.624	30124.140	2395.223	106583.990
$\tau_\delta$	18	24.094	4.106	17.304	33.414

La Figura 4.8 muestra las distribuciones marginales a posteriori. La distribución del efecto fijo ( $\alpha$ ) es aproximadamente simétrica y se concentra alrededor de 3.0, con poca variabilidad alrededor de este valor central. Adicionalmente, las distribuciones de los parámetros de precisión de los efectos temporales y espaciales, indican una alta variabilidad, en especial para el término  $\tau_\gamma$  que captura dinámicas temporales que pueden cambiar considerablemente en diferentes puntos en el tiempo. La precisión del componente de interacción presenta una distribución menos asimétrica, por lo que su variabilidad es moderada.



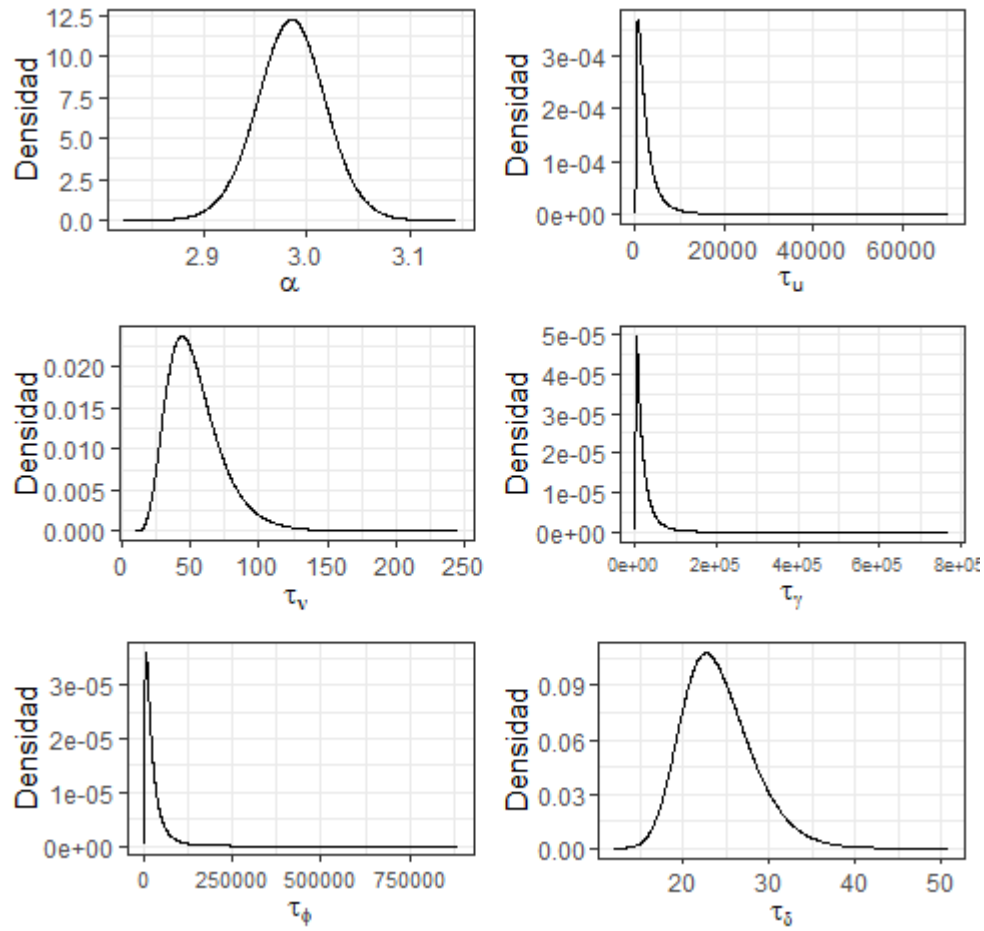


Figura 4.8: Modelo 3: Distribuciones marginales a posteriori.

La Figura 4.9 muestra la estimación a posteriori de la estructura del efecto aleatorio  $\gamma_t$  durante el periodo bajo estudio. Las fluctuaciones observadas reflejan un comportamiento similar al del Modelo 2.

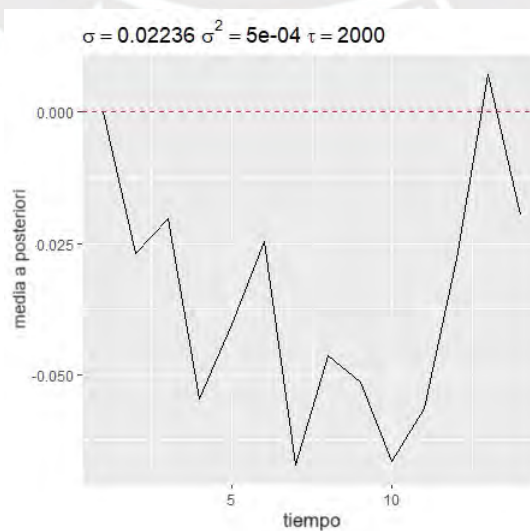


Figura 4.9: Estimación de la media a posteriori de  $\gamma_t$  según el Modelo 3.



## Capítulo 5

# Aplicación

En los capítulos previos se detallaron tres modelos espacio-temporales. Un primer modelo conocido como modelo paramétrico, introducido por Bernardinelli et al. (1995), cuya formulación incluye componentes espaciales estructurados y no estructurados, y asume un efecto lineal del tiempo en cada área ( $\delta_i$ ). Un segundo modelo no paramétrico, donde no se asume la linealidad temporal, el término  $\gamma_t$  representa el efecto temporal estructurado que es modelado utilizando un paseo aleatorio de orden 1 y el término  $\phi_t$  representa el efecto temporal no estructurado. Un tercer modelo que es una expansión del segundo modelo y que agrega la interacción  $\delta_{it}$  entre el componente espacial y temporal no estructurados.

En este capítulo se ajustan estos modelos espacio-temporales a datos reales de casos de dengue en el Perú. La información epidemiológica de Perú fue obtenida a partir de los registros de dengue, tanto confirmados como probables, provistos por el Instituto Nacional de Salud (INS) y los registros climáticos provinieron de los datos de libre disposición POWER (Prediction Of Worldwide Energy Resources), los cuales pueden ser consultados a través del enlace <https://power.larc.nasa.gov/>.

En la presente tesis, se presentan dos aplicaciones para los modelos espacio temporales: i) una primera aplicación, que incluyó la incidencia de dengue reportada en 25 áreas (correspondientes a 24 departamentos y la provincia constitucional del Callao) a través de las 52 semanas epidemiológicas del año 2023 y ii) una segunda aplicación, que incluyó la incidencia anual de dengue reportada en 196 locaciones, a través de un periodo de tiempo de catorce años (i.e. 2010 a 2023), evaluando el ajuste de los modelos al agregar covariables climáticas.

## 5.1. Aplicación 1: Dengue por semanas epidemiológicas

Para la primera aplicación de los modelos, los datos de dengue, correspondieron a los casos reportados por semana epidemiológica, en los 24 departamentos, y la provincia constitucional de Callao (Figura 4.1) durante el año 2023.

### 5.1.1. Análisis exploratorio de los datos

Los valores máximos de casos de dengue, en el año 2023, se registraron en el departamento de Lima, siendo el valor más alto el registrado en la semana 21 (5318 casos), seguido del departamento de Piura con registros de 4578 y 4276 casos, durante las semanas 19 y 29 respectivamente. En el Apéndice A, se incluyen los diagramas de caja de los casos de dengue por semana epidemiológica.

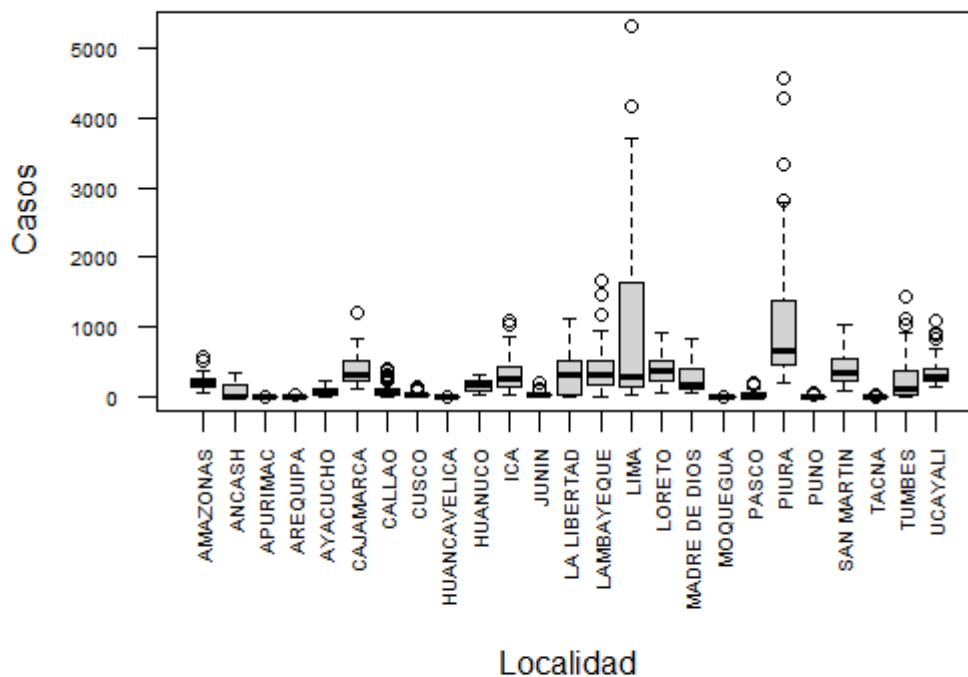


Figura 5.1: Casos de dengue por departamento y el Callao en el año 2023.

Entre las semanas epidemiológicas 1 a 13, en Piura se observó un incremento progresivo de casos hasta alcanzar las 2784 observaciones, valor muy superior al Q3 (416 casos). Mientras que Ucayali alcanzó el pico de casos durante la semana 8 para luego decrecer en sus registros. Los departamentos ubicados en la región Amazónica, además de aquellos ubicados en la costa norte del país, como Piura, fueron los que presentaron mayor cantidad de registros. En la

Figura 5.2, se incluyen los resultados correspondientes a los casos de dengue reportados entre las semanas epidemiológicas 1 a la 13.

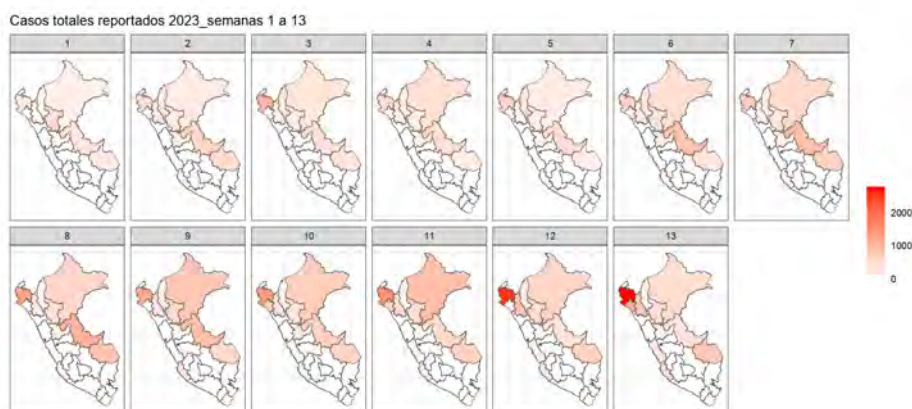


Figura 5.2: Casos de dengue - Semanas 1 a 13, Año 2023.

Entre las semanas epidemiológicas 14 a 26, siguiendo la tendencia de la semana 13, Piura fue el departamento con mayor cantidad de registros entre las semanas 14 a la 16; seguido del departamento de Lima. En la semana 21, Lima reportó su mayor registro del año (5318 casos). En la Figura 5.3, puede observarse que los departamentos de la Amazonía disminuyeron sus registros, mientras que en la costa central y norte se observaron los mayores registros.

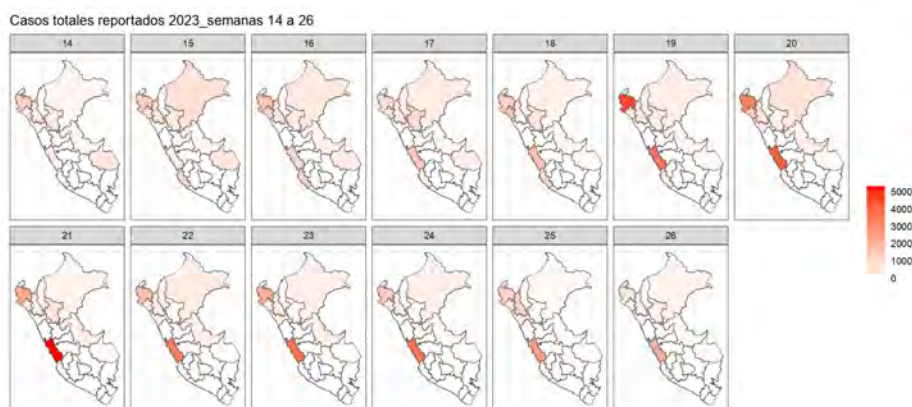


Figura 5.3: Casos de dengue - Semanas 14 a 26, año 2023.

Entre las semanas epidemiológicas 27 a 36, el departamento de Lima alcanzó la mayor cantidad de registros, mientras Piura, obtuvo su mayor registro del año, 4276 casos, en

la semana 29. La Figura 5.4, indica que la mayor cantidad de casos se concentró en los departamentos de la costa central y costa norte.

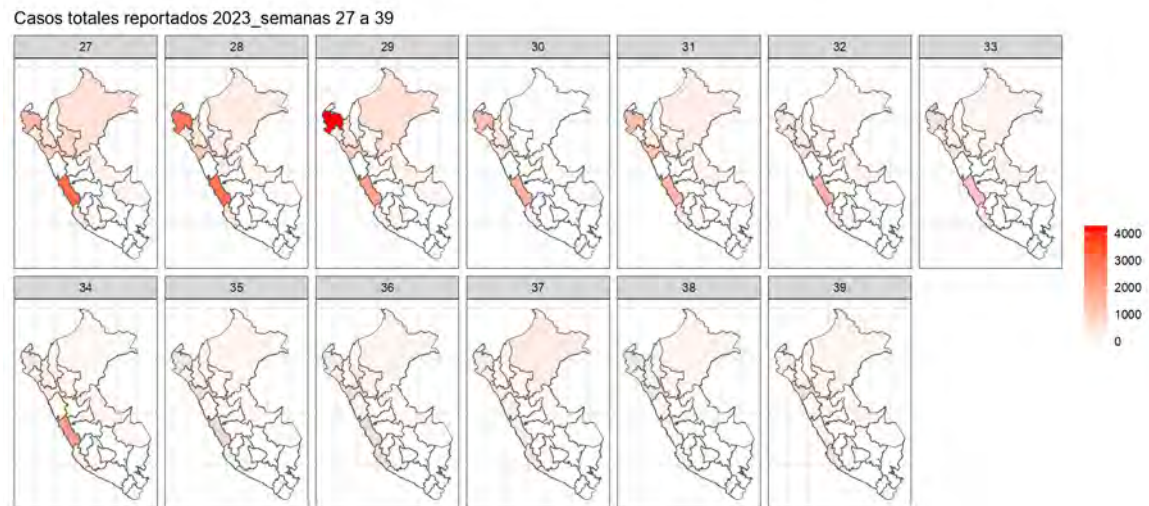


Figura 5.4: Casos de dengue - Semanas 14 a 26, año 2023.

