PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Modelo bayesiano geoestadístico beta-inflacionado utilizando NNGP con aplicación a datos de cobertura forestal

TESIS PARA OPTAR EL GRADO DE MAGÍSTER EN ESTADÍSTICA

Presentado por:

Alfonso Carlos Cesar Barriga Pozada

Asesora: Zaida Jesus Quiroz Cornejo

Miembros del jurado:

Dr. Luis Valdivieso

Dr. Cristian Bayes

Dra. Zaida Quiroz

Lima, Julio 2019

Dedicatoria

A mi esposa e hijo por comprenderme cada vez que tenía que dedicarle horas de familia a la maestría.
A mis padres por su apoyo incondicional en todos mis proyectos y etapas de la vida.

Agradecimentos

A mi asesora, Zaida Quiroz, por toda la paciencia y guía que me ha brindado en el desarrollo de la tesis.
A los miembros del jurado, el Dr. Luis Valdivieso y el Dr. Cristian Bayes, por sus críticas y correcciones oportunas.

Resumen

En esta tesis proponemos un nuevo modelo geoestadístico beta inflacionado en ceros y unos utilizando NNGP (del inglés Nearest Neighbor Gaussian Process). La ventaja principal de modelar los efectos espaciales utilizando NNGP es la reducción del elevado tiempo computacional que conlleva modelar un proceso gaussiano, ya que no necesita trabajar con todos los vecinos sino solo con un grupo reducido. La estimación de los parámetros se llevó a cabo desde una perspectiva bayesiana. Además, se llevo a cabo un estudio de simulación en el cual se hicieron pruebas con diferentes cantidades de vecinos para evaluar en términos de RMSE y tiempo computacional la ganancia en la estimación del modelo al agregar más vecinos. Finalmente, se modeló la proporción de cobertura forestal en Hiroshima utilizando el modelo geoestadístico desarrollado, obteniendo buenos resultados.

Palavras-clave: Deforestación, distribución beta inflacionada, geoestadística, inferencia bayesiana, NNGP, RStan.

Abstract

In this thesis, we propose a new geostatistical beta inflated zero-one model using NNGP (Nearest Neighbor Gaussian Process). The main advantage of using NNGP in the modeling of spatial effects is the reduction of the large computing time it takes to model a Gaussian process since it does not need to work with all the neighbors, but only with a small group. The estimation of the parameters is done from a bayesian perspective since the posterior distribution does not have a known shape. In addition, a simulation study was carried out in which tests were done with different amounts of neighbors to evaluate in terms of RMSE and computational time the gain in the models when adding more neighbors. Finally, the proportion of forest cover in Hiroshima was modeled using the developed geostatistical model, obtaining good results.
Keywords: Bayesian inference, deforestation, geostatistics, inflated beta distribution, NNGP, RStan.

Índice general

Índice de figuras viii

Índice de cuadros x

1. Introducción 1
   1.1. Motivación . . . 1
   1.2. Revisión de la literatura . . . 2
   1.3. Objetivos de la tesis . . . 4
   1.4. Organización del Trabajo . . . 5

2. Marco teórico 6
   2.1. Distribución beta y sus distribuciones inflacionadas . . . 6
      2.1.1. Distribución beta . . . 6
      2.1.2. Distribución beta inflacionada en cero y uno . . . 7
   2.2. Inferencia Bayesiana . . . 8
      2.2.1. El teorema de Bayes . . . 8
      2.2.2. Métodos MCMC . . . 8
      2.2.3. Métricas de ajuste y comparación de modelos bayesianos . . . 10
   2.3. Conceptos de Geoestadística . . . 12
      2.3.1. Semivariograma . . . 13
      2.3.2. Krigeaje ordinario . . . 16
   2.4. Proceso Gaussiano de vecinos más cercanos (NNGP) . . . 18

3. Modelos geoestadísticos 21
   3.1. Modelo geoestadístico beta . . . 21
      3.1.1. Definición del modelo . . . 21
      3.1.2. Inferencia Bayesiana . . . 22
   3.2. Modelo geoestadístico beta inflacionado en cero y uno . . . 23
      3.2.1. Definición del modelo propuesto . . . 23
      3.2.2. Inferencia Bayesiana . . . 24

4. Estudio de Simulación 26
4.1. Simulación del proceso Gaussiano 26
4.2. Simulación del modelo geoestadístico beta inflacionado en cero y uno . 27

5. Aplicación 33
5.1. Descripción de los datos 33
5.2. Análisis exploratorio: Semivariograma y Krigeaje 35
5.3. Modelamiento de la PCF 38
5.4. Resultados..... 39
5.5. Predicción 45

6. Conclusiones 47
6.1. Comentarios finales 47
6.2. Sugerencias para investigaciones futuras..... 47

Bibliografía 49

Apéndice A: Cadenas del estudio de simulación 51

Apéndice B: Código de Aplicación 54

Índice de figuras

1.1. Interpolación de los datos de cobertura forestal en Hiroshima. La escala indica la PCF, marrón cuando la región está completamente deforestada, y verde claro cuando la región está intacta. 2

2.1. Formas de la distribución beta según sus parámetros (elaboración propia). 6
2.2. Semivariograma empírico (círculos) usando datos simulados a partir de los parámetros originales del modelo teórico Matérn. Modelo teórico Matérn (línea negra) con parámetros: efecto pepita $\tau^2 = 0.1$ (línea roja), varianza $\sigma^2 = 0.9$, rango $1/\delta = 0.1$, parámetro de suavización $\upsilon = 0.5$ y la meseta parcial (línea verde). 16
2.3. Izquierda: Grafo de 14 aristas (denso). Derecha: Grafo de 7 aristas (disperso o no denso). 20

4.1. Izquierda: Generación aleatoria de 1000 locales en el cuadrado unitario. Derecha: Mapa de la matriz de distancias entre todas las observaciones. 26
4.2. Proceso espacial gaussiano $f(s)$ simulado con rango $1/\delta = 0{,}08$ y varianza marginal $\sigma^2 = 2$. 27
4.3. Histograma de los datos simulados a partir de una distribución beta inflacionada en ceros y unos, asumiendo dependencia espacial. 28
4.4. Convergencia de las cadenas de los parámetros y algunos efectos espaciales para M = 5 locales vecinos 29

5.1. Izquierda: Histograma de la proporción de cobertura forestal. Derecha: Datos de PCF muestreados, donde se muestran los locales completamente deforestados (rojos), locales parcialmente cubiertos de árboles (plomos) , y los locales completamente cubiertos de arboles (verdes). 34
5.2. Relación entre la PCF y las covariables diferencia de altitud máxima y mínima (izquierda), y la densidad poblacional (derecha). 34
5.3. Izquierda: Histograma de la variable de densidad poblacional (N). Derecha: Histograma de la variable N estandarizada. 35
5.4. Izquierda: Histograma de la variable de diferencia de altitud máxima y mínima (R). Derecha: Histograma de la variable R estandarizada. 35

5.5. Semivariograma empírico calculado a partir de las PCF muestreadas. 36
5.6. Ajuste del Modelo Exponencial (izquierda), Gaussiano (derecha) y Esférico (abajo). 37
5.7. Mapa de interpolación de la predicción de PCF (izquierda). Mapa de interpolación de los datos observados (derecha). 37
5.8. Evaluación del krigeaje ordinario. Comparación de las PCF observadas y predichas. 38
5.9. Convergencia de las cadenas de los parámetros estimados y algunos efectos espaciales 41
5.10. Modelo sin efecto espacial. Completamente deforestados y completamente cubiertos de arboles (puntos rojos), parcialmente cubiertos (puntos negros). 43
5.11. Modelo Geoestadístico usando 5 locales vecinos. Completamente deforestados y completamente cubiertos de arboles (puntos rojos), parcialmente cubiertos (puntos negros). 43
5.12. Izquierda: Mapa de interpolación de PCF observado. Derecha: Mapa de estimación de PCF con el modelo geoestadístico usando 5 locales vecinos. 44
5.13. Validación del modelo geoestadístico con M=5 locales vecinos. PCF predicho vs PCF observado en 1000 locales 46

1. Convergencia de las cadenas de los parámetros y algunos efectos espaciales para M=3 locales vecinos 51
2. Convergencia de las cadenas de los parámetros y algunos efectos espaciales para M=7 locales vecinos 52
3. Convergencia de las cadenas de los parámetros y algunos efectos espaciales para M=10 locales vecinos 52
4. Convergencia de las cadenas de los parámetros y algunos efectos espaciales para M=15 locales vecinos 53

Índice de cuadros

4.1. Porcentaje de valores simulados de la variable respuesta Y acotada en [0,1]. 28
4.2. Métricas de ajuste y tiempos de ejecución para cada escenario. 30
4.3. Media a posteriori, desviación estándar a posteriori e intervalos de credibilidad (al 95 %) con M = 3 vecinos..... 30
4.4. Media a posteriori, desviación estándar a posteriori e intervalos de credibilidad con M = 5 vecinos..... 31
4.5. Media a posteriori, desviación estándar a posteriori e intervalos de credibilidad con M = 7 vecinos..... 31
4.6. Media a posteriori, desviación estándar a posteriori e intervalos de credibilidad con M = 10 vecinos..... 32
4.7. Media a posteriori, desviación estándar a posteriori e intervalos de credibilidad con M = 15 vecinos..... 32

5.1. Porcentaje de datos por categor´ıa..... 33
5.2. Partes del semivariograma..... 36
5.3. Modelo Geoestadístico de 5 vecinos: comparación de tiempos entre 1500, 2000 y 3000 datos..... 39
5.4. Porcentaje de datos por categor´ıa..... 39
5.5. Comparación del RECM y tiempo de ejecución para cada modelo 40
5.6. Resumen: media a posteriori, intervalo de credibilidad 95 % para los parámetros del modelo geoestadístico. 42
5.7. Resumen: media a posteriori, desviación estandar a posteriori, intervalo de credibilidad 95 % para los parámetros del modelo de referencia 42

Capítulo 1

Introducción

1.1. Motivación

Actualmente, los bosques cubren al menos la tercera parte de la tierra y cumplen un rol central en la preservación de la biodiversidad y los suelos; además, son parte importante del ciclo hidrológico y ayudan a prevenir el cambio climático (Chakravarty et al., 2012). Sin embargo, la expansión de la agricultura y ganadería, así como la tala indiscriminada, las malas políticas forestales y en general el crecimiento poblacional han contribuido a que la deforestación se extienda de forma rápida y a lo largo de todo el planeta, afectando en mayor medida a países no desarrollados o en vías de desarrollo. Una de las consecuencias más importante es el calentamiento global, ya que la deforestación de los bosques no permite eliminar el exceso de dióxido de carbono en la atmósfera. Por ello que resulta de suma importancia entender el problema y aplicar modelos estadísticos adecuados que permitan la cuantificación de la deforestación en diferentes áreas para que de esta forma se puedan establecer estrategias y políticas de acción adecuadas (Puyravaud, 2003).

La deforestación puede ser medida a través de la cobertura forestal (CF), que representa áreas dotadas de plantas con más de dos metros de altura. La proporción de CF (PCF) es calculada dividiendo el área de cobertura entre la superficie de una celda y puede tomar valores en el intervalo cerrado [0,1]. De forma específica, se puede tener un área completamente deforestada, un área completamente cubierta de árboles ó un área parcialmente cubierta de árboles. En particular, en este trabajo se analizarán los datos de PCF en Hiroshima (Figura 1.1). Estos datos fueron usados en (Nishii and Tanaka, 2013), donde la PCF es calculada en una malla regular, de aproximadamente $1\ km^2 \times 1\ km^2$. En la Figura 1.1 podemos observar un mapa que representa la PCF en Hiroshima, donde se evidencia la presencia de muchas regiones intactas, es decir que estan cubiertas de árboles y no han sido deforestadas. Otra característica que observamos en este mapa es que las regiones completamente deforestadas tienden a estar cercanas, así como las regiones cubiertas de árboles entre sí. Esta característica indica que podría existir dependencia espacial entre las regiones vecinas. Esto en la

práctica tiene sentido porque si se talan árboles en una región, es muy probable que se tale más árboles en zonas o regiones cercanas debido a costos y tiempo.

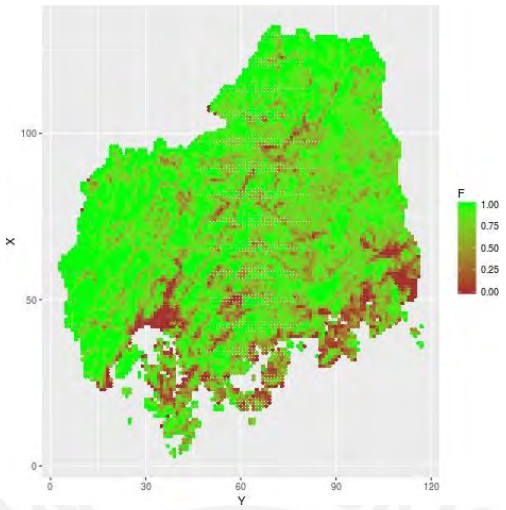

Figura 1.1: Interpolación de los datos de cobertura forestal en Hiroshima. La escala indica la PCF, marrón cuando la región está completamente deforestada, y verde claro cuando la región está intacta.

En particular, los datos que usamos en esta tesis fueron usados por Nishii and Tanaka (2013), quienes proponen modelar la PCF mediante una metodología de dos partes. La primera parte es un modelo log´ıstico trinomial, que sirve para clasificar la PCF en alguno de los tres grupos (cero, uno o en el intervalo abierto entre cero y uno). La segunda parte corresponde a un modelo de regresión logístico-normal para los valores que están en el intervalo $(0,1)$. En ambas partes se añade un componente espacial bajo el concepto de errores normales espacialmente dependientes. Se toma en consideración la influencia del vecindario, el cual se define inicialmente como cuatro regiones adyacentes a la región de medición. Este enfoque es muy restringido, pues es poco probable que una región solo dependa de cuatro regiones vecinas adyacentes, y no se hizo un análisis exploratorio ni se modelaron diferentes modelos que prueben dicha hipótesis. Además no modelaron los datos usando una distribución que considere de forma global la inflación en cero y uno, por el contrario modelaron esta característica usando modelos por partes. En este trabajo proponemos modelar los datos de tal forma que tomemos en cuenta las características de los datos, a través de una distribución adecuada y tomando en cuenta la dependencia espacial a través de efectos aleatorios estructurados, en base a las regiones vecinas próximas, no necesariamente las adyacentes, permitiendo incorporar mayor dependencia entre las regiones.

1.2. Revisión de la literatura

En Ferrari and Cribari-Neto (2004) se propuso un nuevo modelo de regresión en el cual la variable respuesta, en el intervalo (0,1), sigue una distribución beta,

parametrizada de tal forma que se modela la media a través de un predictor lineal. La estimación se lleva a cabo utilizando el método de máxima verosimilitud. Por otro lado, en el trabajo de Ospina and Ferrari (2010) también se trató el problema de modelar datos en el intervalo $[0,1]$ considerando que la variable dependiente puede tomar los valores extremos, cero y uno. Para ello proponen utilizar una distribución mixta que permita capturar la probabilidad en cero, uno o ambos valores. Para modelar el intervalo continuo $(0,1)$ utilizaron una distribución beta y para modelar los extremos plantean utilizar una distribución discreta.

Con respecto a las investigaciones relacionadas al análisis de regresión para variables respuesta acotadas en $[0,1]$, Ospina and Ferrari (2012) proponen un modelo general de regresión beta inflacionado en cero o en uno. Es decir, únicamente cuando uno de los extremos (cero o uno) aparece en los datos. Para ello consideran una mixtura de una distribución beta y una distribución degenerada en cero o en uno dependiendo del caso. La distribución beta se parametriza en términos de la media y el parámetro de precisión. Otro trabajo relacionado a la regresión beta inflacionada es el modelo reparametrizado propuesto por Bayes and Valdivieso (2016). Ellos proponen un modelo de regresión que toma en consideración ambos extremos (0 y 1) y relacionan directamente el valor esperado de la variable respuesta con un predictor lineal a traves de una parametrización especial. Una de las ventajas de este modelo es la mejor interpretación del efecto de las covariables en los odds de la media de la variable respuesta. Fernández et al. (2018) extiende el trabajo anterior al incluir efectos mixtos en la regresión y realizar la estimación de los parámetros bajo el enfoque Bayesiano.

Con respecto a modelos espaciales relacionados con la distribución beta, la literatura sobre trabajos que usen modelos de este tipo son poco abordados. En Kalhori and Mohhamadzadeh (2017) se propone un modelo bayesiano de regresión espacial beta para datos de áreas. Para ello extienden el modelo de regresión beta de Ferrari and Cribari-Neto (2004), le añaden un efecto aleatorio a través del cual incluyen la estructura de la correlación espacial. El efecto espacial aleatorio es tratado como un proceso condicional autoregresivo (CAR) comunmente usado para datos en áreas. Una aplicación del modelo bayesiano espacial, también para datos de áreas, asumiendo que la variable respuesta tiene distribución beta inflacionada en cero y uno, es el desarrollado por Parker et al. (2014). Ellos modelan una medida de la enfermedad periodontal, la cual representa la proporción de lugares que presentan un moderado a severo avance de la enfermedad. Casos en los cuales no hay presencia de la enfermedad (cero) y en los cuales ya es demasiado severa (uno), son estudiados a través de una variable respuesta que pueda tomar valores en el intervalo cerrado $[0,1]$. La dependencia espacial entre los datos es considerada debido a que los dientes vecinos a uno que presente la enfermedad son más propensos a también presentarla. Es así que proponen trabajar bajo el marco de un modelo espacial beta inflacionado en cero y uno, donde la probabilidad de

ocurrencia de la enfermedad periodontal es modelada partiendo del modelo de regresión beta propuesto por Ferrari and Cribari-Neto (2004) y un efecto espacial condicional auto-regresivo (CAR) para controlar la dependencia espacial.

Por otro lado, la literatura sobre modelos para datos georeferenciados usando la distribución beta aún no ha sido tan abordada, incluso no hay modelos geoestadísticos que utilicen la distribución beta inflacionada en ceros y unos y que además combinen NNGP. Un ejemplo más simple que el que proponemos es Lagos-Alvarez et al. (2017), donde proponen un modelo espacial para datos georeferenciados donde las aplicaciones tienen como variables respuestas la tasa de migración y la tasa de divorcio, las cuales se asume que siguen una distribución beta.

1.3. Objetivos de la tesis

Debido a las características de la variable PCF, en este trabajo proponemos analizar la dependencia espacial en los datos observados de la PCF a través de modelos para proporciones entre 0 y 1, incluyendo ambos extremos, que además tomen en cuenta la dependencia espacial en estos datos. En este contexto, se propone utilizar un modelo de regresión beta inflacionado aplicado a datos georeferenciados. No proponemos datos de áreas porque tiene más sentido pensar que la dependencia espacial puede ser mayor que solo con áreas adyacentes. Para modelar los efectos espaciales se utilizará el algoritmo NNGP (del inglés *Nearest Neighbor Gaussian Process*) cuya ventaja principal será reducir el elevado tiempo computacional que conlleva modelar un proceso Gaussiano. La estimación de los parámetros se realizará bajo el enfoque bayesiano, el cual permite tener más flexibilidad de modelamiento, sobre todo para variables respuesta no normales, así como una mayor eficiencia computacional. El objetivo principal de la tesis es proponer un modelo espacial beta inflacionado para estimar y predecir la proporción de cobertura forestal (PCF) en Hiroshima. De manera específica:

- Revisar la literatura acerca de diferentes métodos para lidiar con la dependencia espacial en datos georeferenciados, en particular sobre el NNGP.
- Revisar la literatura acerca de los diferentes modelos espaciales propuestos para variables de proporciones acotadas.
- Desarrollar un modelo espacial que combine la distribución beta inflacionada y NNGP.
- Implementar métodos MCMC para estimar los parámetros del modelo espacial beta inflacionado propuesto para datos georeferenciados.
- Finalmente, aplicar el modelo geoestadístico beta inflacionado propuesto usando NNGP al conjunto de datos de PCF de Hiroshima.

1.4. Organización del Trabajo

La presente tesis se organiza de la siguiente manera. En el Capítulo 2, se presentan los conceptos previos y necesarios para el desarrollo del trabajo, como la definición de la distribución beta y sus distribuciones inflacionadas, inferencia bayesiana y los métodos MCMC, conceptos de geoestadística y modelos espaciales. En el Capítulo 3 se define el modelo geoestadístico beta inflacionado en ceros y unos, se presenta su estructura y la estimación de los parámetros bajo el enfoque Bayesiano. El Capítulo 4 muestra los resultados del estudio de simulación para evaluar el buen ajuste del modelo propuesto. El Capítulo 5 muestra el análisis exploratorio de los datos de Hiroshima y la aplicación del modelo propuesto. Finalmente, en el Capítulo 6 se discuten las conclusiones obtenidas del trabajo. En el anexo se presentan las imágenes de la convergencia de las cadenas (Apendice A) y los códigos para ajustar los modelos pertinentes (Apéndice B).

Capítulo 2

Marco teórico

En el presente capítulo se discutiran los principales conceptos teóricos necesarios para el desarrollo de la tesis.

2.1. Distribución beta y sus distribuciones inflacionadas

2.1.1. Distribución beta

La distribución beta es muy flexible para modelar datos en el intervalo abierto (0,1), dado que su función de densidad presenta diferentes formas dependiendo de los valores que tomen los dos parámetros que definen la distribución. Suponga que Y es una variable aleatoria que sigue una distribución beta, a la cual denotaremos por $Y \sim Beta(a,\beta)$, donde los parámetros a y β controlan la forma de la distribución (véase la Figura 2.1). Entonces la función de densidad de probabilidad (f.d.p.) de Y dada por:

$$f(y) = \frac{\Gamma(a+\beta)}{\Gamma(a)\Gamma(\beta)} y^{a-1}(1-y)^{\beta-1}; 0 \leq y \leq 1.$$

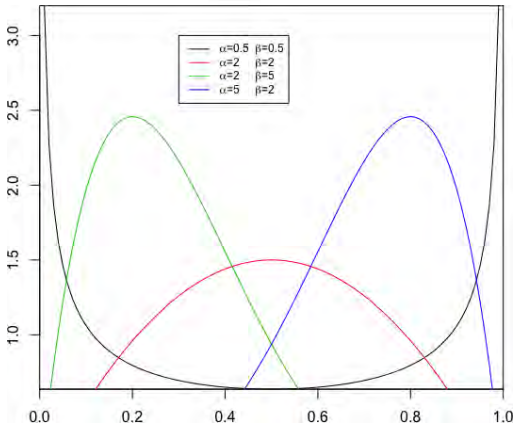

Figura 2.1: Formas de la distribución beta según sus parámetros (elaboración propia).

La esperanza así como la varianza de Y están definidas respectivamente por:

$$E(Y) = \frac{a}{a+\beta};$$
$$V(Y) = \frac{a\beta}{(a+\beta+1)(a+\beta)^2}.$$

La distribución beta puede ser reparametrizada en función de su media μ y el parámetro de precisión φ, tal que $Y \sim Beta(\mu, \varphi)$, donde $\mu = \frac{a}{a+\beta}$ y $\varphi = a+\beta$, (Ferrari and Cribari-Neto, 2004). De esta forma la f.d.p de Y se redefine por:

$$b(y \mid \mu, \varphi) = \frac{\Gamma(\varphi)}{\Gamma(\mu\varphi)\Gamma((1-\mu)\varphi)} y^{\mu\varphi-1}(1-y)^{(1-\mu)\varphi-1}, \quad 0 < y < 1;\ 0 < \mu < 1 \text{ y } \varphi > 0. \tag{2.1}$$

Si $Y \sim Beta(\mu, \varphi)$, la esperanza y varianza de Y, son redefinidas por:

$$E(Y) = \mu;$$
$$V(Y) = \frac{\mu(1-\mu)}{1+\varphi}.$$

2.1.2. Distribución beta inflacionada en cero y uno

La distribución beta no toma en consideración los casos en los cuales la variable dependiente puede tomar valores cero o uno inclusive. La distribución beta inflacionada es básicamente una mixtura discreta entre una distribución beta y una binomial, y permite que la variable respuesta pueda tomar valores en los extremos 0 y 1 (Ospina and Ferrari, 2010).

Suponiendo que Y es una variable aleatoria que toma valores entre $[0,1]$. Se asume que la variable aleatoria Y sigue una distribución beta inflacionada,

$$Y \sim BetaInf(a, \gamma, \mu, \varphi),$$

y su f.d.p es la siguiente:

$$f_Y(y \mid a, \gamma, \mu, \varphi) = \begin{cases} a(1-\gamma) & y = 0 \\ a\gamma & y = 1 \\ (1-a)b(y \mid \mu, \varphi) & y \in (0,1) \end{cases}$$

donde $0 < a, \gamma, \mu < 1$ y $\varphi > 0$ y $b(y \mid \mu, \varphi)$ es la f.d.p. de una distribución beta (Ecuación 2.1); a es un parámetro de mixtura; $a\gamma$ determina la probabilidad de que y sea igual a uno y $a(1-\gamma)$ determina la probabilidad de que y sea igual a cero. La esperanza y la varianza de la distribución de Y son dadas por:

$$E(Y) = a\gamma + (1 - a)\mu,$$

$$V(Y) = a\gamma(1 \_ \gamma) + (1 \_ a)\frac{V(\mu)}{(\varphi + 1)} + a(1 \_ a)(\gamma \_ \mu)^2.$$

2.2. Inferencia Bayesiana

La inferencia bayesiana difiere de la inferencia clásica principalmente en el tratamiento de los parámetros, la primera los trata como variables aleatorias mientras que la segunda los trata como valores fijos (no aleatorios). El enfoque bayesiano permite incorporar información previa o a priori acerca de los parámetros y resulta útil para formular modelos complejos.

En las siguientes secciones de desarrollarán los conceptos básicos y necesarios para aplicar inferencia bayesiana en la estimación de parámetros.

2.2.1. El teorema de Bayes

Todo proceso de inferencia Bayesiana se basa en la actualización de la información del vector de parámetros θ a través del teorema de Bayes

$$\pi(\theta \mid y) = \frac{f(y \mid \theta)\pi(\theta)}{\int f(y \mid \theta)\pi(\theta)\,d(\theta)}, \tag{2.2}$$

donde $f(y \mid \theta)$ es la función de verosimilitud y $\pi(\theta)$ es la f.d.p. a priori del parámetro La expresión del numerador representa la densidad conjunta de θ, Y y el denominador representa la densidad marginal de Y, es decir $f(y)$. La densidad condicional $\pi(\theta \mid y)$ es llamada densidad a posteriori, y se interpreta como la información actualizada acerca de θ luego de observar los datos y la información a priori de θ.

Dado que $f(y)$ no depende del parámetro θ entonces el teorema de Bayes se puede escribir de una forma más compacta:

$$\pi(\theta \mid y) \propto f(y \mid \theta)\pi(\theta),$$

este resultado es la base de todos los procedimientos en inferencia Bayesiana por ello su definición es de extrema importancia.

2.2.2. Métodos MCMC

Los métodos de Monte Carlo vía cadenas de Markov (MCMC) son una alternativa a la integración numérica o a la aproximación analítica cuando la densidad a posteriori no tiene una forma conocida o es demasiado compleja como en los modelos jerárquicos. El MCMC permite lidiar con estas dificultades a través de la generación de valores de

θ a partir de distribuciones aproximadas para luego ir corrigiéndolos hasta obtener una aproximación de la distribución a posteriori $\pi(\theta \mid y)$. Los valores de θ son tomados secuencialmente en donde cada nuevo valor depende únicamente del inmediato anterior, formando una cadena de Markov, es decir, una secuencia de variables aleatorias $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}$ tal que para cualquier t, la densidad de $\theta^{(t)}$, dados todos los θ previos, depende unicamente del ultimo valor observado $\theta^{(t-1)}$. Las distribuciones aproximadas son actualizadas en cada iteración hasta converger a la distribución a posteriori $\pi(\theta \mid y)$ definida en (2.2).

Existen diferentes algoritmos de cadenas de Markov, entre los más destacados se encuentran los de Muestreo de Gibbs y de Metropolis-Hasting.

Algoritmo de Gibbs

Asumiendo que la densidad de interés es $\pi(\theta \mid y)$, donde $\theta = (\theta_1, \ldots, \theta_d)^T$. Cada uno de los componentes θ_i puede ser un escalar, un vector o una matriz. Considerar además que las f.d.p. condicionales $\pi_i(\theta_i) = \pi(\theta_i \mid \theta_{-i})$, para $i = 1, \ldots, d$ y $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)$ son conocidas y pueden ser simuladas. En muchos casos no es posible simular o muestrear directamente de π pero si de π_i. El algoritmo de muestreo de Gibbs ofrece un esquema de generación de muestras a partir de las distribuciones condicionales completas. El algoritmo puede ser descrito en los siguientes pasos:

1. Inicializar el contador de iteraciones en $j = 1$ y definir los valores iniciales $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$.
2. Obtener nuevos valores $\theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_d^{(j)})$ a partir de

$$\theta_1^{(j)} \sim \pi(\theta_1 \mid \theta_2^{(j-1)}, \ldots \theta_d^{(j-1)}),$$

$$\theta_2^{(j)} \sim \pi(\theta_2 \mid \theta_1^{(j-1)}, \theta_3^{(j-1)}, \ldots \theta_d^{(j-1)}),$$

$$\vdots$$

$$\theta_d^{(j)} \sim \pi(\theta_d \mid \theta_1^{(j-1)}, \ldots \theta_{d-1}^{(j-1)}).$$

3. Actualizar el contador j a $j + 1$ y volver al paso 2 hasta lograr la convergencia de todos los θ_i.

Algoritmo de Metropolis-Hasting

Dada una distribución objetivo con f.d.p. $\pi(\theta \mid y)$ no conocida, solo es necesario conocerla en términos de la f.d.p. a priori y la función de verosimilitud. Este algoritmo parte de una distribución con f.d.p. $q(\theta' \mid \theta)$ de la cual es posible simular valores. Cuando el estado actual de una cadena es $\theta = \theta^{(i)}$ se propone un valor θ' para la siguiente iteración a partir de la distribución $q(\theta' \mid \theta)$. Luego, el valor de θ' es aceptado

o rechazado por el algoritmo bajo ciertas condiciones. Si el valor es aceptado el siguiente estado $\theta^{(i+1)}$ será el nuevo θ^{j} , en caso contrario la cadena se queda en el mismo estado. Aceptar o rechazar el valor de θ^{j} depende de la razón:

$$r = r(\theta^{j}, \theta) = \frac{\pi(\theta^{j})q(\theta \mid \theta^{j})}{\pi(\theta)q(\theta^{j} \mid \theta)},$$

donde $\theta = \theta^{(i)}$ es el estado actual y θ^{j} es el estado propuesto. Luego, se genera un valor u de una distribución uniforme$(0, 1)$. Si $u < r$ se acepta el valor propuesto, caso contrario se rechaza. La probabilidad de aceptar el valor propuesto θ^{j}, cuando el estado actual es θ, es dada por

$$P(\text{valor propuesto es aceptado} \mid \theta^{(i)} = \theta, \theta^{j}) = min(1, r(\theta^{j}, \theta)).$$

Estos pasos continuan hasta generar una muestra de valores de θ lo más proxima posible a la distribución estacionaria de la cadena de Markov, que en este caso es la distribución a posteriori $\pi(\theta \mid y)$.