Entre las semanas epidemiológicas 40 a 52, (Figura 5.5), se visualiza como al final del año, se incrementan los casos en la Amazonía, manteniéndose una importante cantidad de registros en la costa norte.

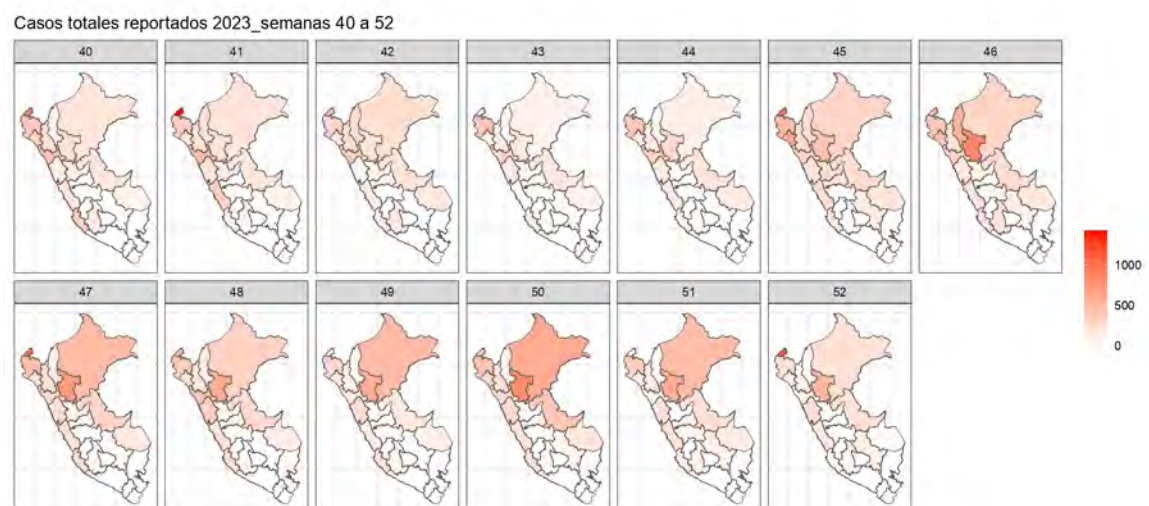


Figura 5.5: Casos de dengue - Semanas 40 a 52, año 2023.



### 5.1.2. Aplicación de modelos - casos de dengue

Sea  $y_{it}$  definida como el número de casos de dengue en el departamento  $i = 1, 2, \dots, 25$ , durante las semanas epidemiológicas  $t = 1, 2, \dots, 52$ ; una variable respuesta cuya densidad se muestra en la Figura B.1. (Apéndice B), y para la cual se tuvieron en cuenta, las siguientes distribuciones de conteo:

- Poisson:  $y_{it} \sim \text{Poisson}(\lambda_{it})$ ,
- Binomial Negativa (BN):  $y_{it} \sim \text{BN}(\lambda_{it}, \Gamma)$ , donde  $\Gamma$  es un parámetro de sobredispersión,
- Poisson Cero Inflacionada (ZIP):  $y_{it} \sim \text{ZIP}(\lambda_{it}, p_{it})$ , donde  $p_{it}$  es la probabilidad cero inflacionada,
- Binomial Negativa Cero Inflacionada (ZINB):  $y_{it} \sim \text{ZINB}(\lambda_{it}, p_{it}, \Gamma)$ .

Se utilizaron modelos espacio-temporales para ajustar los datos; que tras el análisis exploratorio visualizaron una distribución espacial semanal, en la que los departamentos cercanos alcanzaron cantidades similares de casos, observándose además patrones temporales. En ese contexto, para el análisis se definió el predictor lineal del Modelo 1 de acuerdo a la fórmula de la ecuación (3.4), mientras que los predictores lineales de los Modelos 2 y 3, fueron ejecutados según las fórmulas de las ecuaciones (3.5) y (3.6). Se realizó la inferencia bayesiana utilizando el paquete R-INLA ([www.r-inla.org](http://www.r-inla.org)) ajustando los tres modelos planteados para cada distribución de  $y_{it}$  y asumiendo en todos los casos que  $\lambda_{it} = \theta_{it}$ .

### Resultados

De acuerdo a los valores WAIC (Cuadro 5.1), el Modelo 3 evaluado para los casos de dengue, considerando la distribución Poisson, fue el que logró el mejor ajuste pues tiene el menor valor de WAIC, seguido del Modelo 3 evaluado para la distribución ZIP.

Cuadro 5.1: Criterios de información (WAIC) para los diferentes modelos y distribuciones

Modelos	Poisson	BN	ZIP	ZINB
<b>Modelo 1</b>	58817.13	12945.35	57083.53	13417.36
<b>Modelo 2</b>	84354.30	13209.86	80058.44	13570.70
<b>Modelo 3</b>	<b>9053.57</b>	13209.80	<b>9678.72</b>	13566.77

La Figura 5.6 muestra la comparación entre los valores observados y los valores estimados por los diferentes modelos y distribuciones. Observándose que las mejores correlaciones fueron

detectadas en el Modelo 3 con distribución Poisson y en el Modelo 3 con distribución ZIP. En el resto de modelos, la dispersión de los puntos, no permitió capturar adecuadamente la estructura espacio-temporal de los datos. Un resumen de las estimaciones a posteriori de todos los modelos evaluados se añade en el Apéndice C.

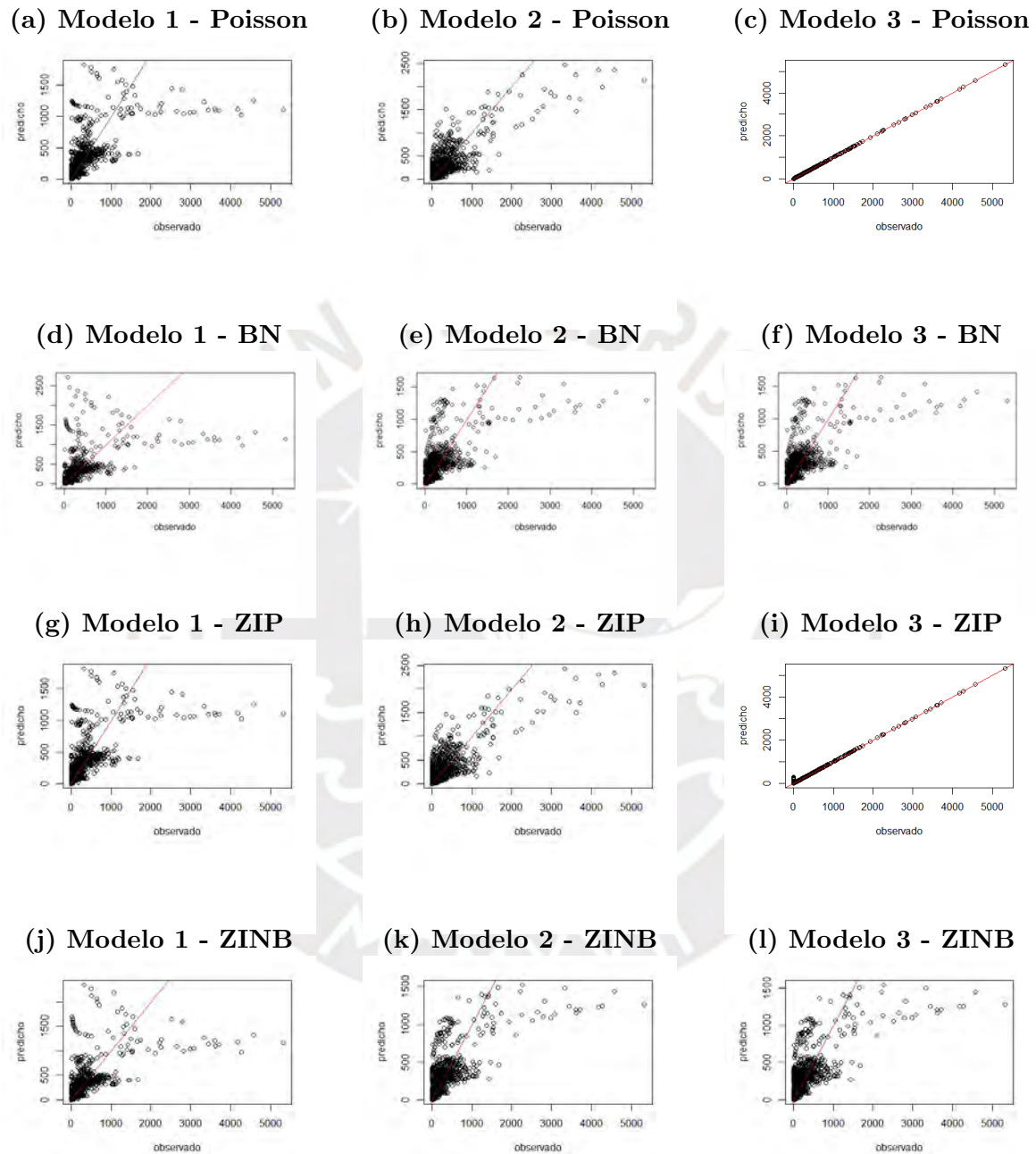


Figura 5.6: Valores estimados versus valores observados. Casos de dengue.

Para el Modelo 3 con distribución Poisson, el resumen de la estimación a posteriori del efecto fijo y los hiperparámetros se detallan en el Cuadro 5.2. El valor del efecto fijo se analizó



tras la transformación exponencial de los parámetros del predictor lineal. En ese sentido, el efecto fijo  $\alpha$  indicó que el número promedio de casos de dengue fue de  $\exp(3.259) = 26.02$ , con un intervalo de credibilidad de 9 a 75 casos, en ausencia de los otros efectos. Por otro lado, los hiperparámetros informaron que la precisión del componente temporal no estructurado alcanzó el mayor valor (3351.526); el alto valor de precisión sugiere la baja variabilidad temporal de este efecto ( $\tau_\phi$ ). A su vez, la interacción espacio-temporal ( $\tau_\delta = 0.602$ ) influyó considerablemente en la presencia de los casos dengue, siendo su estimación bastante precisa de acuerdo al intervalo de credibilidad (0.548 a 0.668).

Cuadro 5.2: Estimaciones a posteriori de los parámetros del Modelo 3 - Poisson.

Casos de dengue -Modelo 3 distribución Poisson					
Parámetros	media	sd	$Q_{0.025}$	$Q_{0.975}$	mediana
$\alpha$	3.259	0.537	2.199	4.316	3.259
$\tau_\nu$	0.154	0.038	0.097	0.246	0.148
$\tau_u$	2719.540	1871.891	770.719	7700.0	2223.005
$\tau_\gamma$	128.163	78.273	46.138	337.0	107.733
$\tau_\phi$	3351.526	6204.184	334.204	17300.0	1668.464
$\tau_\delta$	0.602	0.031	0.548	0.668	0.599

### 5.1.3. Aplicación de modelo - incidencia de dengue semanal

De los modelos ajustados en la sección 5.1.2, se seleccionó como mejor, al modelo con menor WAIC, es decir el Modelo 3 con distribución Poisson. Bajo esta distribución se ajustó nuevamente el modelo pero se consideró como variable respuesta la incidencia de dengue. Esta opción se utilizó para estimar la probabilidad de que una persona en la población desarrolle la enfermedad durante un período específico, siendo una medida más útil para comparar entre diferentes poblaciones.

Sea  $y_{it}$  el número de casos de dengue en el departamento  $i = 1, 2, \dots, 25$ , durante la semana epidemiológica  $t = 1, 2, \dots, 52$ . Se tuvo en cuenta, la distribución de Poisson para la variable respuesta  $y_{it} \sim \text{Poisson}(\lambda_{it} = E_{it}\theta_{it})$ , donde el *offset*  $E_{it}$  representa la población en riesgo en cada departamento  $i$  en el tiempo  $t$ , y  $\theta_{it}$  representa el riesgo de dengue en el área  $i$  y la semana epidemiológica  $t$ , tal que  $\log(\theta_{it}) = \alpha + u_i + \nu_i + \gamma_t + \phi_t + \delta_{it}$ . Se realizó la inferencia bayesiana utilizando INLA.

## Resultados

El uso del *offset* mejoró ligeramente el valor del criterio de información WAIC, el cual fue de 9053.45 (Apéndice C). En el Cuadro 5.3 se muestran las estimaciones a posteriori

del efecto fijo e hiperparámetros. Según el intercepto del Modelo 3, la tasa de incidencia de dengue fue igual a  $\exp(0.1741) = 1.190$ , cuando los efectos espaciales y temporales no estuvieron presentes. La precisión del efecto espacial no estructurado  $\tau_\nu = 0.198$  fue la más baja, indicando una alta variabilidad espacial de la incidencia de dengue explicada por el efecto aleatorio espacial. La precisión del efecto de interacción espacio-temporal  $\tau_\delta = 0.600$  fue baja, e inversamente proporcional a la varianza de la interacción, indicando que la existencia de una alta variabilidad en la interacción espacio-temporal, la que refleja variaciones específicas en la incidencia de dengue debido a combinaciones entre los departamentos y las semanas epidemiológicas. El componente temporal no estructurado presentó una alta precisión ( $\tau_\phi = 18700$ ), por lo que la variación temporal de la incidencia de dengue fue reducida entre las semanas epidemiológicas. En términos generales los hiperparámetros interacción ( $\tau_\delta$ ) y el componente espacial no estructurado ( $\tau_\nu$ ) observaron las menores incertidumbres en sus estimaciones.

Cuadro 5.3: Estimaciones posteriores de hiperparámetros en modelos espacio-temporales - Poisson con *offset*

<b>Incidencia de dengue - Modelo 3 Poisson con <i>offset</i></b>					
Parámetros	media	sd	$Q_{0.025}$	$Q_{0.975}$	mediana
$\alpha$	-1.741	0.482	-2.692	-0.792	-1.741
$\tau_\nu$	0.198	0.05	0.120	0.313	0.191
$\tau_u$	14.7	4.56	7.7195	25.55	14.1
$\tau_\gamma$	69.0	27.76	26.755	134	64.7
$\tau_\phi$	18700	15228.32	5760.133	59600	14200
$\tau_\delta$	0.600	0.03	0.545	0.661	0.600

La Figura 5.7 muestra las distribuciones marginales a posteriori. Se observa que los valores del intercepto, y las precisiones del componente interacción y del componente espacial estructurado presentan poca incertidumbre en su estimación. Mientras que las precisiones de los componentes temporales y del componente espacial no estructurado presentan mayor incertidumbre en su estimación.

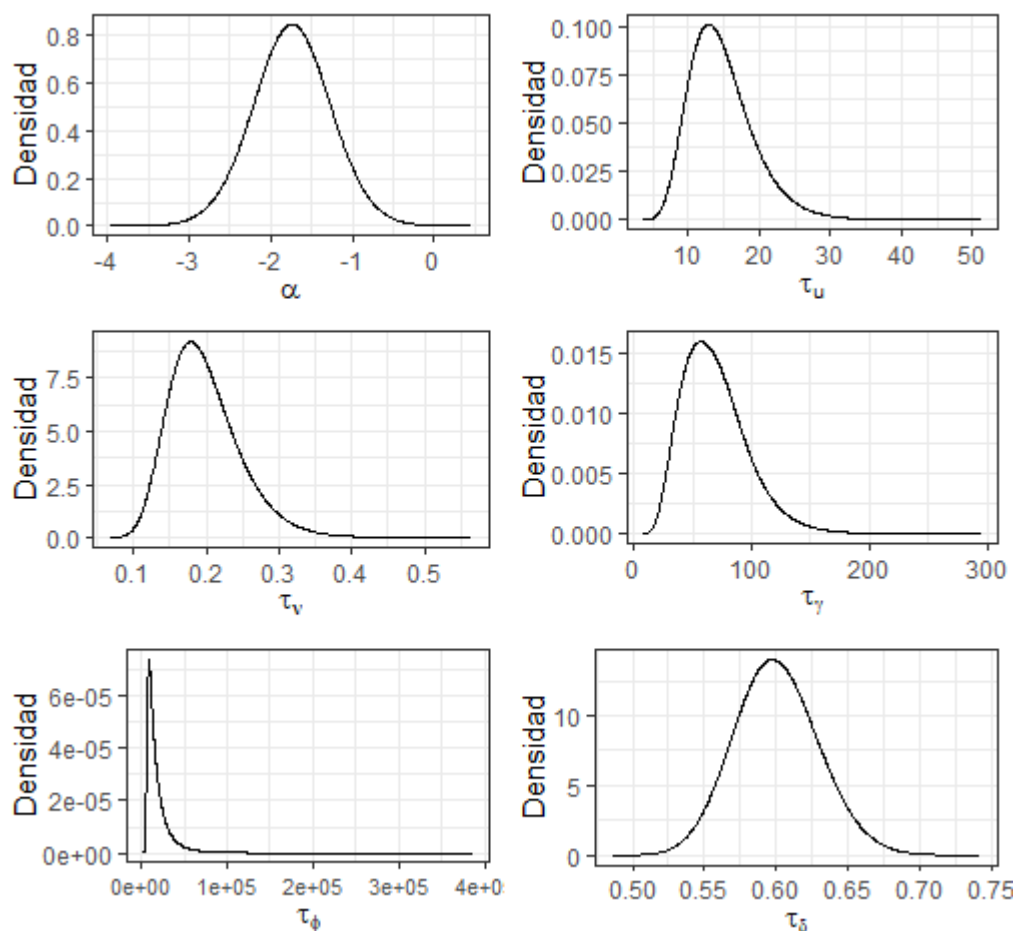


Figura 5.7: Modelo 3 Poisson con *offset*: Distribuciones posteriores

## 5.2. Aplicación 2: Dengue anual

Para la segunda aplicación de los modelos, los datos analizados correspondieron a los casos de dengue reportados anualmente en las 196 provincias señaladas en la Figura 5.8, durante el año 2010 al 2023.

### 5.2.1. Análisis exploratorio de los datos

La Figura 5.9 muestra año a año, los valores de la variable respuesta (casos de dengue) en las provincias del Perú. El incremento de casos registrados puede observarse a lo largo del periodo estudiado, lo cual se visualiza tanto al incrementarse la cantidad de localidades afectadas (i.e. incorporación de nuevas localidades, principalmente aquellas ubicadas al sur del país), como al incrementarse el número de casos por localidad (i.e. incremento en la



Figura 5.8: Mapa del Perú a nivel provincial.

intensidad de la coloración rojiza).

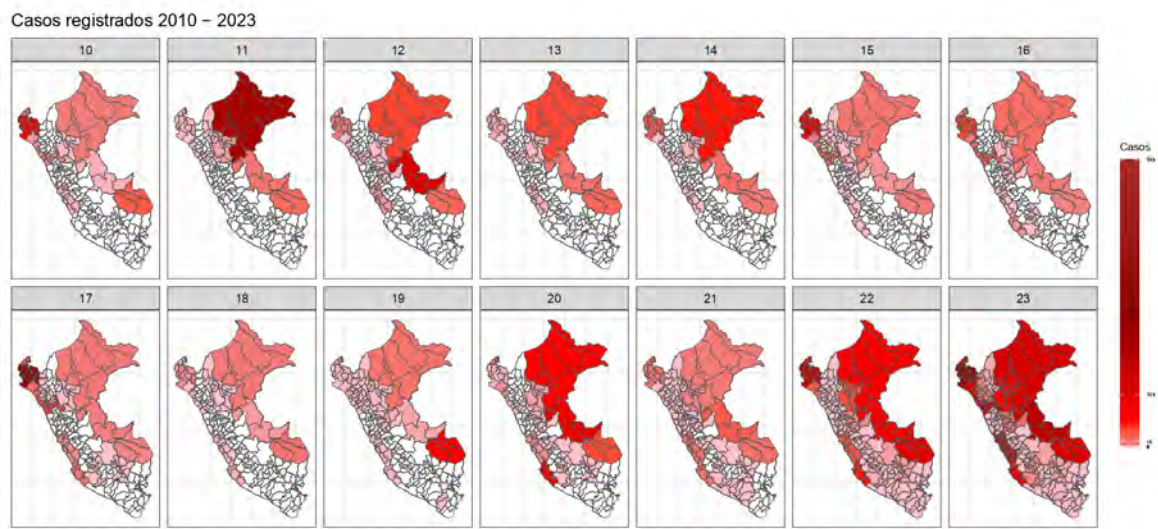


Figura 5.9: Casos reportados de dengue según provincias del año 2010 al 2023.

En ese aspecto, la Figura 5.10 permite visualizar la incorporación de provincias con casos, tanto confirmados como probables, a lo largo del periodo evaluado, de esta forma de las 196 provincias evaluadas, el 68.4 % (134) presentaron registros en el año 2023.

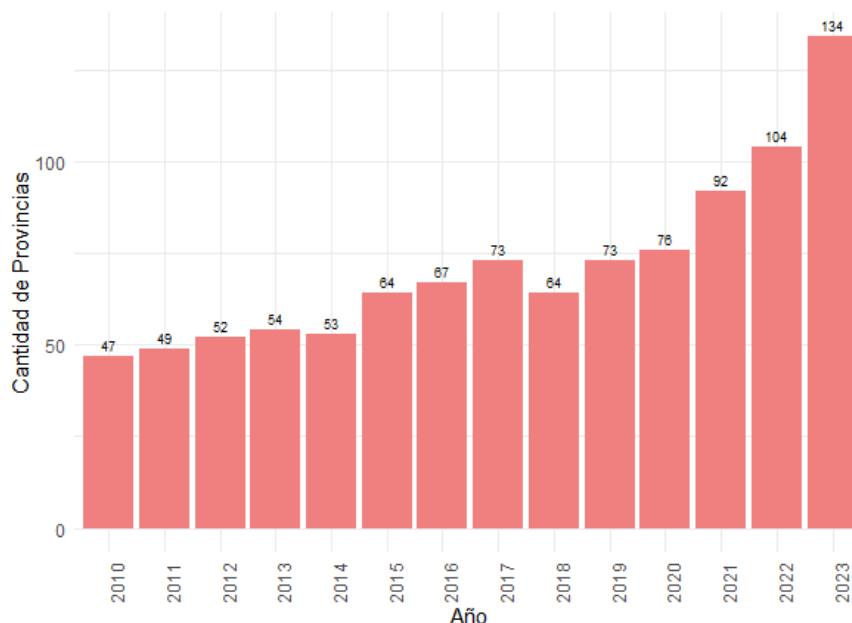


Figura 5.10: Cantidad de provincias que registraron casos de dengue del año 2010 al 2023.

### 5.2.2. Aplicación de modelos - casos de dengue anual

Sea  $y_{it}$  el número de casos de dengue en las provincias  $i = 1, 2, \dots, 196$ , y años  $t = 1, 2, \dots, 14$ . Se tuvo en cuenta, diferentes distribuciones de conteo para la variable respuesta  $y_{it}$ , entre ellas:

- Poisson:  $y_{it} \sim \text{Poisson}(\lambda_{it})$ ,
- Binomial Negativa (BN):  $y_{it} \sim \text{BN}(\lambda_{it}, \Gamma)$ ,
- Poisson Cero Inflacionada (ZIP):  $y_{it} \sim \text{ZIP}(\lambda_{it}, p_{it})$ ,
- Binomial Negativa Cero Inflacionada (ZINB):  $y_{it} \sim \text{ZINB}(\lambda_{it}, p_{it}, \Gamma)$ .

Según el análisis exploratorio, la distribución espacial anual de casos, estableció un número similar de casos en las provincias cercanas. En concordancia con lo señalado, se usaron los tres modelos espacio-temporales planteados para ajustar los datos, utilizando el paquete R-INLA ([www.r-inla.org](http://www.r-inla.org)) para cada distribución de  $y_{it}$ , y considerando en todos los casos que  $\lambda_{it} = \theta_{it}$ .

### Resultados

El ajuste de todos los modelos se midió mediante el criterio de WAIC. De acuerdo a



los valores señalados en el Cuadro 5.4, los modelos con mejor ajuste al evaluar la variable respuesta casos de dengue, fueron los que consideraron la distribución Binomial Negativa, correspondiendo los menores valores al Modelo 2 (15748.09) y al Modelo 3 (15790.30).

Cuadro 5.4: Criterios de información (WAIC) para los diferentes modelos y distribuciones considerando los casos de dengue.

Modelos	Poisson	BN	ZIP	ZINB
<b>Modelo 1</b>	502261.63	15869.94	430960.57	16450.22
<b>Modelo 2</b>	473170.79	<b>15748.09</b>	388626.86	16365.77
<b>Modelo 3</b>	743377.85	15790.30	641428.61	16363.86

En el Cuadro 5.5, se detallan los resultados de las estimaciones a posteriori correspondientes al Modelo 2 con distribución Binomial Negativa. El número promedio de casos de dengue fue de 26, con un intervalo de credibilidad de 7 a 92 casos. Las precisiones asociadas al componente espacial estructurado ( $\tau_u$ ) y el componente temporal no estructurado ( $\tau_\phi$ ) fueron las mayores, indicando que la varianza espacial explicada por el modelo es baja y que la variabilidad debido a la interacción entre las provincias y años, también es baja. Por otro lado, la precisión asociada al efecto temporal RW1 es baja ( $\tau_\gamma$ ), indicando que la varianza de los efectos aleatorios temporales con estructura RW1 es alta. La sobredispersión de los datos de acuerdo al modelo ajustado fue de 10.99 (i.e.  $1/\Gamma = 0.091$ ), este parámetro captura la variabilidad extra, e indica que hay una considerable variabilidad adicional de los datos, determinando que la media y la varianza no sean iguales. Resultados adicionales para los demás modelos se detallan en el Apéndice C.

Cuadro 5.5: Modelo 2 BN - Estimaciones a posteriori para los casos de dengue por provincia y año - Periodo 2010 a 2023.

Casos de dengue -Modelo 2 distribución BN					
Parámetros	media	sd	$Q_{0.025}$	$Q_{0.975}$	mediana
$\alpha$	3.241	0.649	1.963	4.517	3.241
$\tau_\nu$	0.100	0.027	0.059	0.164	0.096
$\tau_u$	1720.886	1003.049	401.168	4202.180	1507.618
$\tau_\gamma$	0.825	0.316	0.289	1.478	0.797
$\tau_\phi$	1515.216	1874.885	169.686	6337.742	957.293
$1/\Gamma$	0.091	0.003	0.084	0.097	0.090

### 5.2.3. Aplicación de modelos - incidencia de dengue anual

En esta sección se ajustaron los modelos de la sección 5.2.2, considerando como variable respuesta la incidencia de dengue, y teniendo en cuenta para las diferentes distribuciones de



conteo el *offset*  $E_{it}$  que representa la población en riesgo en cada provincia  $i$  en el año  $t$ ; y  $\theta_{it}$  que representa el riesgo de dengue en la provincia  $i$  y año  $t$ .

- Poisson:  $y_{it} \sim \text{Poisson}(\lambda_{it} = E_{it}\theta_{it})$ ,
- Binomial Negativa (BN):  $y_{it} \sim \text{BN}(\lambda_{it} = E_{it}\theta_{it}, \Gamma)$ ,
- Poisson Cero Inflacionada (ZIP):  $y_{it} \sim \text{ZIP}(\lambda_{it} = E_{it}\theta_{it}, p_{it})$ ,
- Binomial Negativa Cero Inflacionada (ZINB):  $y_{it} \sim \text{ZINB}(\lambda_{it} = E_{it}\theta_{it}, p_{it}, \Gamma)$ .

Asimismo, en todos los predictores lineales se agregó el término  $Z_{it}^\top \beta$ , para considerar el vector de covariables  $Z_{it}^\top$  y los coeficientes de regresión  $\beta$  respectivos. Y se asignó una a priori normal no informativa para los coeficientes de regresión. La inferencia bayesiana se realizó utilizando INLA.

### Resultados

Los valores WAIC (Cuadro 5.6), indicaron que cuando se utilizaron los datos de dengue ajustados por el tamaño de la población, el mejor ajuste correspondió al Modelo 2 con distribución Binomial Negativa (15711.77). La estimación para la incidencia de dengue se ajustó mejor que cuando solo se modeló la cantidad de casos de dengue.

Cuadro 5.6: Criterios de información (WAIC) para los diferentes modelos y distribuciones considerando la incidencia de dengue.

Modelos	Poisson	BN	ZIP	ZINB
<b>Modelo 1</b>	495694.41	15803.05	425820.97	16444.76
<b>Modelo 2</b>	468546.96	<b>15711.77</b>	383570.14	16351.27
<b>Modelo 3</b>	730028.79	15754.67	629693.77	16355.36

En el Cuadro 5.7 se muestran las estimaciones a posteriori para el Modelo 2 usando la distribución Binomial Negativa. Los resultados fueron muy similares a los obtenidos para el Modelo 2 ajustado a los casos de dengue sección 5.2.2. El valor del parámetro intercepto implicó una incidencia de dengue igual a 0.907 cuando los otros efectos aleatorios fueron iguales a cero. Entre los hiperparámetros, la precisión del componente temporal no estructurado ( $\tau_\phi$ ) tuvo el mayor valor, y la amplitud del intervalo de credibilidad indicó incertidumbre en su cálculo. La precisión asociada al efecto temporal RW1 fue una de las más bajas ( $\tau_\gamma$ ), indicando que la varianza de los efectos aleatorios temporales con estructura RW1 fue alta. La precisión asociada al componente espacial estructurado ( $\tau_u$ ) fue alta, por lo que la varianza

espacial explicada por el modelo fue baja. Los resultados a posteriori para todos los modelos se encuentran en el Apéndice C.