Algoritmo Hamiltoniano

Es un método de Metropolis, aplicable a espacios de estados continuos, que hace uso de la información del gradiente. En este método, el espacio de estados x es aumentado mediante las variables momento p y se alternan dos tipos de propuestas. La primera propuesta aleatoriza la variable momento, dejando el estado x inalterado. La segunda propuesta cambia x y p usando dinámica Hamiltoniana simulada. El Hamiltoniano viene dado por:

$$H(\mathrm{x}, \mathrm{p}) = E(\mathrm{x}) + K(\mathrm{p}),$$

donde $K(p) = \frac{p^T p}{2}$ representa la energía cinética. Estas dos propuestas se usan para crear asintóticamente muestras de la densidad conjunta:

$$f(\mathrm{x}, \mathrm{p}) = exp(-H(\mathrm{x}, \mathrm{p})) = f(x)\varphi(p).$$

Esta distribución es separable así que la distribución marginal de x es la distribución deseada. Por tanto, solo basta descartar las variables momento para obtener una secuencia de muestras $x^{(t)}$ que proceden asintóticamente de $f(x)$.

2.2.3. Métricas de ajuste y comparación de modelos bayesianos

A continuación se definirán las métricas más utilizadas en la evaluación y comparación de modelos bayesianos.

Deviance information criterion (DIC)

Es la versión bayesiana de la conocida métrica AIC (Akaike information criterion) que sirve para comparar el ajuste entre modelos, a menor valor, mejor ajuste del modelo. Esta métrica reemplaza la estimación por máxima versimilitud de los parámetros por la media a posteriori. La formula es la siguiente:

$$DIC = log\ p(y \mid \hat{\theta}) - p_{DIC}$$

donde $\hat{\theta}$ es la media a posteriori y p_{DIC} es la correción por el número efectivo de parámetros, definida como:

$$p_{DIC} = 2\left[log\ p(y \mid \hat{\theta}) - \frac{1}{S}\sum_{s=1}^{S} p(y \mid \theta_s)\right]$$

Widely Applicable Information Criterion (WAIC)

Es una métrica de ajuste introducida por Watanabe (2010) que permite estimar el valor esperado fuera de muestra. Teóricamente se define de la siguiente manera:

$$WAIC = -\frac{1}{n}\sum_{i=1}^{n} log E_\theta\left[p(X_i \mid \theta)\right] + \frac{1}{n}\sum_{i=1}^{n} E_\theta\left(log(X_i \mid \theta)\right)^2 - \left[E_\theta\left(log(X_i \mid \theta)\right)\right]^2$$

A traves de muestras de la densidad a posteriori se puede estimar computacionalmente su valor:

$$\hat{WAIC} = \sum_{i=1}^{n} log\ \frac{1}{S}\sum_{s=1}^{S} p(y_i \mid \theta^s) - p_{\hat{WAIC}}$$

donde S es el número de simulaciones y $p_{\hat{WAIC}}$ es una correción introducida para evitar el sobreajuste, la cual representa el número efectivo de parámetros, definida como:

$$\hat{p} = \sum_{i=1}^{n} V_{WAIC}\left(\sum_{s=1}^{n} log\ p(y_i \mid \theta^s)\right)$$

En comparación con el DIC, el WAIC tiene la propiedad deseable de promediar sobre la densidad a posteriori en lugar de condicionar en una estimación puntual. Esto se vuelve especialmente relevante en un contexto de predicción, ya que se evaluan realmente las predicciones sobre nuevos datos.

La teoría dada aquí es un resumen de conceptos de (Gamerman and Lopes, 2006), (Koistinen, 2010) y (Ghosh et al., 2007).

2.3. Conceptos de Geoestadística

En general se hace referencia a estadística espacial cuando las mediciones de las características de interés en un estudio tienen implícitamente asociadas las coordenadas de las posiciones en donde estas fueron tomadas, y dichas posiciones dependen entre sí debido a su localización en el espacio. En particular, la geoestadística es un área de la estadística espacial que trata el análisis de datos georeferenciados en un dominio espacial continuo $D \in \mathbb{R}^2$.

Sea $Z(s) = \{Z(s_1), \ldots, Z(s_n)\}$ una realización de un proceso estocástico conocido como campo espacial $Z(s) : s \in D$, en donde $D \subset \mathbb{R}^2$ representa una región en el plano. Denotamos como $Z(s_i)$ a una variable aleatoria en la posición $s_i \in D$. A continuación presentamos dos definiciones que se asumen comúmnente para definir campos espaciales.

Definición 1: Un campo espacial $Z(s), \forall s \in D \subset \mathbb{R}^2$ es estrictamente estacionario si cualquier conjunto de locales $\{s_1, \ldots, s_n\}$ y cualquier $h \in \mathbb{R}^2$, la distribución de $(Z(s_1), \ldots, Z(s_n))$ es la misma que para $(Z(s_1 + h), \ldots, Z(s_n + h))$.

Definición 2: Un campo espacial $Z(s), \forall s \in D \subset \mathbb{R}^2$ es débilmente estacionario si la media $\mu(s) = \mu$ y la covarianza $Cov(Z(s), Z(s + h)) = C(h)$, para todo $h \in \mathbb{R}^2$ tal que $s, s + h \in D$.

Esta definición implica que la media es constante para todo local s dentro de la región D, y la covarianza de dos locales cualquiera, cuya distancia entre si es h, tendrán la misma covarianza, así la covarianza es una función de solo h.

Definición 3: Un campo o proceso espacial es llamado proceso espacial Gaussiano si la distribución conjunta de $Z = (Z(s_1), \ldots, Z(s_n))^\intercal$ tiene una distribución normal multivariada con media $\mu = (\mu(s_i), \ldots, \mu(s_n))^\intercal$ y matriz de covarianza Σ compuesta por $Cov(Z(s_i), Z(s_j)) = Cij$ para todo local $s_i, s_j \in D$, tal que

$$f_Z(z) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{-1/2}} \exp\{-\frac{1}{2}(Z - \mu)^\intercal \Sigma^{-1}(Z - \mu)\}.$$

En particular, para un proceso Gaussiano se cumple que estacionariedad débil implica estacionariedad estricta.

La primera etapa en el estudio de datos georeferenciados es el análisis exploratorio, a través de semivariogramas con los cuales se describe la autocorrelación entre posiciones en el espacio. Luego se procede a estimar los parámetros de interés asociados a las dependencias espaciales usando modelos geoestadísticos bajo una serie de supuestos.

Y finalmente se realizan estimaciones y predicciones de la variable de interés.

2.3.1. Semivariograma

En todo análisis geoestadístico la primera etapa exploratoria corresponde al estudio de la dependencia espacial entre las observaciones debido a sus ubicaciones en el espacio, es decir la autocorrelación espacial. Se asume que la variabilidad espacial es finita y puede ser explorada. La función de semivarianza o semivariograma, denotada por $\gamma(h)$, esencialmente caracteriza las propiedades de dependencia del proceso espacial y es definida por:

$$\gamma(h) = \frac{1}{2} var(Z(s_i + h) - Z(s_i)), \ \forall s_i, \ s_i + h \in D,$$

donde asumimos que h es la distancia entre los locales de dos variables $Z(s_i+h)$ y $Z(s_i)$.

Proposición 1: Cuando el campo espacial es un campo espacial débilmente estacionario, se cumple que:

$$\gamma(h) = C(0) - C(h), \tag{2.3}$$

donde $C(0) = \sigma^2$ *es la varianza marginal o "pura" del proceso espacial.*

Prueba de la proposición 1:
Un proceso $Z(s)$ débilmente estacionario implica que $Z(s)$ es estacionario intrínseco. Un campo espacial $Z(s), \ \forall s \in D$ es estacionario intr´ınseco si

$$E[Z(s + h) - Z(s)] = 0 \text{ y}$$

$$E[Z(s + h) - Z(s)]^2 = V[Z(s + h) - Z(s)] = 2\gamma(h).$$

Luego,

$$V[Z(s+h) - Z(s)] = 2\gamma(h) = V[Z(s+h)] + V[Z(s)] - 2Cov[Z(s+h), Z(s)] = 2\gamma(h). \tag{2.4}$$

Como el proceso es débilmente estacionario, por la definición 2 se cumple que

$$Cov[Z(s + h), Z(s)] = C(h), V[Z(s + h)] = C(0) \ \ y V[Z(s)] = C(0), \tag{2.5}$$

Entonces reemplazando la eq. (2.5) en la eq. (2.4) se tiene que :

$$C(0) + C(0) - 2C(h) = 2\gamma(h),$$

luego

$$C(0) - C(h) = \gamma(h).$$

Para más detalles de procesos Gaussianos en estadística espacial ver Banerjee et al. (2004).

El semivariograma está compuesto por tres partes fundamentales:

- Efecto pepita (τ^2): es el comportamiento en el origen, que detecta la variabilidad a muy pequeña escala, es decir, representa estructuras que no son accesibles por la resolución muestral. Representa una discontinuidad puntual en el semivariograma empírico. Este valor es asociado a errores de medición de la variable en estudio.
- Meseta ($\tau^2 + \sigma^2$): representa la cota superior del semivariograma, o, en otras palabras, es el límite superior cuando la distancia h tiende al infinito. Caracteriza el comportamiento del semivariograma a largas distancias.
- Rango ($1/\delta$): se dice que es la distancia a partir de la cual las observaciones no son dependientes espacialmente, cuando el semivariograma se vuelve constante. El rango se puede interpretar como la distancia hasta la cual existe dependencia espacial entre las observaciones.

Existen diversos modelos teóricos para modelar la semivarianza, entre los más representativos se encuentran los procesos estacionarios que solo dependen de la distancia h entre los locales:

- Modelo Matérn: Donde K_υ es una función de Bessel modificada de orden υ. Y υ es un parámetro de suavización. Su expresión matemática es la siguiente:

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2 \left[1 - \frac{(2\psi t \sqrt{(\upsilon)})^\nu}{2^{\nu-1}\Gamma(\upsilon)} K_\upsilon(2\psi h \sqrt{(\upsilon)})\right], & \text{si } h > 0 \\ \tau^2, & \text{caso contrario.} \end{cases}$$

- Modelo Exponencial: Este modelo se aplica cuando la dependencia espacial tiene un crecimiento exponencial respecto a la distancia entre las observaciones. Este modelo es ampliamente usado. Su expresión matemática es la siguiente:

$$\gamma(h) = \tau^2 + \sigma^2 \left[1 - \exp(-\delta h)\right], \text{ si } h > 0,$$

donde τ^2 representa el efecto pepita, $\tau^2 + \sigma^2$ representa la meseta, $1/\delta$ es llamado rango y h es la distancia entre dos posiciones en el espacio. Se debe tener en cuenta que el rango efectivo ($r*$) queda definido a partir del modelo Mattern para $\upsilon = 1/2$ de la siguiente forma:

$$r* = \frac{\sqrt{8\nu}}{\delta} = \frac{2}{\delta}.$$

- Modelo Gaussiano: Al igual que en el modelo exponencial, la dependencia espacial se desvanece solo en una distancia que tiende a infinito. El principal distintivo de este modelo es su forma parabólica cerca al origen. Su expresión matemática es:

$$\gamma(h) = \tau^2 + \sigma^2 \left(1 - \exp\left(-\delta^2 h^2\right)\right).$$

- Modelo Esférico: Tiene un crecimiento rápido cerca al origen, pero los incrementos marginales van decreciendo para distancias grandes, hasta que para distancias superiores al rango los incrementos son nulos. Su expresión matemática es la siguiente:

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2 \left(\frac{3}{2}\frac{3\psi h}{2} - \frac{1}{2}\frac{(\delta h)^3}{2}\right), & \text{si } 0 < h \leq a \\ \tau^2 + \sigma^2, & \text{si } h > a \end{cases}$$

La función de semivarianza empírica es llamada semivariograma experimental, y es calculada en base a las distancias entre los locales de los datos observados $z(s_i)$. Se calcula de la siguiente manera:

$$\hat{\gamma}(h) = \frac{1}{2\,|N(h))|} \sum_{(s_i, s_j) \in N(h)} (z(s_i + h) - z(s_i))^2, \tag{2.6}$$

donde $N(h)$ es el conjunto de pares de locales tales que la distancia entre ellos es h y $|N(h)|$ es el número de pares en este conjunto. La función de semivarianza se calcula para diferentes distancias h. En la práctica, se toman diversos intervalos de distancia y el semivariograma experimental se calcula con una distancia promedio entre parejas de dichos intervalos (ver Figura 2.2).

En la práctica, el semivariograma es muy usado para realizar una análisis exploratorio de la autocorrrelación espacial entre las observaciones. Para ello, a partir de los datos observados se grafica el semivariograma experimental y se define "visualmente" δ, τ^2 y σ^2, para luego ajustar un modelo teórico. Para interpretar el semivariograma experimental se parte del criterio de que a menor distancia entre los locales existirá una mayor similitud o correlación espacial entre sí. Entonces, cuando hay presencia de autocorrelación espacial, se espera que para valores de h pequeños el semivariograma experimental tenga magnitudes menores a las que este toma cuando las distancias h se incrementan hasta estabilizarse cuando la distancia entre observaciones se acerca al rango efectivo.

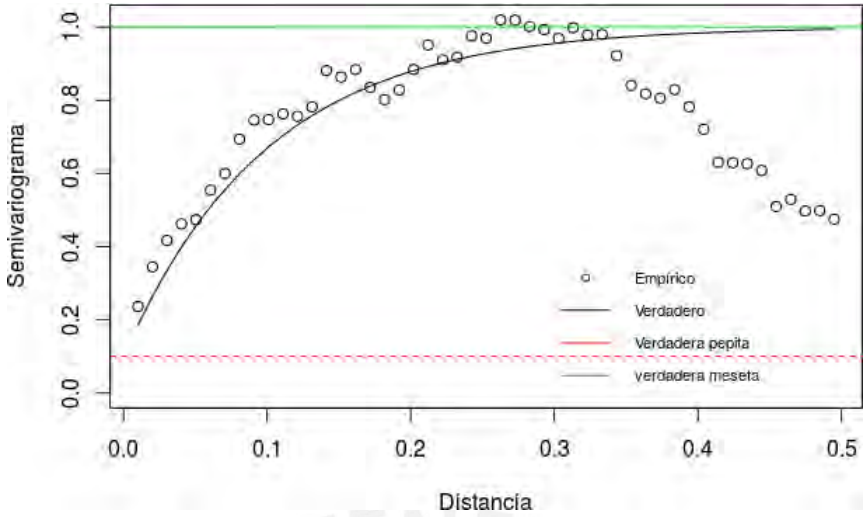

Figura 2.2: Semivariograma empírico (círculos) usando datos simulados a partir de los parámetros originales del modelo teórico Matérn. Modelo teórico Matérn (línea negra) con parámetros: efecto pepita $\tau^2 = 0.1$ (línea roja), varianza $\sigma^2 = 0.9$, rango $1/\delta = 0.1$, parámetro de suavización $\nu = 0.5$ y la meseta parcial (línea verde).