Cuadro 5.7: Modelo 2 BN con *offset* - Estimaciones a posteriori para la incidencia de dengue por localidad y año - Periodo 2010 a 2023.

Incidencia de dengue -Modelo 2 distribución BN con <i>offset</i>					
Parámetros	media	sd	$Q_{0.025}$	$Q_{0.975}$	mediana
$\alpha$	-0.098	0.627	-1.337	1.138	-0.097
$\tau_\nu$	0.108	0.030	0.058	0.174	0.105
$\tau_u$	1250.003	4993.768	22.245	8254.009	316.137
$\tau_\gamma$	0.887	0.374	0.375	1.820	0.815
$\tau_\phi$	2862.001	2612.595	453.823	9789.264	2113.602
$1/\Gamma$	0.093	0.003	0.085	0.099	0.093

Al calcular los valores estimados por los modelos ajustados y compararlos con los valores observados (Figura 5.11), se observa en el Modelo 1 que los valores estimados tienden a ser menores que los valores observados, especialmente cuando los valores observados son altos. Mientras en los Modelos 2 y 3, se observa un mejor ajuste, si bien aún se presentan valores estimados alejados de la línea los cual indica errores en las estimaciones.

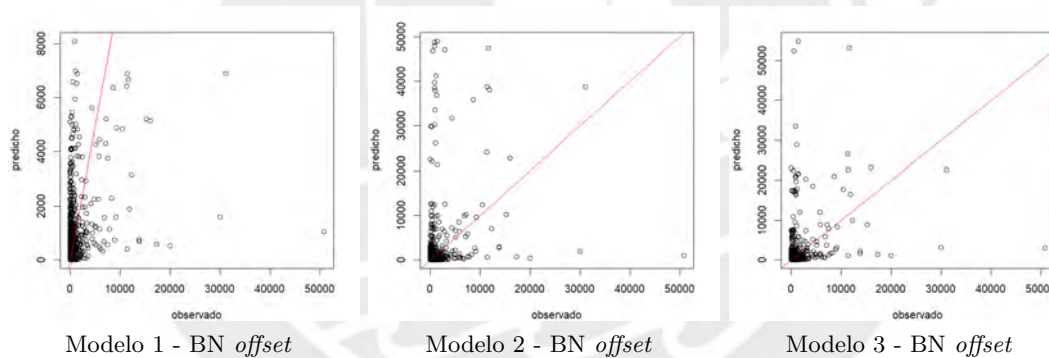


Figura 5.11: Valores estimados versus observados en Modelos sin covariables climáticas.

#### 5.2.4. Aplicación de Modelos con covariables climáticas - incidencia de dengue anual

Tras el análisis de los datos anuales de incidencia de dengue por provincia, los menores valores de WAIC determinaron que los modelos que consideraron una distribución Binomial Negativa fueron los que describieron mejor el comportamiento de la variable respuesta. Se ajustaron nuevamente estos modelos, incluyendo en el predictor lineal como covariables las variables climáticas: Temperatura máxima ( $T_{\max}$ ), Precipitación total ( $P_{\text{total}}$ ) y Humedad Relativa promedio ( $HR_{\text{mean}}$ ), estas covariables son factores que están fuertemente correla-

cionados con la transmisión de esta enfermedad (Tsheten et al., 2020).

En el Cuadro 5.8 se comparan los criterios de información WAIC para los modelos ajustados bajo una distribución Binomial Negativa sin covariables (modelos ajustados en la sección anterior) y los modelos con covariables. Siendo el Modelo 3 con distribución Binomial Negativa y con covariables el que obtuvo el mejor ajuste.

Cuadro 5.8: Criterios de información (WAIC) - Binomial Negativa con *offset*.

Modelo	WAIC sin variables climáticas	WAIC con variables climáticas
Modelo 1	15803.05	15399.95
Modelo 2	15711.77	15320.28
Modelo 3	15754.67	<b>15284.41</b>

A continuación se detallan los resultados obtenidos para el Modelo 3 BN evaluado para la tasa de incidencia del dengue, utilizando covariables climáticas. Los valores resumidos en el Cuadro 5.9 muestran las estimaciones a posteriori para el modelo seleccionado. Los resultados indicaron que las covariables climáticas (Temperatura máxima, Precipitación total y Humedad relativa promedio) tuvieron un efecto positivo significativo sobre la tasa de incidencia del dengue. En ese sentido, cuando la  $P_{total}$  aumentó en una unidad, la incidencia de dengue aumentó en 8.9%, cuando la  $T_{max}$  aumentó en un grado, la incidencia de dengue aumentó en 22.4% y cuando la  $HR_{mean}$  aumentó en una unidad, la incidencia de dengue aumentó en 33.6%.

Cuadro 5.9: Modelo 3 BN con *offset* y covariables climáticas - Estimaciones a posteriori para la incidencia de dengue por provincia y año - Periodo 2010 a 2023.

Incidencia de dengue - Modelo 3 distribución BN con <i>offset</i> y covariables climáticas					
Parámetros	media	sd	$Q_{0.025}$	$Q_{0.975}$	mediana
$\alpha$	-13.691	1.913	-17.439	-13.692	-9.935
$P_{total}$	0.086	0.015	0.057	0.115	0.086
$T_{max}$	0.202	0.072	0.061	0.344	0.202
$HR_{mean}$	0.290	0.094	0.105	0.475	0.290
$\tau_\nu$	0.352	0.102	0.203	0.600	0.335
$\tau_u$	13.435	5.281	5.640	26.099	12.586
$\tau_\gamma$	1.364	0.443	0.645	2.367	1.312
$\tau_\phi$	278.497	114.880	121.973	565.764	256.325
$\tau_\delta$	0.649	0.116	0.467	0.919	0.634
$1/\Gamma$	0.147	0.006	0.136	0.158	0.147

A diferencia de los modelos ajustados previamente, la varianza de los efectos aleatorios espaciales del Modelo 3 con covariables fue mayor. Esto implica que en este modelo se modeló mejor la distribución espacial de los datos. En este sentido, la alta varianza indicó que las

diferencias entre las áreas, en términos de la incidencia del dengue, fueron capturadas adecuadamente por el modelo. De otro modo, los parámetros de precisión asociados a los efectos temporales, establecieron que la mayor parte de la variabilidad temporal capturada por el modelo, fue estructurada (i.e. tendencias temporales) y no aleatoria. Estos resultados sugieren que, aunque las covariables climáticas son importantes, hay otros factores espacio-temporales que también influyen en las tasas de la enfermedad.

La Figura 5.12 muestra las distribuciones marginales a posteriori del Modelo 3 con *offset*, distribución BN y covariables climáticas. Al evaluar las distribuciones posteriores, se observa que tanto el intercepto como las covariables presentaron poca incertidumbre y distribuciones más simétricas, en particular, la covariable Precipitación total. Contrariamente, la precisión del efecto temporal no estructurado y de los efectos espaciales, observaron la mayor asimetría, e incertidumbre dentro del modelo.

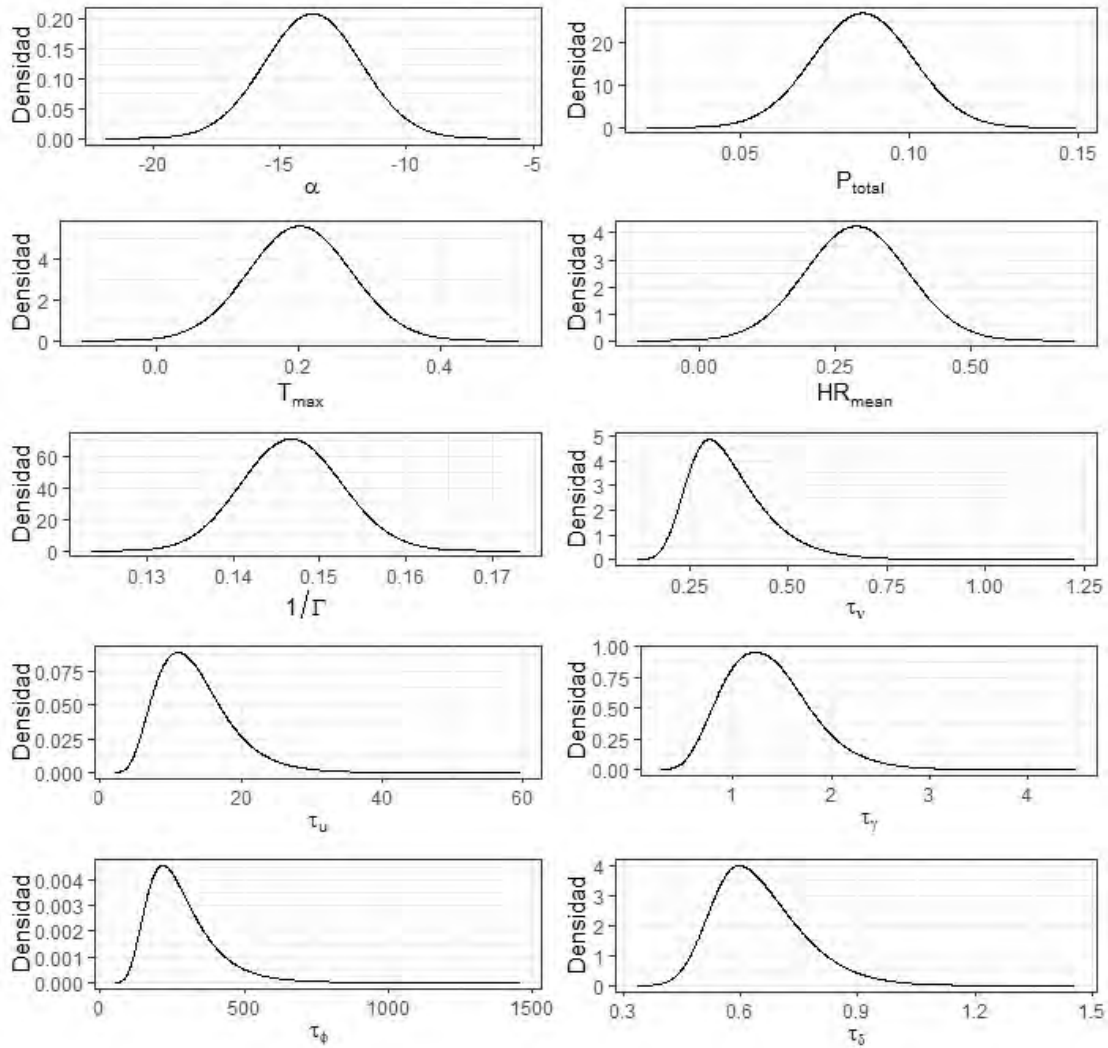


Figura 5.12: Distribuciones marginales a posteriori del Modelo 3 BN con *offset* y covariables climáticas.

La Figura 5.13 muestra la media a posteriori del efecto aleatorio temporal estructurado  $\gamma_t$  a lo largo de los 14 años de evaluación. Primero se observa que estos efectos aleatorios son diferentes de cero, por lo tanto existe un patrón temporal en los datos que debe tomarse en cuenta. Además se observa en los últimos años, el aumento sostenido de la incidencia de dengue.

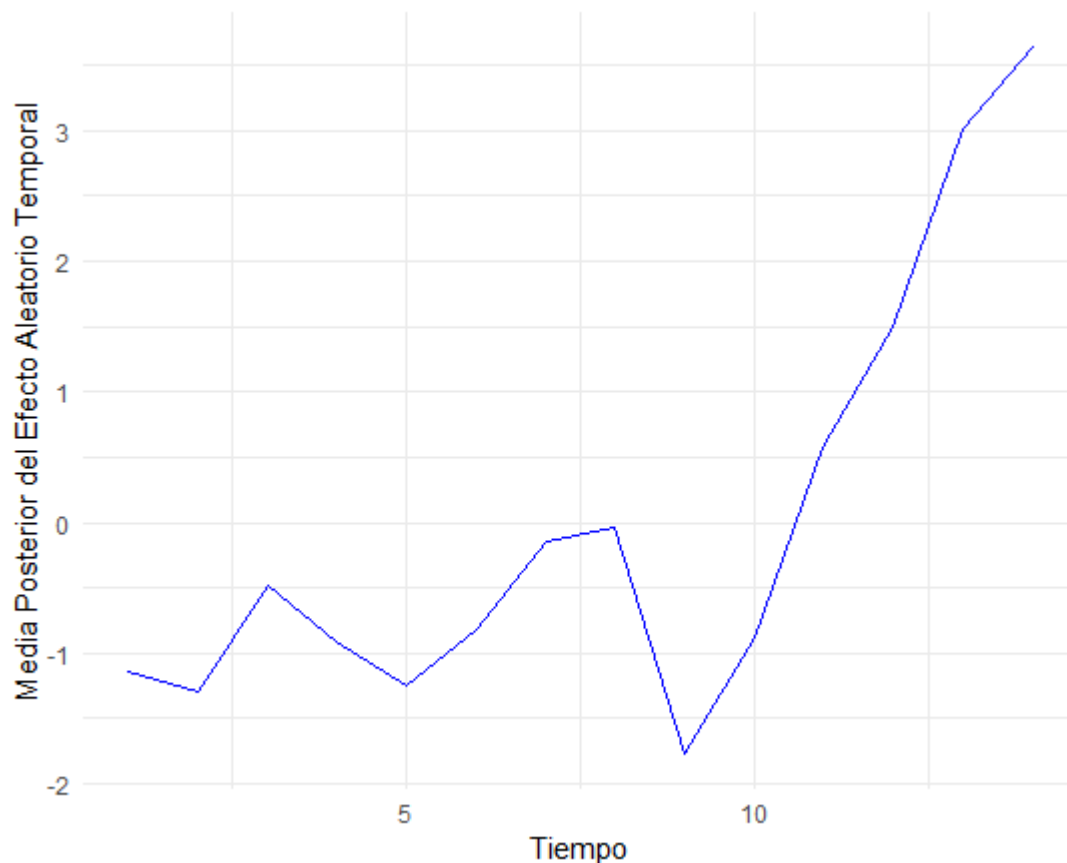


Figura 5.13: Media a posteriori de los efectos temporales estructurados  $\gamma_t$  usando el Modelo 3 BN con *offset*, distribución BN y covariables climáticas.

La estructura del modelo con mejor ajuste (Modelo 3 con *offset*, distribución BN y covariables climáticas), permitió evaluar la incidencia del dengue ajustada por los efectos aleatorios espaciales  $u_i + \nu_i$  (Figura 5.14). Observándose que la mayoría de las provincias tuvieron una incidencia de dengue similar al promedio ajustado (color amarillo a verde claro), mientras que otras provincias como Condorcanqui (Amazonas), Huarmey (Ancash) y Santa (Ancash), presentaron una incidencia ajustada menor al del promedio comparado (azul). En otras provincias se observaron una incidencia ajustada mayor que el promedio, por lo que el riesgo de dengue en estos puntos fue mayor al esperado por el modelo, tales como Luya (Amazonas) cuyos primeros registros fueron en el año 2023, así como Bolognesi (Ancash) y Ocros (Ancash) donde hay registros desde el año 2021.