2.3.2. Krigeaje ordinario

Es una técnica geoestadística que sirve para predecir la variable de interés en locales no observados, la cual unicamente asume que el proceso es débilmente estacionario. Sirve principalmente para predecir la variable de interés en locales no observados, en particular es usado con la finalidad de crear mapas de interpolación. Su ventaja frente a otras técnicas de interpolación radica en incluir el comportamiento de la variable en el espacio, tomando el cuenta los parámetros obtenidos observando el variograma empírico. Sin embargo, la predicción obtenida por el krigeaje no está basada en parámetros estimados asumiendo alguna distribución sobre los datos, por ello solo debe ser usada como referencia para realizar un análisis exploratorio.

Existen diferentes tipos de krigeaje, dependiendo de la estacionariedad del campo espacial. El tipo de krigeaje más común es llamado ordinario, el cual asume que el campo espacial es débilmente estacionario, por consiguiente se asume que $E(Z(s_i)) = \mu$. $\forall i = 1, \ldots, n$. El krigeaje ordinario considera que la predicción de $Z(s_0)$ (siendo $s_0 \in D$ la posición de predicción) es calculada ponderando los datos observados, es decir por una combinación lineal de las demás n variables aleatorias $Z(s_i)$, tal que:

$$Z^*(s_0) = \lambda_1 Z(s_1) + \lambda_2 Z(s_2) + \ldots + \lambda_n Z(s_n) = \sum_{i=1}^{n} \lambda_i Z(s_i), \ \forall s_i \in D,$$

en donde los λ_i representan los pesos o ponderaciones para un variable observada en la i-ésima posición. La suma de los pesos debe ser igual a uno, $\sum_{i=1}^{n} \lambda_i = 1$, para asegurar

que la esperanza de $Z^*(s_0)$ sea igual a μ, es decir, que el proceso sea estacionario. Además, las ponderaciones o pesos están basados en los parámetros observados del semivariograma empírico (δ, τ^2, σ^2). Para minimizar el error cuadrático medio restricto a $\sum_{i=1}^{n} \lambda_i = 1$, se usa multiplicadores de Lagrange, es decir, se minimiza:

$$L = Var(Z(s_0) - Z^*(s_0)) + 2a(\sum_{i=1}^{n} \lambda_i - 1),$$

donde a es el multiplicador de Lagrange. Entonces los pesos $\lambda = (\lambda_1, \ldots, \lambda_n)^T$ se obtienen derivando L con respecto a λ_i, $\forall i = 1, \ldots, n$, y ae igualando a cero, tal que se obtiene el siguiente sistema de ecuaciones:

$$\begin{bmatrix} C_{11} & C_{12} & C_{13} & \ldots & C_{1n} & 1 \\ C_{21} & C & C & \ldots & C & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ C_{n1} & C_{n2} & C_{n3} & \ldots & C_n & 0 \\ 1 & 1 & 1 & \ldots & 1 & a \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \cdots \\ \lambda_n \\ \cdots \end{bmatrix} = \begin{bmatrix} C_{10} \\ C_{20} \\ \cdots \\ C_{n0} \\ 1 \end{bmatrix},$$

que se redefinen por :

$$\begin{bmatrix} C_1 & 1_n & \lambda \\ 1_n^T & 0 & a \end{bmatrix} = \begin{bmatrix} c_0 \\ 1_n \end{bmatrix}, \tag{2.7}$$

donde

$$C_1 = \begin{bmatrix} C_{11} & C_{12} & C_{13} & \ldots & C_{1n} \\ C_{11} & C_{12} & C_{13} & \ldots & C_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ C_{n1} & C_{n2} & C_{n3} & \ldots & C_{nn} \end{bmatrix}, \tag{2.8}$$

Por la proposición 1, como el proceso espacial $Z(s)$ es estacionario, se tiene que $\gamma = C(0) - C(h)$, entonces para dos locales s_i y s_j, podemos definir $\gamma_{ij}(h) = C(0) - C_{ij}(h)$ que denotamos como $\gamma_{ij} = C(0) - C_{ij}$. Luego, tenemos que :

$$C_{ij} = C(0) - \gamma_{ij} \tag{2.9}$$

Si además la varianza marginal de $Z(s_i)$ es σ^2, $\forall i = 1, \ldots, n$, entonces

$$C(0) = \sigma^2 \tag{2.10}$$

Reemplazando la eq. (2.10) en la eq. (2.9) se tiene que :

$$C_{ij} = \sigma^2 - \gamma_{ij} \tag{2.11}$$

Reemplazando la eq. (2.11) en la eq. (2.8) :

$$C_1 = \begin{bmatrix} \sigma^2 & \sigma^2 & \dots & \sigma^2 \\ \dots & \dots & \dots & \dots \\ \sigma^2 & \sigma^2 & \dots & \sigma^2 \end{bmatrix} - \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1n} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \cdots & \gamma_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \gamma_{n3} & \cdots & \gamma_{nn} \end{bmatrix},$$

que redefinimos como

$$C_1 = \sigma^2 1_n 1_n^T - \Gamma_1, \tag{2.12}$$

donde:

$$\Gamma_1 = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1n} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \cdots & \gamma_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \gamma_{n3} & \cdots & \gamma_{nn} \end{bmatrix}.$$

De forma similar

$$c_0 = \sigma^2 1_n - \Gamma_0 \tag{2.13}$$

donde

$$\Gamma_0 = \begin{bmatrix} \gamma_{10} \\ \gamma_{20} \\ \vdots \\ \gamma_{n0} \end{bmatrix}.$$

Finalmente, reemplazando la eq. (2.12) y la eq. (2.13) en la eq. (2.7) se tiene que λ es definido por :

$$\lambda = \Gamma_0 + 1_n \frac{(1 - 1_n^T \Gamma_1^{-1} \Gamma_0)}{1_n^T \Sigma_T 1_n} \Gamma_1^{-1}.$$

2.4. Proceso Gaussiano de vecinos más cercanos (NNGP)

En las siguientes secciones se presentará la teoría necesaria para utilizar un *proceso gaussiano de vecinos más cercanos* (NNGP del inglés *Nearest Neighbor Gaussian Process*), la cual fue desarrollada en (Datta et al., 2016). La ventaja de usar esta aproximación es la gran reducción en tiempo computacional al estimar los parámetros a partir de un proceso espacial que tiene una matriz de precisión dispersa (o llena de ceros).

Dado un proceso espacial gausiano , $Z(s) \sim GP(0, C(.|\theta))$, donde $s \in D \subset \mathbb{R}^2$. Sea $S = \{s_1, s_2, \dots s_k\}$ un conjunto de distintas localizaciones en D, llamado conjunto de referencia. Entonces, $Z_S \sim N(0, C_S(\theta))$, donde $Z_S = (Z(s_1), Z(s_2), \dots, Z(s_k))$ y $C_S(\theta)$ es una matriz definida positiva. Luego, se puede escribir la f.d.p. conjunta de Z_S como el producto de densidades condicionales,

$$p(Z_S) = p(Z(s_1))p(Z(s_2)|Z(s_1)) \ldots p(Z(s_k)|Z(s_{k-1}), \ldots, Z(s_1)),$$

y tomar beneficio de esta notación reemplazando los conjuntos condicionales largos por otros más pequeños, de al menos tamaño M, donde $M \quad k$. Entonces, por cada $s_i \in S$, un conjunto condicional más pequeño $N(s_i) \subset S \setminus \{s_i\}$ es usado para construir:

$$\tilde{p}(Z_S) = \prod_{i=1} p\left(Z(s_i)|Z_{N(s_i)}\right),$$

donde $Z_{N(s_i)}$ es un subconjunto de valores del proceso $Z(s)$ en el conjunto de locales $N(s_i)$.

Dado $N_S = \{N(s_i); i = 1, 2, \ldots, k\}$ un subconjunto de k locales de S. Se puede ver el par $\{S, N_S\}$ como un grafo dirigido G con S y N_S como el conjunto de nodos y aristas respectivamente. Para cada dos nodos s_i y s_j se puede decir que s_j es el vecino directo de s_i si existe una arista dirigida de s_i a s_j. Entonces $N(s_i)$ denota el conjunto de M vecinos de s_i. Un grafo es conocido como *grafo acíclico dirigido* (GAD) si no presenta ciclos. En particular, se cumplirá que si G es un (GAD) entonces $\tilde{p}(Z_S)$ es una f.d.p. conjunta.

A partir de $p(Z_S)$ se puede derivar $\tilde{p}(Z_S)$ usando un grafo acíclico dirigido. Esta propiedad resulta especialmente útil si Z_S sigue una distribución gaussiana y G es suficientemente disperso o no denso (Figura 2.3). Bajo este contexto se cumple que:

$$\tilde{p}(Z_S) = \prod_{i=1} N\left(Z(s_i)|B_{s_i}Z_{N(s_i)}, F_{s_i}\right), \tag{2.14}$$

donde $B_{s_i} = C_{s_i,N(s_i)}C^{-1}_{N(s_i)}$ es la media y $F_{s_i} = C(s_i, s_i) - C_{s_i,N(s_i)}C^{-1}_{N(s_i)}C_{N(s_i),s_i}$ es la varianza de la distribución. Además, $\tilde{p}(Z_S)$ es una f.d.p. conjunta y específicamente en (3.8) una densidad Gaussiana con matriz de covarianza $\tilde{C}_S = B_S F_S^{-1} B_S$ y matriz de precisión $\tilde{C}_S^{-1}$ la cual es dispersa. En Datta et al. (2016) construyen un proceso espacial válido a partir de estas definiciones, el cual es llamado NNGP.

Lo mencionado lineas arriba será cierto para cualquier elección de $N(s_i)$ que asegure que G es un grafo acíclico dirigido; sin embargo, la precisión de la aproximación dependerá de los elementos elegidos y la cantidad de vecinos. En Datta et al. (2016) mencionan que tomar los m elementos más cercanos de acuerdo a la distancia Euclideana es una elección practica que da muy buenos resultados en las estimaciones. Finalmente, $\tilde{p}(Z_S)$ denotará la densidad del vecino más cercano de Z_S.

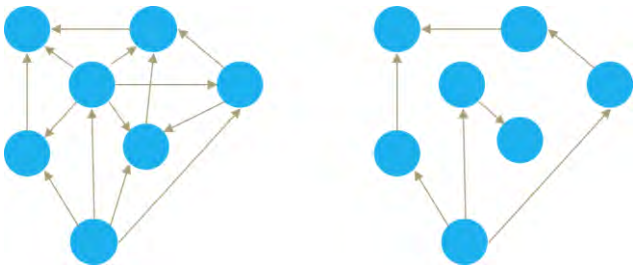

Figura 2.3: Izquierda: Grafo de 14 aristas (denso). Derecha: Grafo de 7 aristas (disperso o no denso).

Capítulo 3

Modelos geoestadísticos

3.1. Modelo geoestadístico beta

Sea $Y_i = Y(s_i)$ la variable aleatoria que representa la variable de interés en el local $s_i \in D$, $i = 1...n$, tal que $0 < y < 1$. Se propone modelar la variable Y_i utilizando una distribución beta con un efecto aleatorio espacial que modele la dependencia espacial entre las observaciones a través de NNGP.

3.1.1. Definición del modelo

Asumimos que Y_i sigue una distribución beta, tal que su función de densidad de probabilidad (f.d.p.) es dada por:

$$f_{Y_i}(y_i \mid \mu_i, \varphi) = b(y_i \mid \mu_i, \varphi) = \frac{\Gamma(\varphi)}{\Gamma(\mu_i\varphi)\Gamma((1-\mu_i)\varphi)} y_i^{\mu_i\varphi-1}(1-y_i)^{(1-\mu_i)\varphi-1}, \qquad (3.1)$$

donde $0 < y_i < 1$, $0 < \mu < 1$ y $\varphi > 0$.
Se puede asociar μ_i con variables explicativas y efectos aleatorios usando funciones de enlace de la siguiente forma:

$$g(\mu_i) = \eta_i = X_i\beta + f_i,\ i = 1, 2, \cdots, n$$

donde g es la función logística $g(x) = \log\left(\frac{x}{1-x}\right)$, la cual enlaza el predictor lineal η_i con el parámetro $\mu_i = E(Y_i \mid 0 < y_i < 1)$. Se tiene que X_i es un vector de covariables, β es un vector de coeficientes (o parámetros de regresión) de X_i. Además, se tiene que $f_i = f(s_i)$ es el efecto espacial aleatorio asociado a μ_i. Se asume que el efecto espacial $f(s) = (f(s_1), \ldots, f(s_n))^T$ usando NNGP tiene la siguiente f.d.p.

$$p(f(s)|\delta, \sigma) = \prod_{i=1}^{n} N\left(f(s_i)|B_{s_i}Z_{N(s_i)}, F_{s_i}\right), \qquad (3.2)$$

donde $B_{s_i} = C_{s_i,N(s_i)}C^{-1}_{N(s_i)}$ y $F_{s_i} = C(s_i, s_i) - C_{s_i,N(s_i)}C^{-1}_{N(s_i)}C_{N(s_i),s_i}$. De donde se deriva

ET LUX IN TENEBRIS LUCET
MCMXVII

$f(s)$ como un proceso gaussiano con vector de medias cero y función de covarianza $\tilde{C}$, es decir, $f(s) \sim NNGP(0, \tilde{C})$, como se definió en la sección 2.4. En particular, la función de covarianza $\tilde{C}$ es definida a partir de una función de covarianza exponencial C, cuyos elementos son definidos por $C(h) = \sigma^2\rho(|h|)$, siendo σ^2 la varianza marginal del campo espacial y $\rho(h) = exp(-\delta h)$ la función de correlación exponencial, donde δ es un parámetro de escala asociado al rango y $|h|$ es la distancia euclidiana entre dos locales.

3.1.2. Inferencia Bayesiana

En esta sección se presenta el proceso para realizar inferencia sobre el modelo 3.1, usando un enfoque bayesiano. Para ello se define el vector de parámetros $\theta = \{f(s), \beta, \varphi, \sigma, \delta\}$ del modelo. Entonces se asumen las siguientes especificaciones para el modelo:

Verosimilitud:

Asumiendo que los $Y_i's$, dado μ_i, φ, son condicionalmente independientes, la función de verosimilitud obtenida a partir del modelo es:

$$L(\theta; y) = \prod_{i=1}^{n} \frac{\Gamma(\varphi)}{\Gamma(\mu_i\varphi)\Gamma((1-\mu_i)\varphi)} y_i^{\mu_i\varphi-1}(1-y_i)^{(1-\mu_i)\varphi-1}. \tag{3.3}$$

Distribuciones a priori:

Se asume que la f.d.p a priori es:

$$p(\theta) = p(\beta)p(\varphi)p(\sigma)p(\delta)p(f(s)|\delta, \sigma);$$

donde el efecto aleatorio es definido por un NNGP, y las f.d.p son definidas por:

$$\begin{aligned} f(s)|\delta, \sigma &\sim \mathrm{N}(0, \tilde{C}) \\ \beta &\sim \mathrm{N}(0, 10^6), \\ \varphi &\sim \mathrm{I}(0, \infty), \\ \sigma &\sim \mathrm{I}(0, \infty), \\ \delta &\sim \mathrm{Gamma}(3, 0.5), \end{aligned} \tag{3.4}$$

.