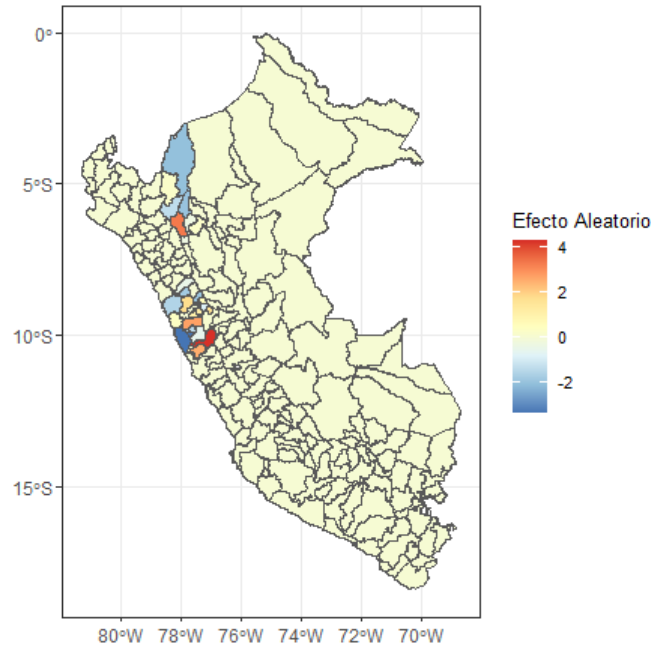


Figura 5.14: Media a posteriori del efecto aleatorio estructurados (BYM) usando el Modelo 3 BN con *offset*, distribución BN y covariables climáticas.

La Figura 5.15 muestra las medias a posteriori del efecto de la interacción de los efectos temporales y espaciales  $\delta_{it}$ . Se observaron provincias con un aumento notable en la incidencia de dengue en comparación con el promedio ajustado, tales como Maynas (Loreto), Putumayo (Loreto), Lamas (San Martín), Bellavista (San Martín) y Piura (Piura).

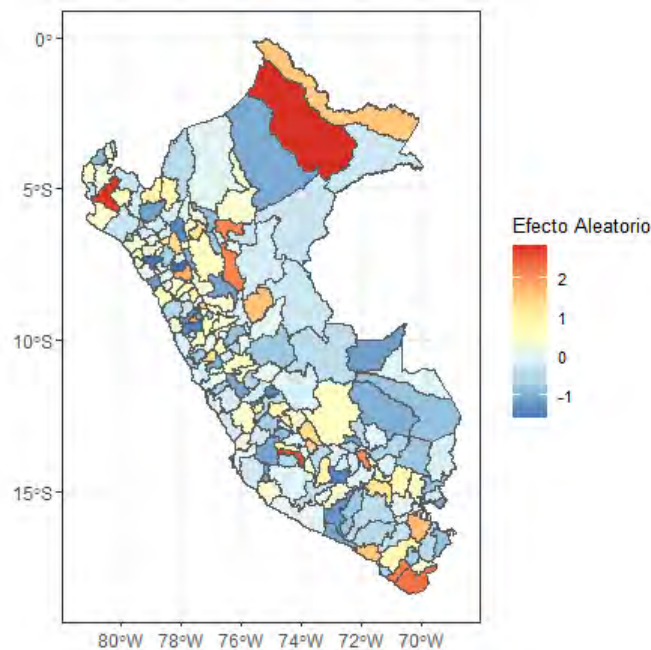


Figura 5.15: Media a posteriori del efecto de la interacción (idarea.idtime) usando el Modelo 3 BN con *offset*, distribución BN y covariables climáticas.



La Figura 5.16 compara los valores originales de casos de dengue respecto a la estimación de los casos de dengue, para los Modelos 1, 2 y 3, ajustados usando la distribución Binomial Negativa y las covariables climáticas. Al incluir las variables climáticas en los modelos, se logró capturar mejor la variabilidad de los datos observados, como se refleja en la cercanía de los puntos a la línea de referencia (roja). No obstante, se redujo la dispersión de los puntos, aún se observan sobreestimaciones en especial cuando los valores observados fueron muy bajos.

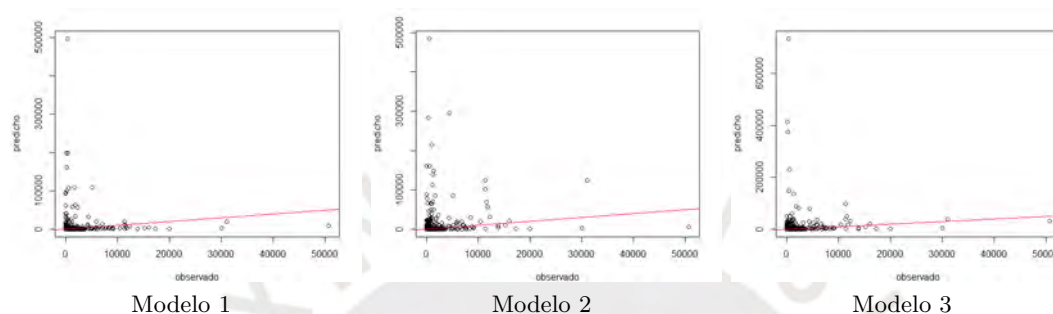


Figura 5.16: Valores estimados versus observados en modelos espacio-temporales ajustados usando la distribución Binomial Negativa y covariables climáticas.

## Capítulo 6

# Conclusiones

El objetivo principal de esta tesis fue aplicar modelos espacio-temporales para estudiar la distribución espacial y temporal de los casos de dengue en el Perú, a través de inferencia bayesiana e INLA. Se ajustaron tres modelos: un modelo clásico paramétrico; un modelo dinámico; y un modelo con interacción. Los modelos fueron ajustados para diferentes agrupaciones de datos: semanas epidemiológicas por departamento y anuales por provincia; considerándose a su vez diferentes distribuciones y características de variables de conteo, como: Poisson, Binomial Negativa, Poisson Cero Inflacionada y Binomial Negativa Cero Inflacionada.

En la primera aplicación, se estudió la incidencia de dengue en un conjunto de 25 locaciones a través de 52 semanas epidemiológicas. La segunda aplicación estudió la incidencia de dengue en un conjunto mayor de datos, que incluyó 196 provincias a través de 14 años, considerando además la inclusión de covariables climáticas, las cuales juegan un papel importante en la propagación del dengue (Ebi y Nealon, 2016). En ambos casos, el Modelo con más efectos aleatorios (espaciales, temporales e interacción) fue el que ofreció estimaciones más precisas.

En el caso de la primera aplicación (a nivel departamental), el modelo con distribución Poisson fue el de mejor ajuste; mientras en el caso de la segunda aplicación cuya agrupación de datos incluyó una resolución espacial más detallada (a nivel provincial), el modelo con distribución Binomial Negativa fue el de mejor ajuste. Al dividir el área en 196 locaciones, el modelo pudo capturar una mayor variabilidad espacial, lo que permitió revelar patrones locales y heterogeneidad que no fueron considerados cuando se utilizó una división departamental. Debe considerarse asimismo la sobredispersión de la data, la cual fue muy superior

al considerar las 196 provincias en comparación con los 25 departamentos (i.e. 9.50 vs. 1.20). La estimación de las tasas de incidencia de dengue a través del Modelo 3 del mismo modo indicó, un efecto temporal no estructurado más fuerte que el efecto espacial.

Los resultados mostraron, además, en el análisis semanal que existe una estructura de autocorrelación en el tiempo en la distribución espacial de los casos de dengue. De otra forma en el análisis anual, se puede observar a lo largo del periodo analizado, un patrón claro de propagación a nuevas áreas, al incrementarse la cantidad de locaciones afectadas por el dengue; y el aumento en la cantidad de los casos de dengue, que se visualiza en la intensidad de la coloración. Además, los resultados también mostraron que la inclusión de las variables climáticas  $T_{\max}$ ,  $P_{\text{total}}$  y  $HR_{\text{mean}}$ , mejoraron el ajuste del modelo, lo cual es coincidente con lo indicado por Tsheten et al. (2020) quienes refieren que la transmisión de esta enfermedad está fuertemente correlacionada con las fluctuaciones en variables como la precipitación, la temperatura y la humedad relativa.

Entender la modelización adecuada de los datos de dengue es relevante porque proporciona información esencial para la planificación sanitaria estratégica, que permite a las autoridades de salud comprender la dinámica de la transmisión de la enfermedad, y aplicar estrategias de control más efectivas. Como trabajos futuros, teniendo en cuenta que la implementación de la metodología INLA dentro del enfoque bayesiano, proporciona estimaciones confiables y no es costosa en términos computacionales, puede extenderse el uso de estos modelos espacio temporales a datos distritales y semanales, que permitan captar mejor la heterogeneidad local; evaluar otros tipos de interacciones, así como estudiar qué otras covariables o factores puedan explicar adecuadamente el comportamiento de los patrones observados en la distribución espacial y temporal de la enfermedad.

## Apéndice A

# Análisis exploratorio de casos de dengue por semanas

Entre las semanas 1 a la 13, se observa un incremento en el total de casos en las semanas 9 a la 13, siendo el departamento de Piura el que alcanzó los mayores registros (i.e. 2784, 2520, 1545, 1370 y 1162 casos respectivamente), mientras Ucayali obtuvo los mayores registros durante las semanas 8, 7 y 6 (i.e. 1083, 909 y 838) (Figura A.1).

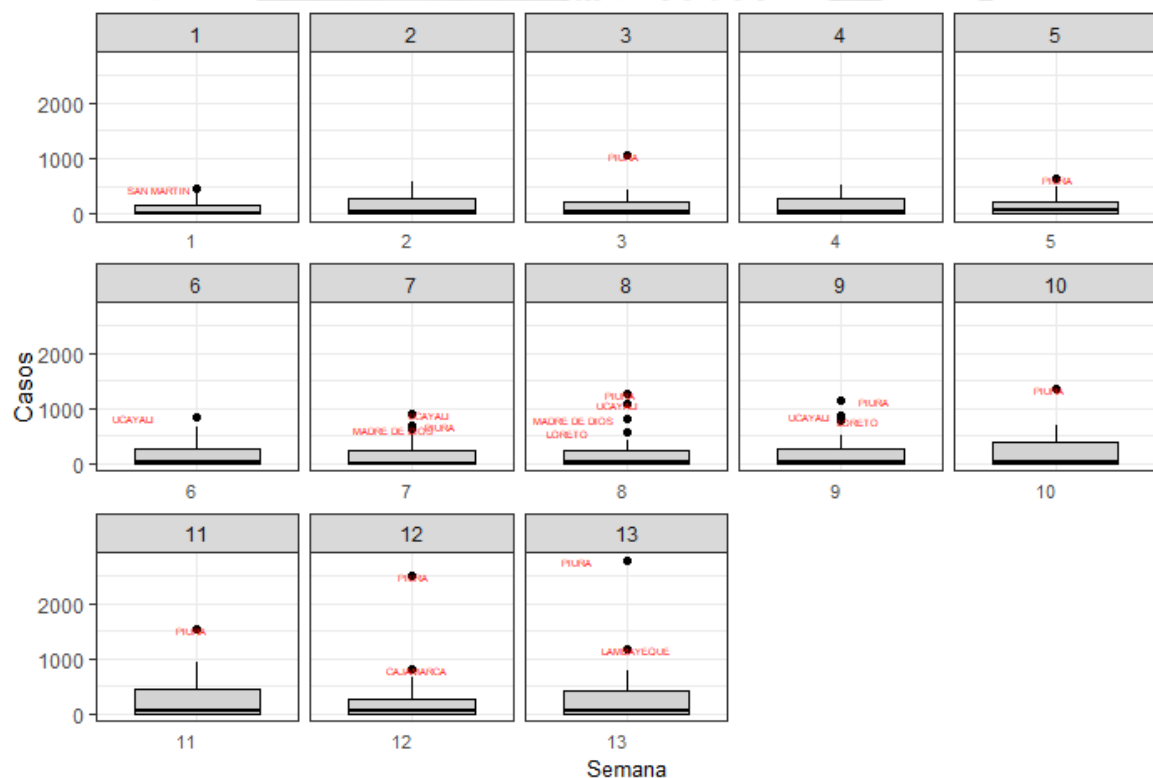


Figura A.1: Boxplot de Casos de dengue - Semanas 1 a 13, año 2023

Siguiendo la tendencia de la semana 13, Piura fue el departamento con mayor cantidad de registros entre las semanas 14 a la 16; seguida del departamento de Lima; este último departamento obtuvo los mayores registros entre las semanas 20 a la 26. En la semana 21, Lima reportó la mayor cantidad de casos del año 2023, con un total de 5318. Durante este periodo se observaron los mayores registros del año 2023 (Figura A.2).

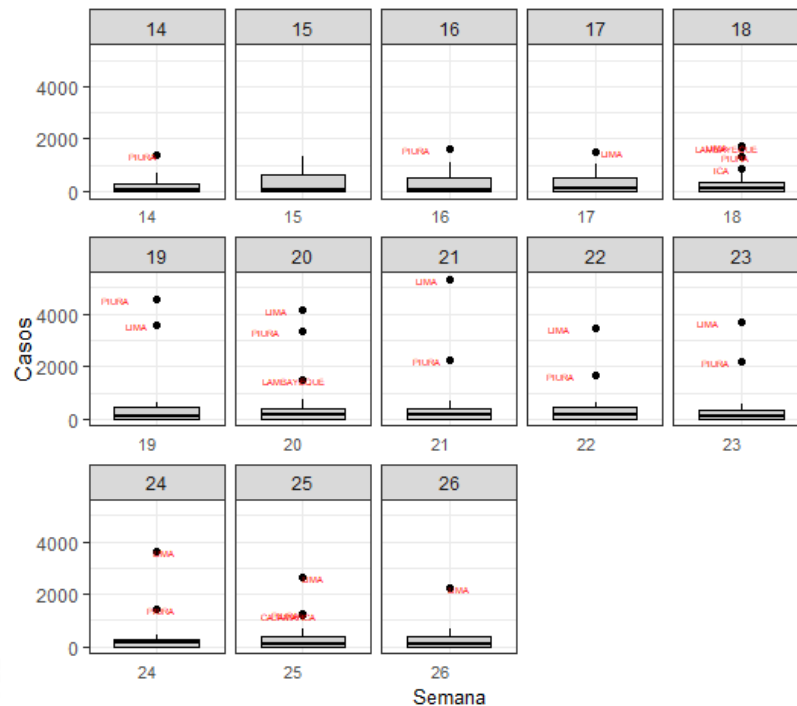


Figura A.2: Boxplot de Casos de dengue - Semanas 14 a 26, año 2023

Entre las semanas 27 a la 36 el departamento de Lima alcanzó la mayor cantidad de registros. Mientras que el departamento de La Libertad obtuvo los mayores valores entre la semana 37 y 39. Piura, alcanzó su mayor registro del año, 4276 casos, durante la semana 29.