F.d.p de la distribución a posteriori conjunta:

El objetivo del proceso de inferencia, es obtener la distribución a posteriori conjunta de todos los parámetros del modelo, y a través de esta, obtener estimaciones (puntuales e intervalares) de cada uno de ellos. A través del teorema de Bayes, se obtiene la siguiente

p.d.f. de la distribución a posteriori conjunta:

$$p(\theta|y) \propto L(\theta|y)p(f(s)|\delta, \sigma)p(\theta),$$
$$p(\theta|y) \propto p(\beta)p(\varphi)p(\delta)p(\sigma)p((f(s)|\delta, \sigma)L(\theta; Y),$$
$$p(\theta|y) \propto p(\beta)p(\varphi)p(\delta)p(\sigma)p((f(s)|\delta, \sigma) \prod_{i=1}^{n} \frac{\Gamma(\varphi)}{\Gamma(\mu_i\varphi)\Gamma((1-\mu_i)\varphi)} y_i^{\mu_i\varphi-1}(1-y_i)^{(1-\mu_i)\varphi-1}.$$

Una vez definido el modelo completamente se procede a estimar los parámetros. Dado que las condicionales completas no tienen forma conocida, se usan métodos de MCMC como los descritos en la sección 2.2.

3.2. Modelo geoestadístico beta inflacionado en cero y uno

En este apartado se extenderá el modelo presentado en la sección anterior al caso inflacionado en cero y uno. Sea $Y_i = Y(s_i)$ la variable aleatoria que representa la variable de interés en el local $s_i \in D$, $i = 1...n$, tal que $0 \leq y_i \leq 1$. Se propone modelar la variable Y_i utilizando una distribución beta inflacionada en ceros y unos modelando la dependencia espacial entre las observaciones a través de NNGP.

3.2.1. Definición del modelo propuesto

Asumimos que Y_i sigue una distribución beta inflacionada en cero y uno, tal que su f.d.p. es dada por:

$$f_Y(y_i \mid a_i, \gamma_i, \mu_i, \varphi) = \begin{cases} a_i(1-\gamma_i) & y = 0, \\ a_i\gamma_i & y = 1, \\ (1-a_i)b(y_i \mid \mu_i, \varphi) & y \in (0,1), \end{cases} \tag{3.5}$$

donde $0 < a_i, \gamma_i, \mu_i < 1$ y $\varphi > 0$ y $b(y_i \mid \mu_i, \varphi)$ es la función de densidad de una distribución beta (Ecuación 2.1). Luego, asumimos que la distribución de Y_i es una mistura discreta, que tiene la siguiente función de densidad:

$$f_Y(y_i \mid p_0, p_1, \mu, \varphi) = a_i(1-\gamma_i)\delta_0 + a_i\gamma_i\delta_1 + (1-a_i)b(y_i \mid \mu_i, \varphi)I_{[0<y_i<1]}, \tag{3.6}$$

donde $\delta_0 = 1$ si $y_i = 0$ y $\delta_1 = 1$ si $y_i = 1$ y b es la f.d.p de la distribución beta.

Se pueden asociar los parámetros a_i, γ_i y μ_i con las variables explicativas o efectos

aleatorios usando funciones de enlace de la siguiente forma:

$$\begin{aligned} g_1(a_i) &= \eta_i^{(1)} = X_i^{(1)}\beta^{(1)}, \\ g_2(\gamma_i) &= \eta_i^{(2)} = X_i^{(2)}\beta^{(2)}, \\ g_3(\mu_i) &= \eta_i^{(3)} = X_i^{(3)}\beta^{(3)} + f_i, \end{aligned} \tag{3.7}$$

en donde g_k $(k = 1, 2, 3)$ son las funciones log´ısticas que enlazan el predictor lineal $\eta_i^{(k)}$ con γ_i, a_i y $\mu_i = E(Y_i \mid 0 < y_i < 1)$. Para cada predictor lineal se tiene que $X^{(k)} = (X_1^{(k)}, X_2^{(k)} \ldots X_n^{(k)})^T$ es una matriz de covariables, donde $(X_i)^{(k)} = (x_{i1}^{(k)}, x_{i2}^{(k)}, \ldots, x_{ip}^{(k)})^T$, luego $\beta^{(k)}$ es un vector de coeficientes (o parámetros de regresión) de $X^{(k)}$. Además, se tiene que $f_i = f(s_i)$ es el efecto espacial aleatorio asociado a μ_i. Se asume que el efecto espacial $f(s) = (f(s_1), \ldots, f(s_n))^T$ tiene la siguiente p.d.f.

$$p(f(s)|\delta, \sigma) = \prod_{i=1}^{n} N\left(f(s_i)|B_{s_i}Z_{N(s_i)}, F_{s_i}\right), \tag{3.8}$$

donde $B_{s_i} = C_{s_i,N(s_i)}C_{N(s_i)}^{-1}$ y $F_{s_i} = C(s_i, s_i) - C_{s_i,N(s_i)}C_{N(s_i)}^{-1}C_{N(s_i),s_i}$. De donde se deriva $f(s)$ como un proceso gaussiano con vector de medias cero y función de covarianza $\tilde{C}$, es decir, $f(s) \sim NNGP(0, \tilde{C})$, como se definió en la sección 2.4. Como en el modelo anterior, la función de covarianza $\tilde{C}$ es definida a partir de una función de covarianza exponencial C, definida por $C(h) = \sigma^2\rho(|h|)$, siendo σ^2 la varianza marginal del campo espacial y $\rho(h) = exp(-\delta h)$ la función de autocorrelación exponencial. Cabe resaltar que por un criterio de parsimonia, se asume que solo la media de la variable acotada entre cero y uno está asociada a una dependencia espacial. Sin embargo, es factible asumir un efecto espacial asociado a cada predictor lineal.

3.2.2. Inferencia Bayesiana

En esta sección se presenta el proceso para realizar inferencia sobre el modelo 3.5, usando un enfoque bayesiano. Para ello se define el vector de parámetros $\theta = \{f(s), \beta^{(1)}, \beta^{(2)}, \beta^{(3)}, \varphi, \sigma, \delta\}$ del modelo. Entonces se asumen las siguientes especificaciones para el modelo:

Verosimilitud:

Asumiendo que los Y_is, dado μ_i, φ, son condicionalmente independientes, la función de verosimilitud obtenida a partir del modelo es:

$$L(\theta; y) = \prod_{i=1}^{n} a_i(1 - \gamma_i)\delta_0 + a_i\gamma_i\delta_1 + (1 - a_i)b(y_i \mid \mu_i, \varphi)I_{[0<y_i<1]}. \tag{3.9}$$

Distribuciones a priori:
Se asume que la f.d.p a priori es:

$$p(\theta) = p(\beta^{(1)})p(\beta^{(2)})p(\beta^{(3)})p(\varphi)p(\sigma)p(\delta)p(f(s)|\delta, \sigma);$$

donde el efecto aleatorio es definido por un NNGP, y las f.d.p son definidas por:

$$\begin{aligned} f(s)|\delta, \sigma &\sim \mathrm{N}(0, \tilde{C}) \\ \beta^{(k)} &\sim \mathrm{N}(0, 10^6); \ \forall k = 1, 2, 3 \\ \varphi &\sim \mathrm{I}(0, \infty), \\ \sigma &\sim \mathrm{I}(0, \infty), \\ \delta &\sim \mathrm{Gamma}(3, 0.5), \end{aligned} \tag{3.10}$$

.

F.d.p de la distribución a posteriori conjunta:

El objetivo del proceso de inferencia, es obtener la distribución a posteriori conjunta de todos los parámetros del modelo, y a través de esta, obtener estimaciones (puntuales e intervalares) de cada uno de ellos. A través del teorema de Bayes, se obtiene la siguiente p.d.f. de la distribución a posteriori conjunta:

$$\begin{aligned} p(\theta|y) &\propto L(\theta|y)p(f(s)|\delta, \sigma)p(\theta), \\ p(\theta|y) &\propto p(\beta)p(\varphi)p(\delta)p(\sigma)p((f(s)|\delta, \sigma)L(\theta; Y), \end{aligned}$$

$$p(\theta|y) \propto p(\beta)p(\varphi)p(\delta)p(\sigma)p((f(s)|\delta, \sigma) \prod_{i=1}^{n} a_i(1-\gamma_i)\delta_0 + a_i\gamma_i\delta_1 + (1-a_i)b(y_i \mid \mu_i, \varphi)I_{[0<y_i<1]} \tag{3.11}$$

Una vez definido el modelo completamente se procede a estimar los parámetros, pero dado que las condicionales completas no tienen forma conocida, se usan métodos de MCMC, como los descritos en la sección 2.2.

Capítulo 4

Estudio de Simulación

Se simulan 1000 datos a partir de los modelos geoestadísticos beta y beta inflacionado definidos en las secciones 3.1 y 3.2 respectivamente. Luego se estiman los parámetros mediante inferencia bayesiana. Finalmente, se evalua la efectividad en la estimación de los parámetros variando el número de vecinos usados para aproximar el proceso espacial. Para todos los procedimientos se utilizó una máquina de alto rendimiento con procesador Intel Xeon core i7 y 32GB de RAM.

4.1. Simulación del proceso Gaussiano

Primero se generan 1000 locales aleatoriamente en $[0,1] \times [0,1]$, luego se calcula la matriz de distancias H entre todos los locales (Figura 4.1).

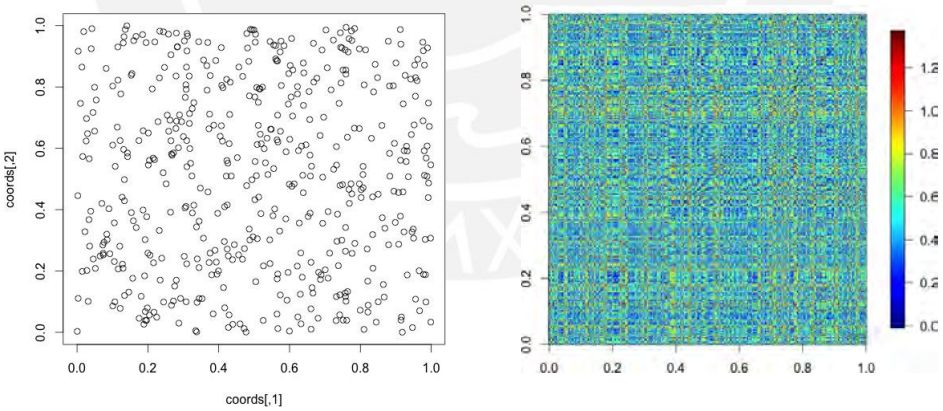

Figura 4.1: Izquierda: Generación aleatoria de 1000 locales en el cuadrado unitario. Derecha: Mapa de la matriz de distancias entre todas las observaciones.

Se asume que el efecto espacial aleatorio sigue un proceso espacial gaussiano con vector de medias cero y matriz de covarianza C, $f(s) \sim GP(0, C)$ donde la función de covarianza exponencial es definida por $C(h) = \sigma^2\rho(|h|)$, siendo σ^2 la varianza marginal del campo espacial, $\rho(h) = exp(-\delta h)$ la función de autocorrelación exponencial. Luego, la matriz de distancias H es utilizada para calcular la matriz de covarianza C. Finalmente, se simulan los efectos espaciales aleatorios asumiendo como parámetros $\delta = 12$ y $\sigma^2 = 2$.

Entonces el rango efectivo definido para $\upsilon = \frac{1}{2}$ es 0.167, un valor pequeño de rango, para el cual se sabe que el NNGP funciona bien. El proceso Gaussiano simulado para los efectos espaciales, considerando los parámetros anteriores, se observa en la Figura 4.2.

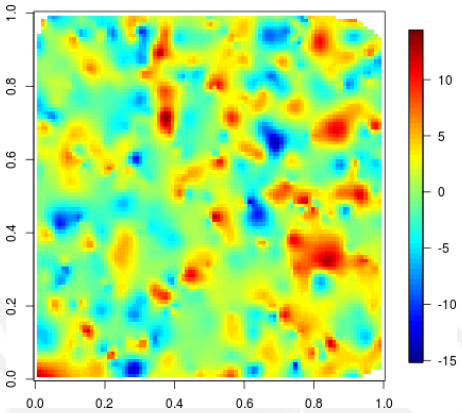

Figura 4.2: Proceso espacial gaussiano $f(s)$ simulado con rango $1/\delta = 0.08$ y varianza marginal $\sigma^2 = 2$.

4.2. Simulación del modelo geoestadístico beta inflacionado en cero y uno

Para simular los valores observados de la variable respuesta asumimos que provienen de una distribución beta inflacionada en cero y uno, cuya f.d.p. es la definida en la Ecuación (3.6) donde:

$$
\begin{aligned}
g_1(\alpha(s_i)) &= \beta_0^{(1)} + \beta_1^{(1)} X_i \\
g_2(\gamma(s_i)) &= \beta_0^{(2)} + \beta_1^{(2)} X_i \\
g_3(\mu(s_i)) &= \beta_0^{(3)} + \beta_1^{(3)} X_i + f(s_i),
\end{aligned}
$$

para $i = 1, \ldots, n$,se tiene que $g_k(.)$ con $k = 1, 2, 3$ son funciones log´ısticas, β_0 es el intercepto, $\beta^{(k)}$ son coeficientes de regresión, X_i es un vector de variables explicativas definidas en el local s_i, y $f(s_i)$ es un efecto espacial estructurado proveniente de un proceso gaussiano tal como se definió en la sección 4.1. Simulamos los datos considerando los siguientes valores para los parámetros: $\beta^{(1)} = (-2, -5)$, $\beta^{(2)} = (0.25, -2)$, $\beta^{(3)} = (1, 2)$, $\varphi = 50$ y el efecto espacial simulado anteriormente con parámetros $\sigma^2 = 2$ y $\delta = 12$. En la Figura 4.3 se observa el histograma del conjunto de datos simulados. Se simuló aproximadamente 18 % de unos, 18 % de ceros y 64 % de valores de la variable respuesta entre cero y uno (Cuadro 4.1).

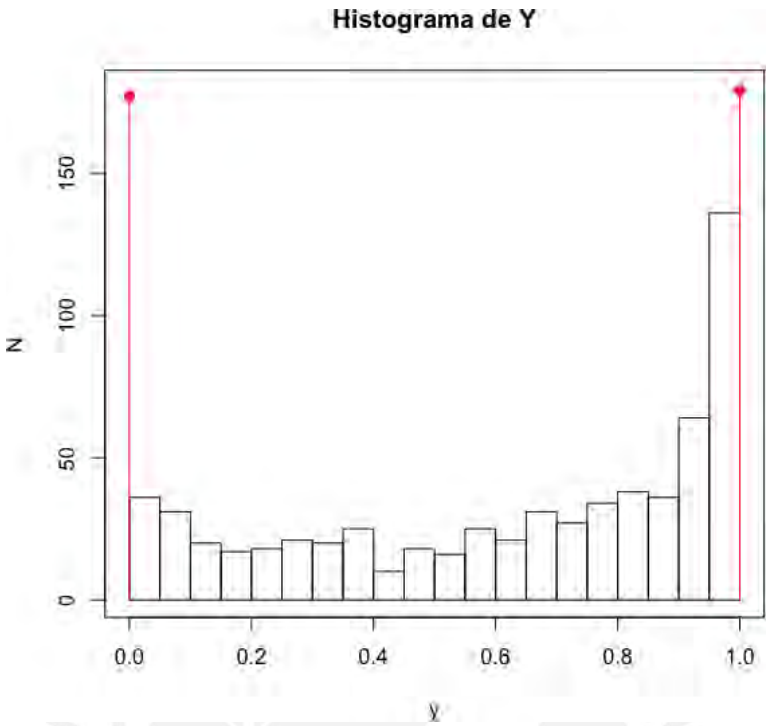

Figura 4.3: Histograma de los datos simulados a partir de una distribución beta inflacionada en ceros y unos, asumiendo dependencia espacial.

$(y = 1)$	177	17.7 %
$(y = 0)$	179	17.9 %
$y \in (0,1)$	644	64.4 %

Cuadro 4.1: Porcentaje de valores simulados de la variable respuesta Y acotada en [0,1].

Para estimar los parámetros del efecto espacial $f(s)$ se asume que el proceso espacial es un NNGP, es decir, $\tilde{f}(s) \sim NNGP(0, \tilde{C})$. Por lo tanto, para estimar los parámetros se consideró la distribución conjunta a posteriori definida en (3.11) asumiendo las distribuciones a priori en (3.10).

Para obtener muestras de las distribuciones a posteriori se usó el algoritmo adaptativo Halmitoniano MCMC con dos cadenas y 20 mil muestras, el cual es implementado en R en el paquete $RStan$. Se usó un burn-in de 10 mil muestras, es decir, la estimación de los parámetros es realizada a través de una muestra de tamaño 10 mil para cada uno de los parámetros. Como parte del estudio, se realizó la estimación de los parámetros bajo diferentes escenarios variando el número de vecinos usados por el NNGP, con el objetivo de ver como este número mejoraba o empeoraba la estimación de los parámetros. En particular, se evaluaron cinco cantidades de vecinos diferentes M = 3, 5, 7, 10 y 15. En las siguientes tablas se observan las medias a posteriori, intervalos de credibilidad y la

desviación estándar a posteriori de cada parámetro para cada número de vecinos.

Para mostrar la convergencia a la distribución conjunta a posteriori, la Figura 4.3 muestra algunas cadenas generadas para cada parámetro y algunos efectos espaciales para el modelo ajustado con M = 5 vecinos. Se puede verificar la convergencia de las dos cadenas del mismo parámetro, ya que convergen al mismo valor. Las demás figuras para la convergencia de los parámetros en cada escenario se encuentran en el Apendice A.

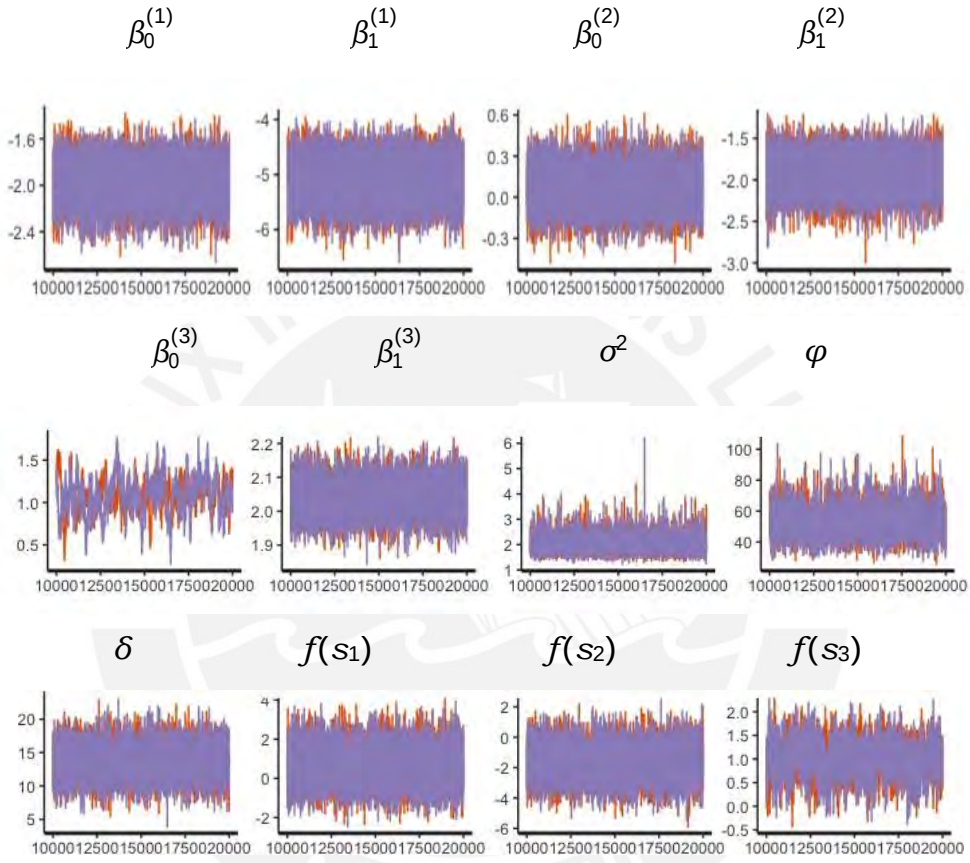

Figura 4.4: Convergencia de las cadenas de los parámetros y algunos efectos espaciales para M = 5 locales vecinos

En el Cuadro 4.2 se observan la raiz del error cuadrático medio (RECM) y los tiempos de ejecución en cada escenario. En términos del RECM la diferencia es mínima entre todos los modelos. Por otro lado, con respecto al tiempo de ejecución, este crece de forma exponencial conforme se incrementa el número M de vecinos, sin dejar una ganancia significativa en la estimación de los parámetros, siendo el modelo con M = 5 vecinos el que obtuvo menor RECM. Esto no significa que siempre sea suficiente usar cinco vecinos para realizar la estimación de los efectos espaciales, pero para estos datos simulados parece ser el número de vecinos más adecuado. En otros datos podría variar, dependiendo del tamaño del rango, ya que a mayor rango probablemente se requiera de un mayor número de vecinos para realizar la estimación adecuadamente.

Podemos observar que para todos los modelos, sin importar el número de vecinos, la estimación de los parámetros fue muy buena, dado que todos los intervalos de

Cuadro 4.2: Métricas de ajuste y tiempos de ejecución para cada escenario.

Número de vecinos (M)	RECM	Tiempo (horas)
M=3	0.266	2.13
M=5	0.267	2.9
M=7	0.266	3.9
M=10	0.266	5.14
M=15	0.266	9.1

credibilidad contienen los verdaderos valores de los parámetros. Como se esperaba, al variar el número de vecinos solo cambian ligeramente los βs asociados al parámetro μ y también cambian el σ^2 y el δ, que son parámetros del NNGP, y el φ que mide la variabilidad restante. Los parámetros correspondientes al efecto espacial, son dificiles de estimar debido a la correlación entre los mismos parámetros, sin embargo, el método de inferencia usado muestra que el NNGP funcionó bien, este resultado se debe a que el valor del parámetro δ produce un rango pequeño, lo cual significa que no se necesitan de demasiados vecinos para estimar el efecto espacial. De los resultados de las simulaciones podemos concluir que el modelo geoestadístico beta inflacionado en ceros y uno propuesto en esta tesis, tiene muchos parámetros para cada predictor lineal, por ello el NNGP es una contribución para este tipo de modelos, porque permite realizar inferencia de una forma más rápida, sin afectar la precisión de las estimaciones.

Cuadro 4.3: Media a posteriori, desviación estándar a posteriori e intervalos de credibilidad (al 95 %) con M = 3 vecinos.

Parámetro		Original	Media	D.E.	IC (95 %)
α	$\beta_0^{(1)}$	-2.00	-1.97	0.17	(-2.32;-1.66)
	$\beta_1^{(1)}$	-5.00	-5.09	0.35	(-5.82;-4.43)
γ	$\beta_0^{(2)}$	0.25	0.07	0.14	(-0.19;0.34)
	$\beta_1^{(2)}$	-2.00	-1.91	0.20	(-2.33;-1.53)
μ	$\beta_0^{(3)}$	1.00	1.10	0.19	(0.75;1.48)
	$\beta_1^{(3)}$	2.00	2.04	0.05	(1.95;2.14)
φ		50.00	50.89	8.85	(35.82;70.48)
σ^2		2.00	2.03	0.29	(1.57;2.69)
δ		12.00	13.07	2.13	(9.05;17.36)

Cuadro 4.4: Media a posteriori, desviación estándar a posteriori e intervalos de credibilidad con M = 5 vecinos.

Parámetro		Original	Media	D.E.	IC (95 %)
a	$\beta_0^{(1)}$	-2.00	-1.97	0.17	(-2.31;-1.66)
	$\beta_1^{(1)}$	-5.00	-5.08	0.35	(-5.81;-4.43)
γ	$\beta_0^{(2)}$	0.25	0.07	0.14	(-0.20;0.34)
	$\beta_1^{(2)}$	-2.00	-1.92	0.21	(-2.34;-1.53)
μ	$\beta_0^{(3)}$	1.00	1.08	0.20	(0.66;1.48)
	$\beta_1^{(3)}$	2.00	2.04	0.05	(1.94;2.14)
φ		50.00	51.93	9.27	(36.03;72.31)
σ^2		2.00	2.01	0.31	(1.53;2.74)
δ		12.00	13.59	2.28	(9.19;18.08)

Cuadro 4.5: Media a posteriori, desviación estándar a posteriori e intervalos de credibilidad con M = 7 vecinos.

Parámetro		Original	Media	D.E.	IC (95 %)
a	$\beta_0^{(1)}$	-2.00	-1.98	0.17	(-2.33;-1.66)
	$\beta_1^{(1)}$	-5.00	-5.09	0.36	(-5.82;-4.42)
γ	$\beta_0^{(2)}$	0.25	0.07	0.14	(-0.20;0.34)
	$\beta_1^{(2)}$	-2.00	-1.91	0.21	(-2.35;-1.52)
μ	$\beta_0^{(3)}$	1.00	1.15	0.22	(0.73;1.58)
	$\beta_1^{(3)}$	2.00	2.04	0.05	(1.94;2.14)
φ		50.00	51.78	9.25	(36.21;72.57)
σ^2		2.00	2.01	0.32	(1.53;2.77)
δ		12.00	13.82	2.38	(9.18;18.52)

Cuadro 4.6: Media a posteriori, desviación estándar a posteriori e intervalos de credibilidad con M = 10 vecinos.

Parámetro		Original	Media	D.E.	IC (95 %)
a	$\beta_0^{(1)}$	-2.00	-1.97	0.17	(-2.32;-1.65)
	$\beta_1^{(1)}$	-5.00	-5.09	0.36	(-5.82;-4.42)
γ	$\beta_0^{(2)}$	0.25	0.07	0.14	(-0.20;0.34)
	$\beta_1^{(2)}$	-2.00	-1.91	0.21	(-2.34;-1.52)
μ	$\beta_0^{(3)}$	1.00	1.20	0.24	(0.78;1.73)
	$\beta_1^{(3)}$	2.00	2.04	0.05	(1.94;2.13)
φ		50.00	51.48	9.15	(36.10;71.77)
σ^2		2.00	2.02	0.32	(1.53;2.79)
δ		12.00	13.71	2.40	(8.95;18.46)

Cuadro 4.7: Media a posteriori, desviación estándar a posteriori e intervalos de credibilidad con M = 15 vecinos.

Parámetro		Original	Media	D.E.	IC (95 %)
a	$\beta_0^{(1)}$	-2.00	-1.98	0.17	(-2.32;-1.66)
	$\beta_1^{(1)}$	-5.00	-5.09	0.36	(-5.83;-4.41)
γ	$\beta_0^{(2)}$	0.25	0.07	0.14	(-0.20;0.33)
	$\beta_1^{(2)}$	-2.00	-1.91	0.21	(-2.33;-1.52)
μ	$\beta_0^{(3)}$	1.00	1.20	0.25	(0.76;1.74)
	$\beta_1^{(3)}$	2.00	2.04	0.05	(1.94;2.13)
φ		50.00	50.95	8.95	(35.85;71.3)
σ^2		2.00	2.07	0.36	(1.54;2.93)
δ		12.00	13.17	2.40	(8.45;17.85)

Capítulo 5

Aplicación

En el presente capítulo se aplicó el modelo propuesto a los datos de proporción de cobertura forestal (PCF) de la prefectura de Hiroshima.

5.1. Descripción de los datos

Los datos de la proporción de cobertura forestal (PCF) utilizados en la presente tesis corresponden a la prefectura de Hiroshima y también fueron analizados en Nishii and Tanaka (2013). Esta variable puede tomar valores en el intervalo cerrado $[0,1]$. En particular, cuando $PCF = 0$ indica un área completamente deforestada (CD), cuando $PCF = 1$ indica un área completamente cubierta de árboles (CC) y, por último, cuando $0 < PCF < 1$ indica un área parcialmente cubierta de árboles (PC). En la Figura 5.1 se observa el histograma de los datos, donde podemos observar que los datos son inflacionados en cero y uno. La Figura 5.1 también muestra a la derecha los datos de PCF observados, donde los puntos rojos corresponden a locales CD, los puntos plomos corresponden a locales PC, y los puntos verdes corresponden a locales CC.

Se calculó el porcentaje de observaciones dentro de cada una de las categorías de la variable, estos porcentajes se observan en el Cuadro 5.1, la variable puede tomar los valores de cero y uno en un 21% de los casos aproximadamente.

Intacto (CC) ($PCF = 1$)	1712	19.68 %
Deforestado (CD) ($PCF = 0$)	159	1.828 %
Parcialmente cubierto (PC) $PCF \in (0, 1)$	6825	78.48 %

Cuadro 5.1: Porcentaje de datos por categor´ıa.

Para explicar la PCF se cuenta con la variable de densidad poblacional (N) en cada uno de los locales s_i, ya que se analizará el impacto del tamaño de la población en un radio de 1 km, en la PCF. También se cuenta con una variable que mide la diferencia entre la altitud máxima y mínima de un local (R), en un radio de 1km, la cual describe las circunstancias del terreno. En la Figura 5.2 se observa cada una de estas variables y su relación con la PCF.

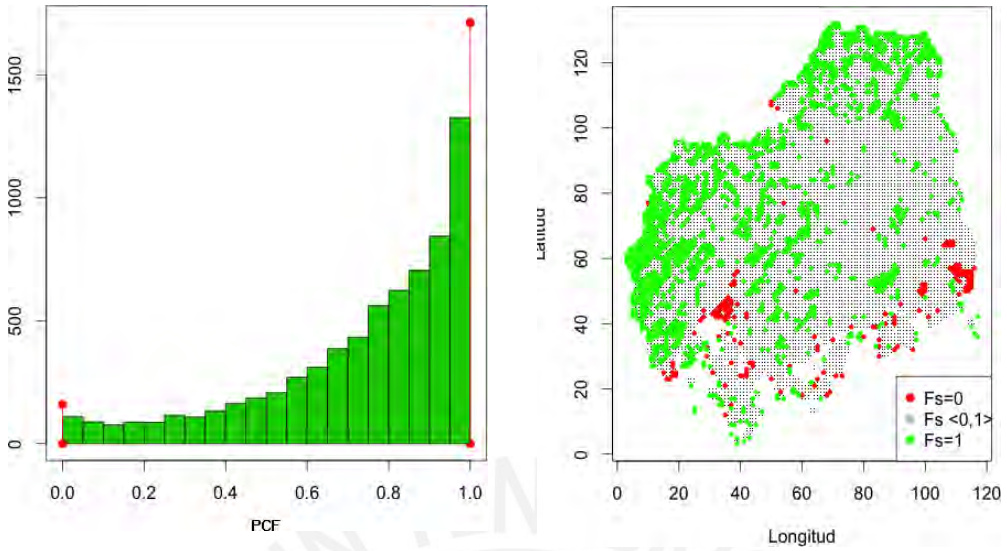

Figura 5.1: Izquierda: Histograma de la proporción de cobertura forestal. Derecha: Datos de PCF muestreados, donde se muestran los locales completamente deforestados (rojos), locales parcialmente cubiertos de árboles (plomos) , y los locales completamente cubiertos de arboles (verdes).