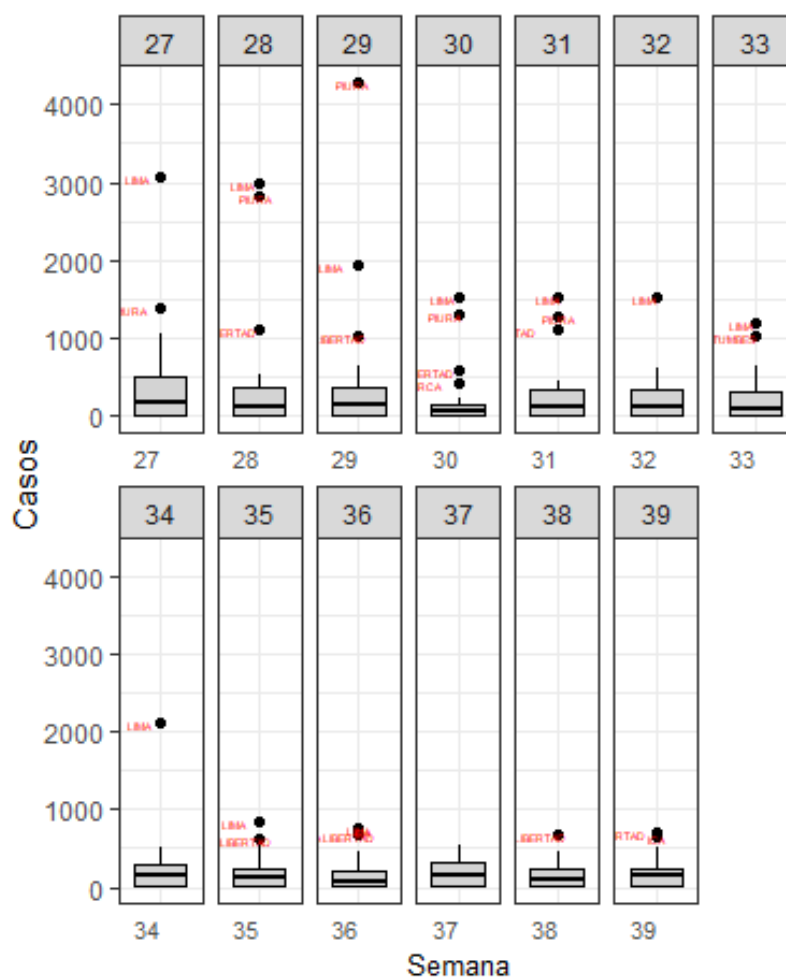


Figura A.3: Boxplot de Casos de dengue - Semanas 27 a 39, año 2023

Entre las semanas 40 y 45 los departamentos de la costa norte (Tumbes y Piura) alcanzaron los mayores registros de casos, en ese sentido el departamento de Tumbes alcanzó su mayor registro anual en la semana 41 (1428 casos). Luego entre la semana 48 y 51 fue San Martín el departamento con mayor cantidad de registros. En la semana 52, nuevamente se observa un repunte de casos en Tumbes, registrándose un total de 1116 casos.



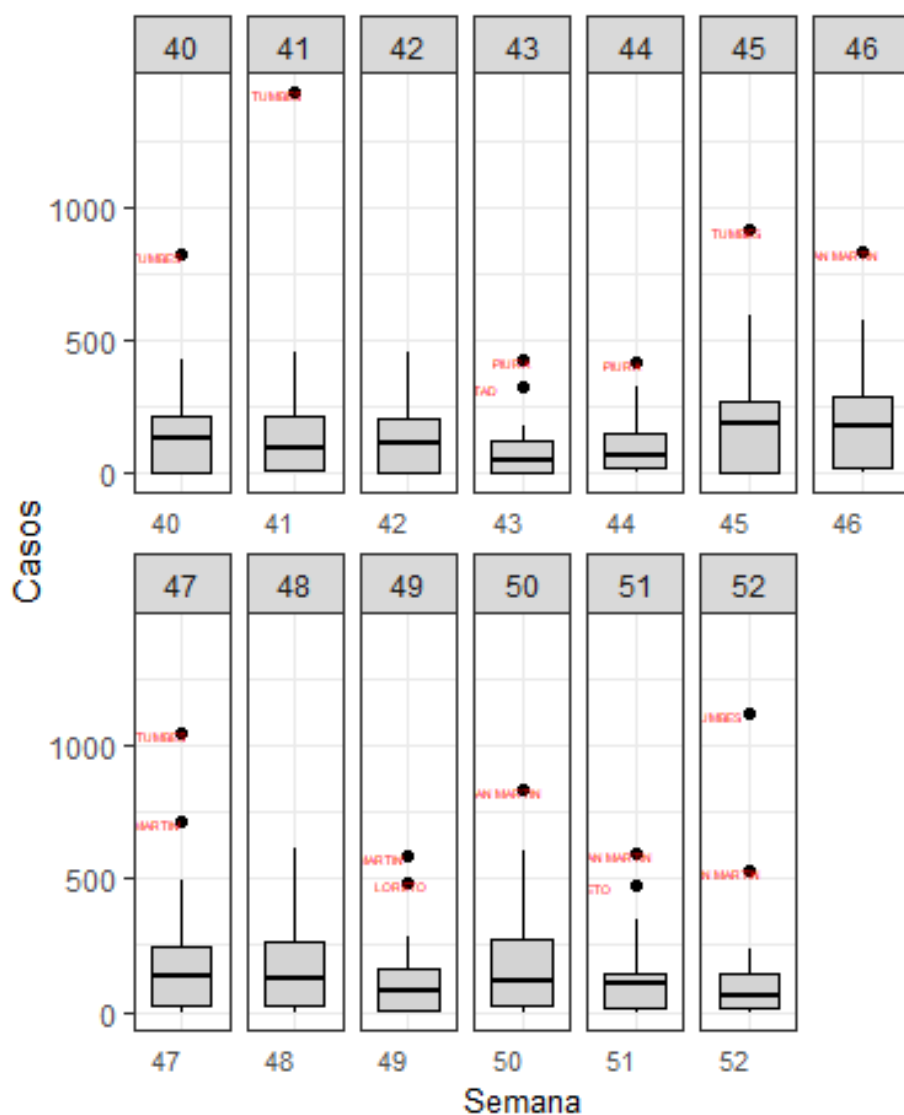


Figura A.4: Boxplot de Casos de dengue - Semanas 40 a 52, año 2023

## Apéndice B

# Distribución de la variable respuesta

Al observar la distribución de la variable casos de dengue (Figura B.1), se observa la típica distribución de una variable de conteo.

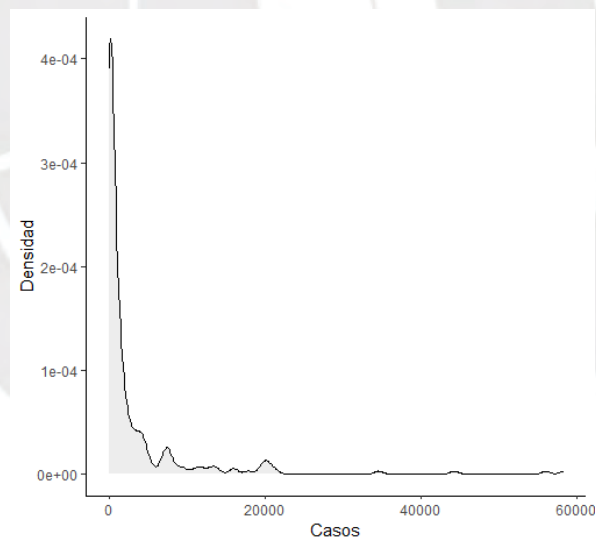


Figura B.1: Distribución de la variable casos de dengue.

El porcentaje de ceros para la aplicación de datos por semanas epidemiológicas por departamento fue del 18.3 %, mientras que para la aplicación de datos de dengue por años en provincias fue de 63.48 %.

## Apéndice C

# Resultados adicionales en la aplicación

### C.1. Aplicación 1: Casos de dengue por semana

Las estimaciones a posteriori para el Modelo 1 se incluyen en el Cuadro C.1.

Cuadro C.1: Estimaciones a posteriori Modelo 1 - Casos año 2023 por semana

Modelo 1				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	3.613	3.449	3.969	3.825
$\beta$	0.009	0.014	0.004	0.005
$\tau_u$	0.423	0.803	2211.930	2909.663
$\tau_v$	0.331	0.141	0.172	0.144
$\tau_\delta$	1042.492	526.716	1337.737	764.015
$1/\Gamma$	-	1.067	-	1.348
p	-	-	0.182	0.182
WAIC	58817.13	12945.35	57083.53	13417.36

Las estimaciones a posteriori para el Modelo 2 se incluyen en el Cuadro C.2.

Cuadro C.2: Estimaciones a posteriori Modelo 2 - Casos año 2023 por semana

Modelo 2				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	3.874	3.947	4.066	4.072
$\tau_u$	0.097	0.126	0.220	0.083
$\tau_v$	7.878	2.01	1.955	411
$\tau_\gamma$	29.788	138	31.758	176
$\tau_\phi$	236.422	2380	226.040	2150
$1/\Gamma$	-	0.834	-	1.16
p	-	-	0.182	0.182
WAIC	84354.30	13209.86	80058.44	13570.70

Las estimaciones a posteriori para el Modelo 3 se incluyen en el Cuadro C.3.

Cuadro C.3: Estimaciones a posteriori Modelo 3 - Casos año 2023 por semana

Modelo 3				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	3.259	3.948	3.619	4.065
$\tau_u$	2719.540	0.072	0.170	0.099
$\tau_\nu$	0.154	2710	1.00	7.60
$\tau_\gamma$	128.163	142	140	186
$\tau_\phi$	3351.526	25500	31400	20400
$\tau_{\delta_{it}}$	0.602	245006	0.849	22500
$1/\Gamma$	-	0.833	-	1.16
p	-	-	0.183	0.183
WAIC	9053.57	13209.80	9678.77	13566.77

## C.2. Aplicación 1: Incidencia de dengue por semana

Cuando se analizaron los datos de incidencia de dengue, el mejor ajuste de acuerdo al criterio WAIC correspondió al Modelo 3 con distribución Poisson (Cuadro 5.1).

Las estimaciones a posteriori para el Modelo 1 se incluyen en el Cuadro C.4.

Cuadro C.4: Estimaciones a posteriori Modelo 1 - Incidencia año 2023 por semana

Modelo 1				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	-1.391	-1.552	-1.038	-1.195
$\beta$	0.009	0.014	0.004	0.006
$\tau_u$	0.057	2264.11	1.758	0.064
$\tau_\nu$	2192.926	0.13	0.233	3479.674
$\tau_\delta$	1042.886	525.22	1348.657	823.115
$1/\Gamma$	-	1.07	-	0.184
p	-	-	0.182	1.358
WAIC	58816.06	12945.81	57083.88	13418.49

Las estimaciones a posteriori para el Modelo 2 se incluyen en el Cuadro C.5.

Cuadro C.5: Estimaciones a posteriori Modelo 2 - Incidencia año 2023 por semana

Modelo 2				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	-1.128	-1.055	-0.938	-0.31
$\tau_u$	0.092	0.091	0.108	8.630
$\tau_\nu$	2224.328	2280	2187.673	16.718
$\tau_\gamma$	29.473	140	31.138	52.965
$\tau_\phi$	278.844	22300	200.586	59.759
$1/\Gamma$	-	0.834	-	0.001
p	-	-	0.182	0.204
WAIC	84353.76	13209.76	80052.17	15799.38

Las estimaciones a posteriori para el Modelo 3 se incluyen en el Cuadro C.6.

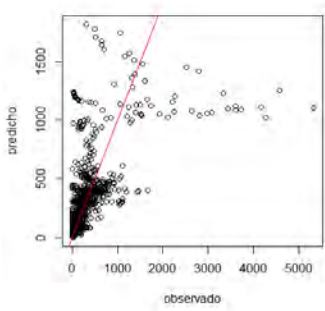
Cuadro C.6: Estimaciones a posteriori Modelo 3 - Incidencia año 2023 por semana

Modelo 3				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	-1.741	-1.054	-1.392	-0.938
$\tau_u$	14.7	0.09	2620	3200
$\tau_\nu$	0.198	2910	0.215	0.224
$\tau_\gamma$	69	138	136	182
$\tau_\phi$	18700	24800	30500	23100
$\tau_{\delta_{it}}$	0.600	27800	0.850	34400
$1/\Gamma$	-	0.833	-	1.16
p	-	-	0.183	0.183
WAIC	9053.45	13210.06	9677.97	13569.84

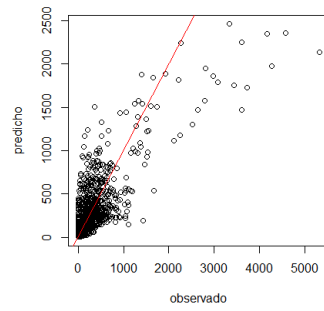
La comparación entre los valores observados y los valores estimados por los diferentes modelos y distribuciones (Figura C.1), también estableció una mejor predicción para el Modelo 3 con distribución Poisson.



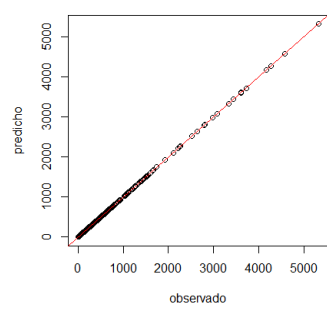
(a) Modelo 1 - Poisson



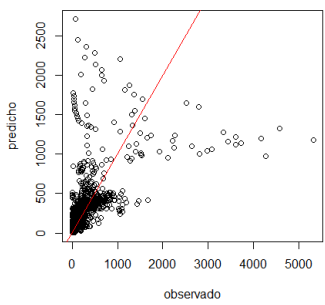
(b) Modelo 2 - Poisson



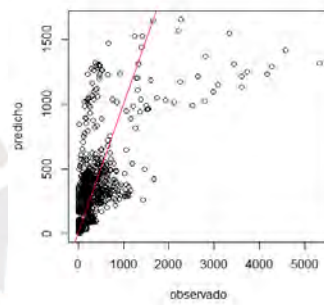
(c) Modelo 3 - Poisson



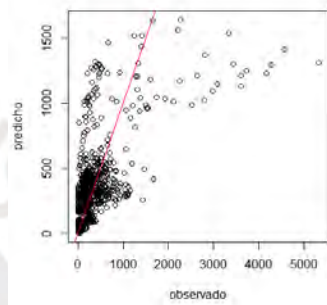
(d) Modelo 1 - BN



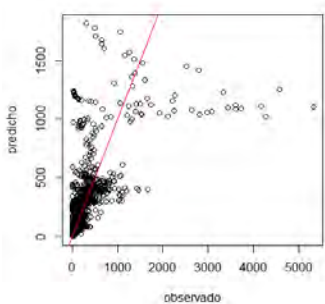
(e) Modelo 2 - BN



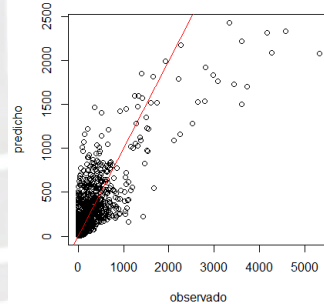
(f) Modelo 3 - BN



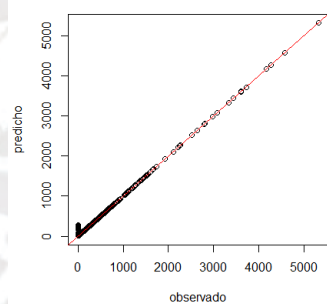
(g) Modelo 1 - ZIP



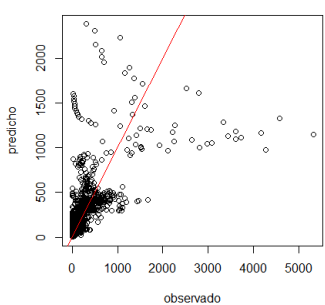
(h) Modelo 2 - ZIP



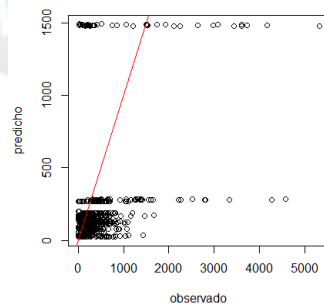
(i) Modelo 3 - ZIP



(j) Modelo 1 - ZINB



(k) Modelo 2 - ZINB



(l) Modelo 3 - ZINB

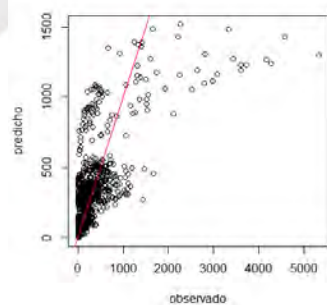


Figura C.1: Valores estimados versus valores observados. Incidencia de dengue.