.

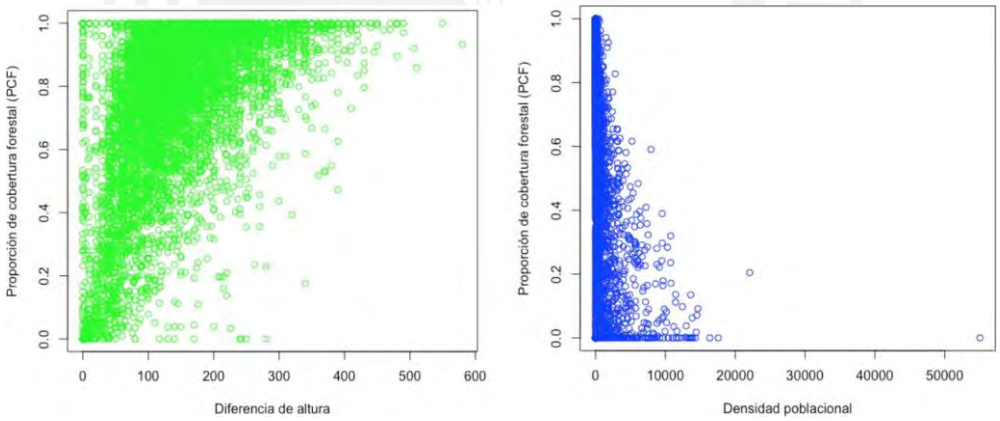

Figura 5.2: Relación entre la PCF y las covariables diferencia de altitud máxima y mínima (izquierda), y la densidad poblacional (derecha).

Luego, las variables fueron estandarizadas de tal forma que la escala en la cual han sido medidas no afecte la estimación. En la Figura 5.3 y 5.4 se observan los histogramas de cada variable antes y despues de ser estandarizadas.

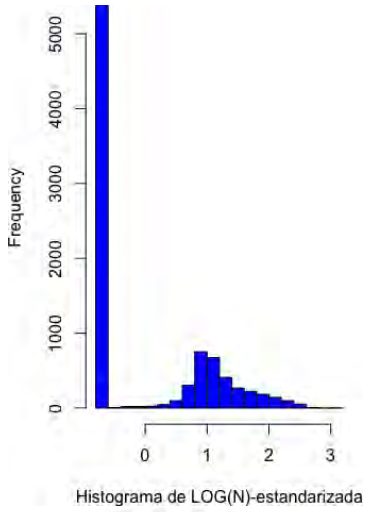

Figura 5.3: Izquierda: Histograma de la variable de densidad poblacional (N). Derecha: Histograma de la variable N estandarizada.

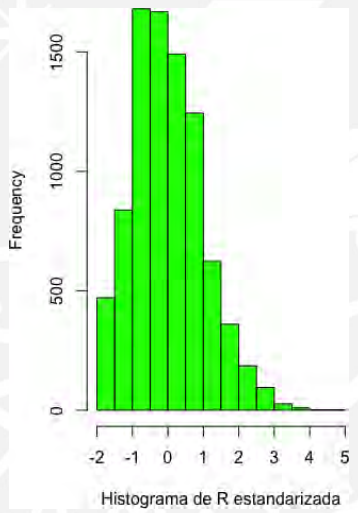

Figura 5.4: Izquierda: Histograma de la variable de diferencia de altitud máxima y mínima (R). Derecha: Histograma de la variable R estandarizada.

5.2. Análisis exploratorio: Semivariograma y Krigeaje

Como primera etapa del análisis exploratorio se debe determinar la existencia de dependencia espacial en los datos. Para ello, se hará uso del semivariograma y posteriormente se usarán los parámetros observados en el semivariograma para usar el krigeaje y realizar predicciones.

El semivariograma se interpreta a partir del criterio de que a menor distancia entre los puntos más parecidos los valores de la variable de interés deberían ser. Entonces, si existe autocorrelación espacial, se espera que para valores de distancia pequeñas la semivarianza sea de menor magnitud que para valores de distancia más grandes. En

la Figura 5.5 se observa el semivariograma emp´ırico de los datos de Hiroshima, en el cual efectivamente observamos que a mayor distancia el valor de la semivarianza se incrementa, hasta llegar a un tope el cual es conocido como meseta. Por otro lado, la discontinuidad en el origen se produce debido a errores no identificables y se conoce como efecto pepita. Otro elemento importante del semivariograma es el rango, que se interpreta como la distancia a partir de la cual los datos no son más dependientes espacialmente, entre más corto sea el rango, menor dependencia espacial existirá. En este caso se observa un rango de entre 10 y 25 grados.

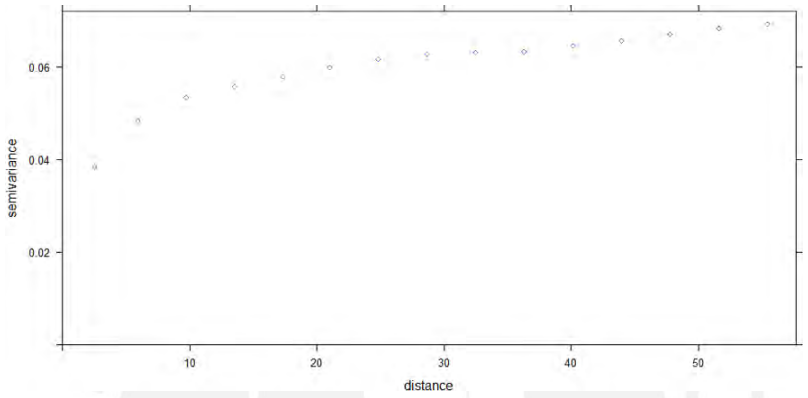

Figura 5.5: Semivariograma emp´ırico calculado a partir de las PCF muestreadas.

Hasta este punto se ha creado el semivariograma emp´ırico o experimental, sin embargo, esta estructura se puede modelar utilizando un modelo específico teórico. En este caso se evaluó el modelo exponencial, esférico y Gaussiano que son los más comunes. En las Figuras 5.6 observamos cada uno de estos modelos ajustados a los datos. Se observa que el modelo exponencial se ajusta mejor, por lo cual será el seleccionado para utilizarse en el krigeaje.

Según lo que observamos en el variograma empírico ajustando un modelo exponencial hasta una distancia de 8 grados hay dependencia espacial fuerte entre observaciones. El efecto pepita es pequeño por lo tanto la varianza debido a efectos no espaciales es pequeña (5.2).

Efecto pepita (τ^2)	0.036
Rango (r)	8.48 grados
Meseta ($\tau^2 + \sigma^2$)	0.07

Cuadro 5.2: Partes del semivariograma

Debido al elevado tiempo de procesamiento que toma realizar el krigeaje con todos los datos ($n = 8696$) y que nuestro objetivo no es predecir con este método, si no solo tener una noción de como luce el mapa interpolado en toda la superficie en estudio,

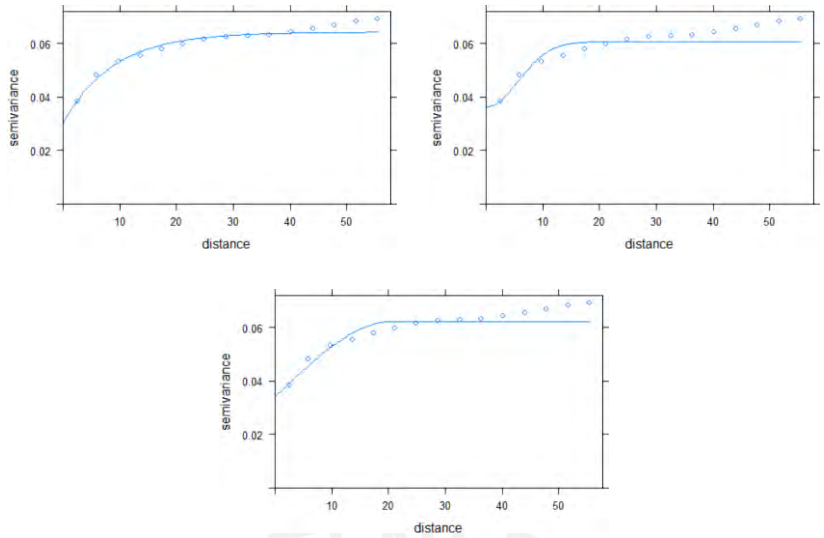

Figura 5.6: Ajuste del Modelo Exponencial (izquierda), Gaussiano (derecha) y Esférico (abajo).

se tomó una muestra aleatoria sin reemplazamiento de 3000 datos y se realizó el procedimiento de krigeaje solo con estos datos. Luego se creó una malla regular, la cual contiene un mapa de los locales sobre los cuales el krigeaje predecirá el valor de la PCF. Con los resultados del krigeaje ordinario se graficó un mapa de interpolación de la PCF. La Figura 5.7 muestra a la izquierda el mapa de predicciones de PCF con la muestra de 3000 posiciones usando el krigeaje, y a la derecha muestra el mapa de interpolación de todos los datos originales observados de la PCF. Las regiones azules indican una deforestación completa ($PCF = 0$), coincidiendo con el mapa original, lo mismo con las regiones completamente deforestadas (rojas) ($PCF = 1$).

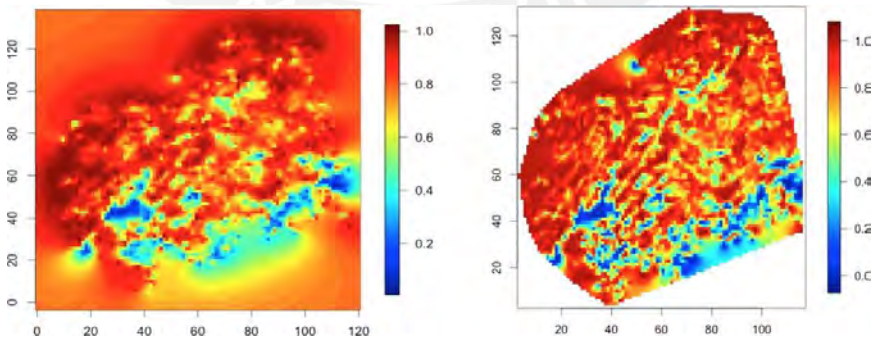

Figura 5.7: Mapa de interpolación de la predicción de PCF (izquierda). Mapa de interpolación de los datos observados (derecha).

Si bien localmente el krigeaje no ha realizado una predicción adecuada de la PCF (Figura 5.8), podemos observar que gráficamente, el krigeaje ordinario (con el modelo exponencial) ha hecho una predicción razonable aproximandose al mapa original.

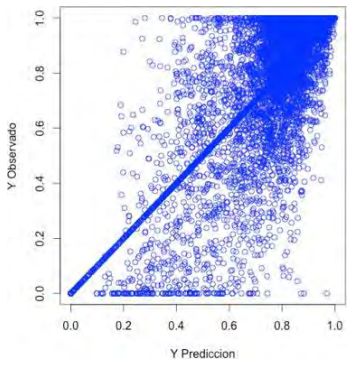

Figura 5.8: Evaluación del krigeaje ordinario. Comparación de las PCF observadas y predichas.

5.3. Modelamiento de la PCF

Sea Y_i la variable aleatoria que representa la proporción de cobertura forestal (PCF) en la celda con centroide s_i, $i = 1...n$. Se asume que Y_i sigue una distribución beta inflacionada en cero y uno, con f.d.p de acuerdo a la Ecuación 3.6. Los parámetros de la distribución se pueden asociar con variables explicativas utilizando una funcion de enlace logística. Por lo tanto, la ecuación (4.1) para el modelo espacial queda específicada de la siguiente manera:

$$
\begin{aligned}
&f_Y(y_i \mid p_0, p_1, \mu, \varphi) = a(1-\gamma)\delta_0 + a\gamma\delta_1 + (1-a)b(y_i \mid \mu, \varphi)I_{[0<y_i<1]},\\
&\text{logistica}(a(s_i)) = \beta_0^{(1)} + \beta_1^{(1)} * latitud(s_i) + \beta_2^{(1)} * longitud(s_i),\\
&\text{logistica}(\gamma(s_i)) = \beta_0^{(2)} + \beta_1^{(2)} * latitud(s_i) + \beta_2^{(2)} * longitud(s_i),\\
&\text{logistica}(\mu(s_i)) = \beta_0^{(3)} + \beta_1^{(3)} * N(s_i) + \beta_2^{(3)} * R(s_i) + f(s_i),
\end{aligned}
$$

donde la $latitud$ y $longitud$ son las coordenadas de cada local s_i, N es la densidad poblacional y R es la diferencia entre la altitud máxima y mínima, definidas en cada local s_i, variables explicativas descritas en la sección 5.1. Además, $\beta_0^{(k)}, \beta_1^{(k)}, \beta_2^{(k)}$ son los coeficientes de regresión de cada parámetro para $k = 1, 2, 3$. Además, se asume que $f(s_i)$ es un efecto espacial aleatorio, tal que $f(s) = (f(s_1), ..., f(s_n))^T \sim NNGP(0, \tilde{C})$. El modelo geoestadístico es completamente definido en el Capítulo 3.

Con el objetivo de tener un modelo de referencia con el cual comparar, se ajustó un modelo con la misma especificación en la ecuación 5.1 pero sin incluir el efecto espacial aleatorio. La media en el local s_i del modelo de referencia queda especificada de la

siguiente forma:

$$f_Y(y_i \mid p_0, p_1, \mu, \varphi) = a(1-\gamma)\delta_0 + a\gamma\delta_1 + (1-a)b(y_i \mid \mu, \varphi)I_{[0<y_i<1]},$$
$$\text{logistica}(a(s_i)) = \beta_0^{(1)} + \beta_1^{(1)} * latitud(s_i) + \beta_2^{(1)} * longitud(s_i),$$
$$\text{logistica}(\gamma(s_i)) = \beta_0^{(2)} + \beta_1^{(2)} * latitud(s_i) + \beta_2^{(2)} * longitud(s_i),$$
$$\text{logistica}(\mu(s_i)) = \beta_0^{(3)} + \beta_1^{(3)} * N(s_i) + \beta_2^{(3)} * R(s_i).$$

Así el modelo de referencia es útil para verificar la importancia de incluir la dependencia espacial para modelar estos datos de PCF.

5.4. Resultados

La base de datos está compuesta por 8696 registros, sin embargo, no todos fueron utlizados ya que la estimación de los parámetros podría tomar mucho tiempo dependiendo del número de vecinos utilizados. Para determinar la viabilidad de la inferencia con el modelo propuesto, primero se determinó el número de datos que se usarán para ajustar el modelo. Para ello se ajustó el modelo geoestadístico utilizando M = 5 vecinos, tomando como referencia el modelo propuesto por Nishii and Tanaka (2013), solo que el NNGP toma en cuenta un rango de mayor tamaño aún considerando cinco vecinos. Se analizaron los tiempos de ejecución con 1500 y 2000 datos y se pudo observar que con tan solo incluir 500 datos más el modelo tardó en ajustar 12 horas más.