### C.3. Aplicación 2: Casos de dengue por año, Periodo 2010 a 2023

Las estimaciones a posteriori para el Modelo 1 se incluyen en el Cuadro C.7.

Cuadro C.7: Estimaciones a posteriori Modelo 1 - Casos durante años 2010 a 2023

Modelo 1				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	-3.843	1.058	-1.748	3.568
$\beta$	0.690	0.319	0.573	0.159
$\tau_u$	2217.25	604.751	84.447	0.144
$\tau_\nu$	0.01	719.361	0.011	2261.497
$\tau_\delta$	2.20	32.069	2.327	2032.624
$1/\Gamma$	-	0.068	-	0.139
p	-	-	0.573	0.634
WAIC	502261.63	15869.94	430960.57	16450.22

Las estimaciones a posteriori para el Modelo 2 se incluyen en el Cuadro C.8.

Cuadro C.8: Estimaciones a posteriori Modelo 2 - Casos durante años 2010 a 2023

Modelo 2				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	3.683	3.241	4.69	4.738
$\tau_u$	2188.401	1720.866	3010	691
$\tau_\nu$	0.138	0.100	0.225	0.393
$\tau_\gamma$	942.115	0.825	30400	1.32
$\tau_\phi$	4.272	1515.216	1.44	46300
$1/\Gamma$	-	0.091	-	0.174
p	-	-	0.635	0.635
WAIC	473170.79	15748.09	388626.86	16365.77

Las estimaciones a posteriori para el Modelo 3 se incluyen en el Cuadro C.9.

Cuadro C.9: Estimaciones a posteriori Modelo 3 - Casos durante años 2010 a 2023

Modelo 3				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	1.888	2.938	4.115	4.638
$\tau_u$	2500	1791.666	3830	2300
$\tau_\nu$	0.087	0.089	0.323	0.568
$\tau_\gamma$	1.32	1.103	1.49	48600
$\tau_\phi$	22400	143.548	58500	2.57
$\tau_{\delta_{it}}$	0.380	0.945	0.600	4.62
$1/\Gamma$	-	0.102	-	0.195
p	-	-	0.631	0.635
WAIC	743377.85	15790.30	641428.61	16363.86

## C.4. Aplicación 2: Incidencia de dengue por año, Periodo 2010 a 2023

Las estimaciones a posteriori para el Modelo 1 se incluyen en el Cuadro C.10.

Cuadro C.10: Estimaciones a posteriori Modelo 1 - Incidencia durante años 2010 a 2023

Modelo 1				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	-7.033	-2.443	-4.982	0.643
$\beta$	0.681	0.328	0.566	0.109
$\tau_u$	2213.96	1595.459	2252.405	3986.324
$\tau_\nu$	0.01	0.071	0.011	1.206
$\tau_\delta$	2.25	55.350	2.381	35931.222
$1/\Gamma$	-	0.088	-	0.140
p	-	-	0.574	0.635
WAIC	495694.41	15803.05	425820.97	16444.76

Las estimaciones a posteriori para el Modelo 2 se incluyen en el Cuadro C.11.

Cuadro C.11: Estimaciones a posteriori Modelo 2 - Incidencia durante años 2010 a 2023

Modelo 2				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	0.39	-0.098	1.365	1.446
$\tau_u$	2221.949	1250.003	1871.192	1604.241
$\tau_\nu$	0.151	0.108	0.251	1.391
$\tau_\gamma$	5999.350	0.887	28.344	2.118
$\tau_\phi$	3.550	2862.001	2.173	8.825
$1/\Gamma$	-	0.093	-	0.173
p	-	-	0.635	0.635
WAIC	468546.96	15711.77	383570.14	16351.27

Las estimaciones a posteriori para el Modelo 3 se incluyen en el Cuadro C.12.

Cuadro C.12: Estimaciones a posteriori Modelo 3 - Incidencia durante años 2010 a 2023

Modelo 3				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	-1.39	-0.418	0.834	1.321
$\tau_u$	2280	10.069	89800	177.619
$\tau_\nu$	0.092	0.100	0.346	0.749
$\tau_\gamma$	1.33	1.173	1.58	1.835
$\tau_\phi$	21100	22.880	341000	99.977
$\tau_{\delta_{it}}$	0.309	0.862	0.594	3.009
$1/\Gamma$	-	0.105	-	0.215
p	-	-	0.638	0.634
WAIC	730028	15754.67	629693.77	16355.36

### C.5. Aplicación 2: Incidencia de dengue por año considerando covariables climáticas, Periodo 2010 a 2023

Las estimaciones a posteriori para el Modelo 1 se incluyen en el Cuadro C.13.

Cuadro C.13: Estimaciones a posteriori Modelo 1 - Incidencia durante años 2010 a 2023, incluyendo covariables climáticas

Modelo 1				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	-10.313	-15.840	-8.420	-4.104
$\beta$	0.600	0.347	0.523	0.161
$P_{total}$	0.005	0.085	0.006	0.014
$T_{max}$	0.075	0.171	0.009	0.039
$HR_{mean}$	0.185	0.325	0.192	0.175
$\tau_u$	2168.866	265.277	2169.438	2231.702
$\tau_v$	0.015	0.129	0.015	1.719
$\tau_\delta$	3.128	25.041	3.128	3067.430
$1/\Gamma$	-	0.114	-	0.162
p	-	-	0.572	0.635
WAIC	476906.03	15399.95	437728.86	16379.04

Las estimaciones a posteriori para el Modelo 2 se incluyen en el Cuadro C.14.

Cuadro C.14: Estimaciones a posteriori Modelo 2 - Incidencia durante años 2010 a 2023, incluyendo covariables climáticas

Modelo 2				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	-3.309	-12.085	-2.478	-4.696
$P_{total}$	0.006	0.086	0.009	0.026
$T_{max}$	0.125	0.206	0.025	0.036
$HR_{mean}$	0.122	0.238	0.164	0.208
$\tau_u$	2450	2409.685	4920	304.958
$\tau_v$	0.278	0.302	0.485	1.005
$\tau_\gamma$	24900	28.538	44400	1.485
$\tau_\phi$	1.15	1.175	1.44	21.392
$1/\Gamma$	-	0.120	-	0.206
p	-	-	0.634	0.634
WAIC	456409.24	15320.28	395899.69	16270.68

Las estimaciones a posteriori para el Modelo 3 se incluyen en el Cuadro C.15.

Cuadro C.15: Estimaciones a posteriori Modelo 3 - Incidencia durante años 2010 a 2023, incluyendo covariables climáticas

Modelo 3				
Parámetro	Poisson	BN	ZIP	ZINB
$\alpha$	-3.322	-13.691	-0.935	-4.360
$P_{total}$	-0.010	0.086	-0.010	0.026
$T_{max}$	0.150	0.202	0.079	0.041
$HR_{mean}$	0.078	0.290	0.094	0.196
$\tau_u$	2680	13.435	1060000	0.228
$\tau_v$	0.149	0.352	0.718	2440
$\tau_\gamma$	1.35	1.364	1.59	1.45
$\tau_\phi$	23600	278.497	17600	27100
$\tau_{\delta_{it}}$	0.318	0.649	0.604	2.01
$1/\Gamma$	-	0.147	-	0.251
p	-	-	0.633	0.635
WAIC	690284.22	15284.41	630864.17	16282.27

## Apéndice D

# Modelos espacio-temporales adicionales

### D.1. Binomial Negativa (BN)

Se asume que  $y_{it} \sim BN(\lambda_{it}, \Gamma)$ , donde  $\lambda_{it}$  es la media del individuo  $i$  en el tiempo  $t$  y  $\Gamma$  es un parámetro de sobredispersión. Además se define  $\lambda_{it} = E_{it}\theta_{it}$ , donde  $E_{it}$  representa la población en riesgo y  $\theta_{it}$  representa la tasa de incidencia.

De forma similar a los modelos espacio-temporales planteados se usa una función de enlace logarítmica para asociar la tasa de la distribución poisson (o la media) con el predictor lineal, por ejemplo para el modelo 1 en la ecuación (3.4),

$$\log(\theta_{it}) = \alpha + \xi_i + (\beta + \delta_i) \times t.$$

De forma similar se procede a definir los modelos 2 y 3.

Además  $\Gamma$  es otro hiperparámetro reparametrizado internamente por:

$$\theta_0 = \log(\Gamma).$$

Para realizar la inferencia bayesiana se le asigna a  $\theta_0$  una distribución a priori penalizada compleja.

Luego se puede definir el modelo BN con la estructura espacio temporal del Modelo 1

como un MGL:

1. Primer nivel: Vector aleatorio,

$$y_{it} \mid \boldsymbol{\eta}, \boldsymbol{\theta} \stackrel{ind}{\sim} \text{BN}(\lambda_{it}, \Gamma),$$

donde  $\lambda_{it} = E_{it}\theta_{it}$  y

$$\log(\lambda_{it}) = \log(E_{it}) + \log(\theta_{it})$$

$$\log(\theta_{it}) = \alpha + u_i + \nu_i + (\beta + \delta_i) \times t.$$

La función de verosimilitud está dada por:

$$L(\mathbf{y} \mid \boldsymbol{\eta}, \boldsymbol{\theta}) = \prod_{t=1}^T \prod_{i=1}^n \pi(y_{it} \mid \boldsymbol{\eta}, \boldsymbol{\theta})$$

2. Segundo nivel: El campo aleatorio gaussiano de Markov, es definido por

$\boldsymbol{\eta} = (\alpha, \beta, u_1, u_2, \dots, u_n, \nu_1, \nu_2, \dots, \nu_n, \delta_1, \delta_2, \dots, \delta_n) = (\alpha, \beta, \mathbf{u}, \boldsymbol{\nu}, \boldsymbol{\delta})$ , donde

$$\alpha \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2),$$

$$\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2),$$

$$\mathbf{u} \sim \mathcal{N}(0, [\tau_u(\mathbf{D}_W - \mathbf{W})]^{-1}),$$

$$\boldsymbol{\nu} \sim \mathcal{N}(0, \tau_\nu^{-1} \mathbf{I}_n),$$

$$\boldsymbol{\delta} \sim \mathcal{N}(0, \tau_\delta^{-1} \mathbf{I}_n).$$

3. Tercer nivel: El vector de hiperparámetros es definido por  $\boldsymbol{\theta} = (\tau_u, \tau_\nu, \tau_\delta, \theta_0)$ . Se asignó una distribución Gamma(1, 0.0005) a todos los parámetros de precisión definidos en  $\boldsymbol{\theta}$ . Se le asignó a  $\theta_0$  una distribución a priori penalizada compleja.

De forma similar se procede a definir los Modelos 2 y 3.

## D.2. Poisson Cero Inflacionada (ZIP)

Se asume que  $y_{it} \sim ZIP(p, \lambda_{it})$  si tiene fdp:

$$\pi(y_{it} \mid \eta, \theta) = p \times 1_{[y_{it}=0]} + (1 - p) \times f_{Poisson}(y_{it} \mid y_{it} > 0)$$

donde  $f_{Poisson}$  es la fdp de una distribución poisson con media  $\lambda_{it} = E_{it}\theta_{it}$  y  $p$  es la probabilidad de  $y_{it} = 0$  (ausencia de dengue) y es otro parámetro que se requiere estimar.

De forma similar a los modelos espacio-temporales planteados se usa una función de enlace logarítmica para asociar la tasa de la distribución Poisson (o la media) con el predictor lineal, por ejemplo para el modelo 1,

$$\log(\theta_{it}) = \alpha + \xi_i + (\beta + \delta_i) \times t.$$

Además  $p$  es otro hiperparámetro reparametrizado internamente por:

$$p = \frac{\exp(\theta_1)}{1 + \exp(\theta_1)}.$$

Para realizar la inferencia bayesiana se le asigna a  $\theta_1$  una distribución a priori  $\mathcal{N}(-1, 0, 2)$ .

Luego se puede definir el modelo ZIP con la estructura espacio temporal del Modelo 1 como un MGL:

1.

$$y_{it} \mid \boldsymbol{\eta}, \boldsymbol{\theta} \stackrel{ind}{\sim} ZIP(p_{it}, \lambda_{it}),$$

donde  $\lambda_{it} = E_{it}\theta_{it}$  y

$$\log(\lambda_{it}) = \log(E_{it}) + \log(\theta_{it})$$

$$\log(\theta_{it}) = \alpha + u_i + \nu_i + (\beta + \delta_i) \times t.$$

La función de verosimilitud está dada por:

$$L(\mathbf{y} \mid \boldsymbol{\eta}, \boldsymbol{\theta}) = \prod_{t=1}^T \prod_{i=1}^n \pi(y_{it} \mid \boldsymbol{\eta}, \boldsymbol{\theta})$$



2. Segundo nivel: El campo aleatorio gaussiano de Markov, es definido por

$$\boldsymbol{\eta} = (\alpha, \beta, u_1, u_2, \dots, u_n, \nu_1, \nu_2, \dots, \nu_n, \delta_1, \delta_2, \dots, \delta_n) = (\alpha, \beta, \mathbf{u}, \boldsymbol{\nu}, \boldsymbol{\delta}), \text{ donde}$$

$$\alpha \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2),$$

$$\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2),$$

$$\mathbf{u} \sim \mathcal{N}(0, [\tau_u(\mathbf{D}_W - \mathbf{W})]^{-1}),$$

$$\boldsymbol{\nu} \sim \mathcal{N}(0, \tau_\nu^{-1} \mathbf{I}_n),$$

$$\boldsymbol{\delta} \sim \mathcal{N}(0, \tau_\delta^{-1} \mathbf{I}_n).$$

3. Tercer nivel: El vector de hiperparámetros es definido por  $\boldsymbol{\theta} = (\tau_u, \tau_\nu, \tau_\delta, \theta_1)$ . Se asignó una distribución Gamma(1, 0.0005) a todos los parámetros de precisión definidos en  $\boldsymbol{\theta}$ . Se le asignó a  $\theta_1$  una distribución a priori  $\mathcal{N}(-1, 0, 2)$ .

De forma similar se procede a definir los Modelos 2 y 3.

### D.3. Binomial Negativa Cero Inflacionada (ZINB)

Se asume que  $y_{it} \sim ZINB(p, \lambda_{it}, \Gamma)$  si tiene fdp:

$$\pi(y_{it} | \eta, \theta) = p \times 1_{[y_{it}=0]} + (1 - p) \times f_{BN}(y_{it} | y_{it} > 0)$$

donde  $f_{BN}$  es la fdp de una distribución binomial negativa con media  $\lambda_{it} = E_{it}\theta_{it}$ , parámetro de dispersión  $\Gamma$  y  $p$  es la probabilidad de  $y_{it} = 0$  (ausencia de dengue) y estos dos últimos son parámetros que se requiere estimar.

De forma similar a los modelos espacio-temporales planteados se usa una función de enlace logarítmica para asociar la tasa de la distribución Poisson (o la media) con el predictor lineal, por ejemplo para el Modelo 1,

$$\log(\theta_{it}) = \alpha + \xi_i + (\beta + \delta_i) \times t.$$

Como  $p$  es otro hiperparámetro es reparametrizado internamente por:

$$p = \frac{\exp(\theta_1)}{1 + \exp(\theta_1)}.$$

Para realizar la inferencia bayesiana se le asigna a  $\theta_1$  una distribución a priori  $\mathcal{N}(-1, 0, 2)$ .

Además  $\Gamma$  es otro hiperparámetro reparametrizado internamente por:

$$\theta_0 = \log(\Gamma).$$

Para realizar la inferencia bayesiana se le asigna a  $\theta_0$  una distribución a priori penalizada compleja.

Luego se puede definir el modelo ZINB con la estructura espacio temporal del Modelo 1 como un MGL:

1.

$$y_{it} \mid \boldsymbol{\eta}, \boldsymbol{\theta} \stackrel{ind}{\sim} \text{ZINB}(p_{it}, \lambda_{it}, \Gamma),$$

donde  $\lambda_{it} = E_{it}\theta_{it}$  y

$$\log(\lambda_{it}) = \log(E_{it}) + \log(\theta_{it})$$

$$\log(\theta_{it}) = \alpha + u_i + \nu_i + (\beta + \delta_i) \times t.$$

La función de verosimilitud está dada por:

$$L(\mathbf{y} \mid \boldsymbol{\eta}, \boldsymbol{\theta}) = \prod_{t=1}^T \prod_{i=1}^n \pi(y_{it} \mid \boldsymbol{\eta}, \boldsymbol{\theta})$$

2. Segundo nivel: El campo aleatorio gaussiano de Markov, es definido por

$$\boldsymbol{\eta} = (\alpha, \beta, u_1, u_2, \dots, u_n, \nu_1, \nu_2, \dots, \nu_n, \delta_1, \delta_2, \dots, \delta_n) = (\alpha, \beta, \mathbf{u}, \boldsymbol{\nu}, \boldsymbol{\delta}), \text{ donde}$$

$$\alpha \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2),$$

$$\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2),$$

$$\mathbf{u} \sim \mathcal{N}(0, [\tau_u(\mathbf{D}_W - \mathbf{W})]^{-1}),$$

$$\boldsymbol{\nu} \sim \mathcal{N}(0, \tau_\nu^{-1} \mathbf{I}_n),$$

$$\boldsymbol{\delta} \sim \mathcal{N}(0, \tau_\delta^{-1} \mathbf{I}_n).$$

3. Tercer nivel: El vector de hiperparámetros es definido por  $\boldsymbol{\theta} = (\tau_u, \tau_\nu, \tau_\delta, \theta_1, \theta_0)$ . Se asignó una distribución Gamma(1, 0.0005) a todos los parámetros de precisión definidos

en  $\theta$ . Se le asignó a  $\theta_1$  una distribución a priori  $\mathcal{N}(-1, 0, 2)$  y a  $\theta_0$  una distribución a priori penalizada compleja.

De forma similar se procede a definir los Modelos 2 y 3.



# Bibliografía

- Anscombe, F. (1949). The statistical analysis of insect counts based on the negative binomial distribution, *Biometrics* **5**: 165—173.
- Anselin, L. (2010). Local indicators of spatial association-lisa, *Geographical Analysis* **27**(2): 93—115.
- Arab, A. (2015). Spatial and spatio-temporal models for modeling epidemiological data with excess zeros, *Int J Environ Res Public Health* **12**(9): 10536—10548.
- Austin, P., Brunner, L. y Hux, J. (2002). Bayeswatch: an overview of bayesian statistics, *Journal of Evaluation in Clinical Practice* **12**(9): 10536—10548.
- Bell, B. y Broemeling, L. (2000). A bayesian analysis for spatial processes with application to disease mapping, *Statistics in Medicine* **19**: 957—974.
- Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. y Songini, M. (1995). Bayesian analysis of space-time variation in disease risk, *Stat Med* **14**(21-22): 2433—2443.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(2): 192—225.
- Bivand, R., Gomez-Rubio, V. y Rue, H. (2015). Spatial data analysis with r-inla with some extensions, *J Stat Softw* **63**(20): 1—31.
- Blangiardo, M. y Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*, first edition edn, John Wiley Sons, Ltd, Online ISBN:9781118950203 —DOI:10.1002/9781118950203.
- Blangiardo, M., Cameletti, M., Baio, G. y Rue, H. (2013). Spatial and spatio-temporal model with inla, *Spat Spatiotemporal Epidemiol* **4**(1): 33—44.

- Brooks, S., Gelman, A., Jones, G. y Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*, Boca Raton, FL: Chapman Hall/CRC Press.
- Cabezas, C., Fiestas, V., García-Mendoza, M., Palomino, M., Mamani, E. y Donaires, F. (2015). Dengue en el Perú: a un cuarto de siglo de su reemergencia, *Rev Peru Med Exp Salud Publica* **32**: 146—56.
- Casella, G. y George, E. (1992). Explaining the gibbs sampler, *American Statistician* **46**: 167—174.
- CDC (2022). *Alertas epidemiológicas*, Centro Nacional de Epidemiología Prevención y Control de Enfermedades. <https://www.dge.gob.pe/portalnuevo/centros/alerta-y-respuesta/alerta-y-respuesta/#tab-content-5>,.
- DIGESA (2022). *Distritos infestados por Aedes aegypti a nivel nacional - Vigilancia y Control Vectorial*, Dirección General de Salud Ambiental.
- Dunson, D. (2001). Practical advantages of bayesian analysis of epidemiologic data, *American Journal of Epidemiology* **153**(12): 1222–1226.
- Ebi, K. y Nealon, J. (2016). Dengue in a changing climate, *Environ Res.* **151**: 115–123.
- Geary, R. (1954). The contiguity ratio and statistical mapping, *The Incorporated Statistician* (5): 115–145.
- Gelman, A., Hwang, J. y Vehtari, A. (2014). Understanding predictive information criteria for bayesian models, *Statistics and Computing* **24**(6): 997–1016.
- Gilks, W., Richardson, S. y Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*, Chapman Hall/CR.
- Hasting, W. (1970). Monte carlo sampling methods using markov chains and their applications, *Biometrika* **57**(1): 97–109.
- Hilbe, J. (2007). *Negative binomial regression*, Cambridge: Cambridge University Press.
- INS (2018). *Eficacia y seguridad de la vacuna contra el dengue*, Instituto Nacional de Salud. <http://www.bvs.minsa.gob.pe/local/MINSA/4511.pdf>,.
- Jaya, I. y Folmer, H. (2020). Bayesian spatiotemporal mapping of relative dengue disease risk in bandung, indonesia, *J Geogr Syst* **22**: 105–142.

- Knorr-Held, L. (2000). Bayesian modeling of inseparable space–time variation in disease risk, *Stat Med* **19**(17-18): 2555–2567.
- Kouri, G., Pelegrino, J. y Guzmán, B. (2007). Sociedad, economía, inequidades y dengue, *Rev Cuba Med Trop [Internet]* **59**(3): 177–185.
- Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing, *Technometrics* **34**(1): 1–14.
- Lana, R., Gomes, M., Lima, T., Honório, N. y Codeço, C. (2017). The introduction of dengue follows transportation infrastructure changes in the state of acre, brazil: A network-based analysis, *PLoS Negl Trop Dis* **11**(e0006070): 610.  
**URL:** <http://doi:10.1371/journal.pntd.0006070>
- Lowe, R., Bailey, T., Jupp, T., Graham, R., Barcellos, C. y Carvalho, M. (2013). The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in southeast brazil, *Statistics in Medicine* **32**: 864–883.
- Lowe, R., Bailey, T., Stephenson, D., Graham, R., Coelho, C. y Carvalho, M. (2011). Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in brazil, *Computers and Geosciences* **37**(3): 371–381.
- Lowe, R., Barcellos, C., Coelho, C., Bailey, T., Coelho, G., Graham, R., Jupp, T., Ramalho, W. M., Stephenson, D. y Rod, X. (2014). Dengue outlook for the world cup in brazil: an early warning model framework driven by real-time seasonal climate forecasts, *The Lancet Infectious Diseases* **14**(7): 619–626.
- Lowe, R., Cazelles, B., Paul, R., Coelho, Bailey, T., Coelho, G., Graham, R. y Rod'o, X. (2016). Quantifying the added value of climate information in a spatio-temporal dengue model, *Stochastic Environmental Research and Risk Assessment* **30**(8): 2067–2078.
- Martínez-Beneito, M., López-Quílez, A. y Botella-Rocamora, P. (2008). An autoregressive approach to spatio-temporal disease mapping, *Statistics and Medicine* **27**: 2874–2889.
- McMichael, A., Lendrum, C., Corvalán, C., Ebi, K., Githeko, A., Scheraga, J. y Woodward, A. (2013). *Climate change and human health risks and responses*, World Health Organization.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. y Teller, E. (1953). Equation of state calculations by fast computing machines, *The Journal of Chemical Physics* **21**(6): 1087–1092.



- Monath, T. y Tsai, T. (1997). *Clinical Virology*, In: Richman DD, Whitley RJ, Hayden FG (ed.), New York: Churchill Livinstone Inc.
- Moraga, P. (2020). *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*, ChapmanHall/CRC Biostatistics Series, Taylor Francis Group, LLC.
- Morin, C., Comrie, A. y Ernst, K. (2013). Climate and dengue transmission: Evidence and implications, *Environ Health Perspect* **121**: 1264–1272.
- Morán, P. (1950). Notes on continuous stochastic phenomena, *Biometrika* **37**: 17–23.
- Mostorino, R., Rosas, A., Gutiérrez, V., Anaya, E., Cobos, M. y García, M. (2002). Manifestaciones clínicas y distribución geográfica de los serotipos de dengue en el Perú. año 2001, *Rev Peru Med Exp Salud Pública* **19**(4): 171–180.
- OMS (2023). *Organización Mundial de la Salud. Dengue y dengue grave*. <https://www.who.int/es/news-room/fact-sheets/detail/dengue-and-severe-dengue#:~:text=E1%20dengue%20es%20una%20enfermedad,albopictus>.
- OMS (2024). *Organización Mundial de la Salud. Dengue y dengue grave*. <https://www.who.int/es/news-room/fact-sheets/detail/dengue-and-severe>.
- Reinhold, J., Lazzari, C. y Lahondère, C. (2018). Effects of the environmental temperature on *Aedes aegypti* and *Aedes albopictus* mosquitoes, *A Review. Insects* **9**: pii: E158.
- Rice, C. (1996). *Flaviviridae: The viruses and their replication*, 3rd ed. Philadelphia edn, Lippincott-Raven Publisher, New York City. En: Fields BN, Knipe DM, Howley PM, Chanock RM, Melnick JL, Monath TP, et al (ed.).
- Ripley, B. (1981). *Spatial Statistics*, John Wiley Sons, Inc., New York City.
- Ross, S. (2014). *Introduction to Probability Models*, 11th edition edn, Academic Press, San Diego, CA.
- Rue, H. y Held, L. (2005). *Gaussian Markov random fields: theory and applications*, CRC Press.
- Rue, H. y Martino, S. (2007). Approximate bayesian inference for hierarchical gaussian markov random field models, *Journal of Statistical Planning and Inference* **137**(10): 3177–319.

- Rue, H., Martino, S. y Chopin, N. (2009). Approximate bayesian inference for latent gaussian models using integrated nested laplace approximations (with discussion), *Journal of the Royal Statistical Society: Series B (Methodological)* **71**(2): 319–392.
- Rue, H., Martino, S. y Lindgren, F. (2012). The r-inla project.  
**URL:** <http://www.r-inla.org>
- Shaddick, G., Zidek, J. V. y Schmidt, A. (2024). *Spatio–Temporal Methods in Environmental Epidemiology with R*, second edition edn, CRC Press.
- Smith, A. y Roberts, G. (1993). Bayesian computation via the gibbs sampler and other related markov chain monte carlo methods, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **55**: 3–23.
- Stewart-Ibarra, A., Muñoz, A., Ryan, S., Ayala, E., Borbor-Cordova, M., Finkelstein, J., Mejía, R., Ordoñez, T., Recalde-Coronel, G. y Rivero, K. (2014). Spatiotemporal clustering, climate periodicity, and social-ecological risk factors for dengue during an outbreak in Machala, Ecuador, in 2010, *BMC Infect Dis* **25**(14): 610.
- Tsheten, T., Clements, A., Gray, D., Wangchuk, S. y Wangdi, K. (2020). Spatial and temporal patterns of dengue incidence in bhutan: a bayesian analysis, *Emerg Microbes Infect.* **9**(1): 1360–1371.
- Ugarte, M., Adin, A., Goicoa, T. y Militino, A. (2014). On fitting spatio-temporal disease mapping models using approximate bayesian inference, *Statistical Methods in Medical Research* **23**(6): 507–530.
- Wakefield, J. (2004). A critique of statistical aspects of ecological studies in spatial epidemiology, *Environmental and Ecological Statistics* **11**(1): 31—54.
- Waller, L. y Carlin, B. (2010). Disease mapping, *Chapman Hall CRC Handb Mod Stat Methods* pp. 217–243.  
**URL:** [doi: 10.1201/9781420072884-c14](https://doi.org/10.1201/9781420072884-c14). PMID: 25285319; PMCID: PMC4180601
- Wang, X., Yue, Y. R. y Faraway, J. J. (2018). *Bayesian Regression Modeling with Inla*, 1st ed. CRC Press.
- Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research* **11**(12): 3571—3594.

Zellweger, R., Cano, J., Mangeas, M., Fo, T., mercier, A., Despinoy, M., Menkès, C., Dupont-Rouzeyrol, M., Nikolay, B. y Teurlai, M. (2017). Socioeconomic and environmental determinants of dengue transmission in an urban setting: an ecological study in noumea, new caledonia, *PLoS Negl Trop Dis* **11**(4): 1–18.