Cuadro 5.3: Modelo Geoestadístico de 5 vecinos: comparación de tiempos entre 1500, 2000 y 3000 datos

Datos	Tiempo (horas)
1500	7.32
2000	19.71
3000	32.4

Se decidió tomar un tamaño de muestra aleatoria igual a 3000 datos de PCF, verificandose que la proporción de ceros y unos se mantenga similar a la distribución original (Cuadro 5.4).

Intacto (CC) ($F = 1$)	556	18.53 %
Deforestado (CD) ($F = 0$)	62	2.07 %
Parcialmente cubierto (PC) $PCF \in (0, 1)$	2382	79.40 %

Cuadro 5.4: Porcentaje de datos por categor´ıa

Como segundo paso se debe determinar el número de vecinos. Para ello se ajustó

el modelo geoestadístico utilizando diferentes cantidades de vecinos con el objetivo de comparar sus resultados y tiempos de ejecución. Para este punto se utilizó una muestra de tan solo 1500 datos. Además, se busca resaltar que el NNGP se aproxima a un proceso espacial Gaussiano utilizando grafos dispersos, es decir, con la menor cantidad de vecinos posible. En el Cuadro 5.5 se observan los resultados obtenidos para cada cantidad de vecinos.

Cuadro 5.5: Comparación del RECM y tiempo de ejecución para cada modelo

Nro. Vecinos (M)	RECM	Tiempo (horas)
3	0.1175	1.01
5	0.1165	2.64
10	0.1163	7.32
20	0.1169	21.21
30	0.1175	49.61

La principal y más resaltante diferencia entre cada modelo es el incremento en tiempo de ejecución, mientras que la estadística de ajuste RECM se mantiene similar en todos los casos, la cual se incrementa mínimamente con el máximo de vecinos. Se observa que un incremento del 100 % en el número de vecinos supone un incremento del 300 % en tiempo de ejecución y un incremento del 300 % en el número de vecinos supone un incremento del 700 % en tiempo de ejecución. Es así que el modelo con 30 vecinos tomó casi 50 horas en ajustarse mientras que el modelo con 10 vecinos tomó solo 7 horas. En términos del RECM los modelos de 5 y 10 vecinos dieron mejores resultados (0.1165 y 0.1163 respectivamente).

Con estos resultados se optó por modelar utilizando 5 vecinos pero con la muestra de 3000 datos. Se ajustaron tanto el modelo de referencia como el modelo geoestadístico utilizando la misma muestra de datos.

Para mostrar la convergencia a la distribución conjunta a posteriori, la figura 5.9 muestra las cadenas generadas para cada parámetro y algunos efectos espaciales para el modelo geoestadístico ajustado con $M = 5$ vecinos. Se puede verificar la convergencia de las dos cadenas del mismo parámetro, ya que convergen al mismo valor.

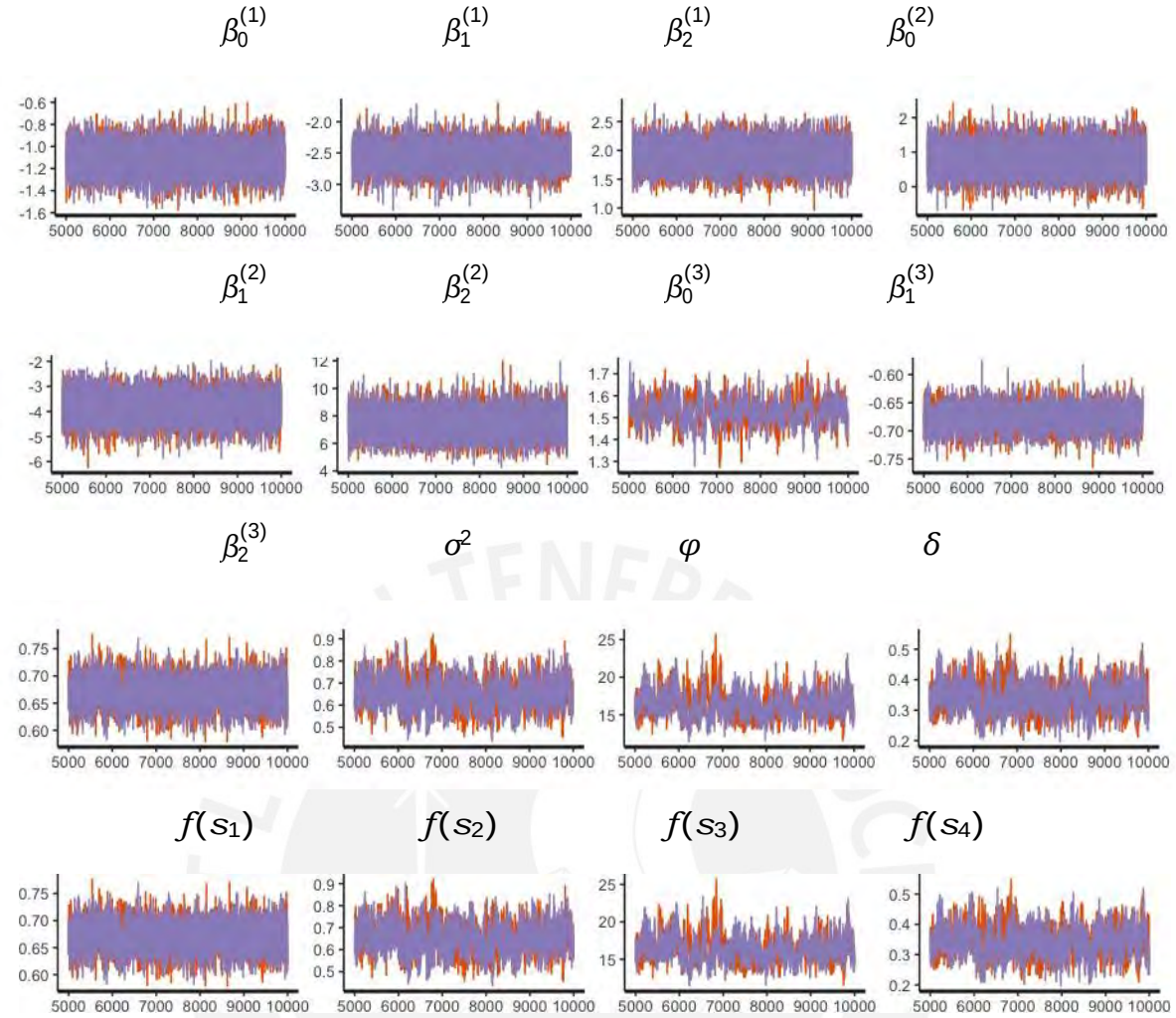

Figura 5.9: Convergencia de las cadenas de los parámetros estimados y algunos efectos espaciales

Analizando los resultados obtenidos para ambos modelos, con respecto al RECM del modelo de referencia, este fue de 0.170, mayor valor que el del modelo geoestadístico (de 5 vecinos); por lo cual en base a este criterio concluimos que la estimación a posteriori de PCF en el modelo con dependencia espacial es mejor que en el modelo de referencia. En el Cuadro 5.7 se observan los resultados de las estimaciones puntuales e intervalares a posteriori del modelo de referencia (sin efectos espaciales). Por otro lado, en el Cuadro 5.6 se observa la media y la desviación estandar a posteriori, así como el intervalo de credibilidad al 95 % de los parámetros estimados para el modelo geoestadístico ajustado con M = 5 vecinos. Como era de esperarse el modelo de referencia y el modelo geoestadístico se diferencian en los coeficientes de regresión de μ y en el parámetro de precisión φ.

Cuadro 5.6: Resumen: media a posteriori, intervalo de credibilidad 95 % para los parámetros del modelo geoestadístico.

Parámetros	Media	d.e.	2.50 %	97.50 %
$\beta_0^{(1)}$	-1.113	0.136	-1.381	-0.847
$\beta_1^{(1)}$	-2.550	0.210	-2.968	-2.144
$\beta_2^{(1)}$	1.908	0.238	1.439	2.379
$\beta_0^{(2)}$	0.811	0.433	-0.024	1.661
$\beta_1^{(2)}$	-3.838	0.572	-4.971	-2.734
$\beta_2^{(2)}$	7.428	0.986	5.606	9.449
$\beta_0^{(3)}$	1.530	0.061	1.407	1.649
$\beta_1^{(3)}$	-0.678	0.021	-0.718	-0.637
$\beta_2^{(3)}$	0.671	0.027	0.617	0.723
φ	16.299	1.706	13.452	20.096
σ^2	0.646	0.065	0.526	0.780
δ	0.342	0.048	0.254	0.439

Cuadro 5.7: Resumen: media a posteriori, desviación estandar a posteriori, intervalo de credibilidad 95 % para los parámetros del modelo de referencia

Parámetros	Media	d.e.	2.50 %	97.50 %
$\beta_0^{(1)}$	-1.113	0.136	-1.381	-0.847
$\beta_1^{(1)}$	-2.550	0.210	-2.968	-2.144
$\beta_2^{(1)}$	1.908	0.238	1.439	2.379
$\beta_0^{(2)}$	0.811	0.433	-0.024	1.661
$\beta_1^{(2)}$	-3.838	0.572	-4.971	-2.734
$\beta_2^{(2)}$	7.428	0.986	5.606	9.449
$\beta_0^{(3)}$	1.320	0.0213	1.278	1.362
$\beta_1^{(3)}$	-0.721	0.018	-0.756	-0.686
$\beta_2^{(3)}$	0.579	0.022	0.534	0.623
φ	5.739	0.169	5.419	6.074

En la Figura 5.10 se observan los valores estimados de PCF por el modelo de referencia (beta inflacionado sin efecto espacial) versus los valores reales de PCF donde

se puede apreciar que no se ajustan adecuadamente a los datos.

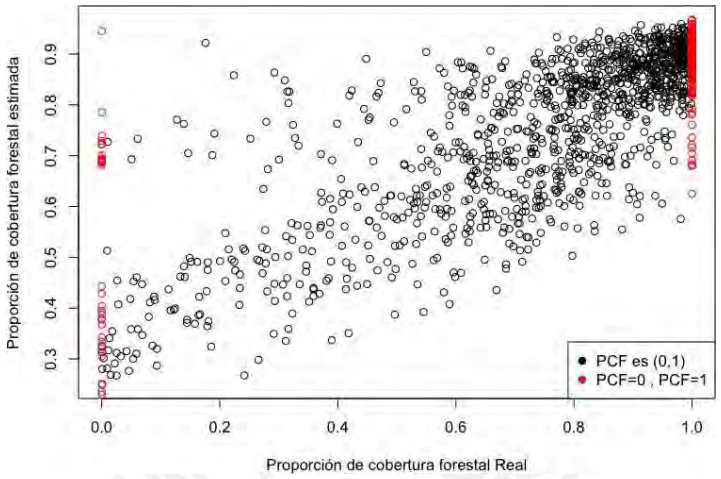

Figura 5.10: Modelo sin efecto espacial. Completamente deforestados y completamente cubiertos de arboles (puntos rojos), parcialmente cubiertos (puntos negros).

Por otro lado, en la Figura 5.11 se observan los valores estimados de PCF por el modelo geoestadístico versus los valores reales de PCF donde se puede apreciar el buen desempeño del modelo.

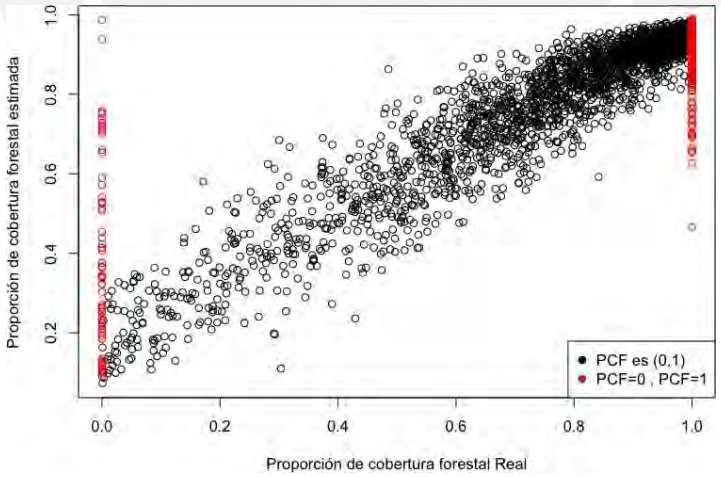

Figura 5.11: Modelo Geoestadístico usando 5 locales vecinos. Completamente deforestados y completamente cubiertos de arboles (puntos rojos), parcialmente cubiertos (puntos negros).

Dado que la única diferencia entre ambos modelos es el efecto espacial, se puede concluir que considerar la dependencia espacial en los datos es necesario dado que ha mejorado la precisión de la estimación de la PCF.

En la Figura 5.12 comparamos el mapa de Hiroshima reconstruído con los datos reales de PCF y el mapa reconstruído con los datos estimados por el modelo geoestadístico beta inflacionado. Visualmente se aprecia que el modelo se ha ajustado correctamente detectando mejor las regiones con alta PCF.

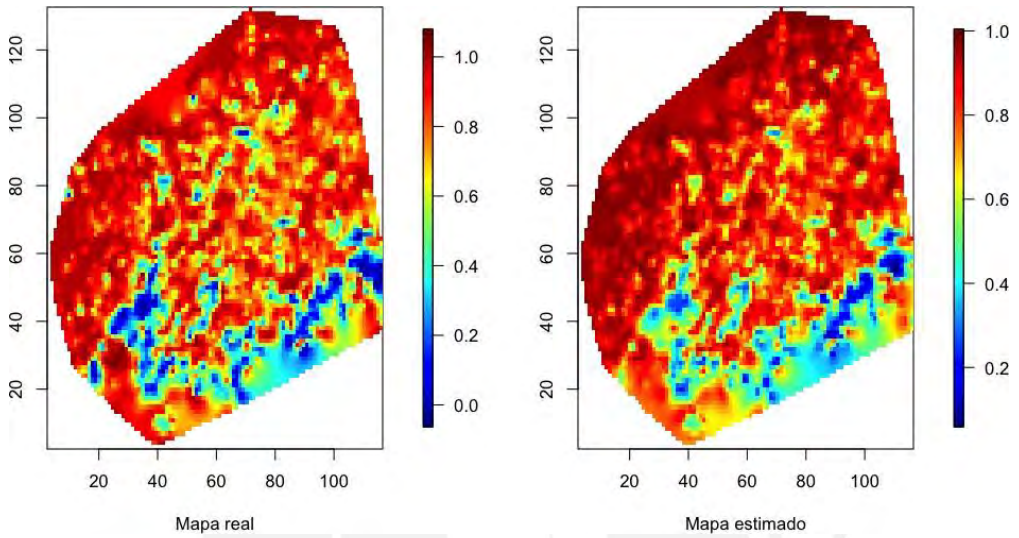

Figura 5.12: Izquierda: Mapa de interpolación de PCF observado. Derecha: Mapa de estimación de PCF con el modelo geoestadístico usando 5 locales vecinos.

Para interpretar los coeficientes estimados en la regresión es importante recordar que la esperanza que se modela está definida como $E(Y) = a\gamma + (1-a)\mu$, entonces los coeficientes asociados a a, parámetro que nos indica la probabilidad de que Y sea binaria (cero o uno), son $\beta_1^{(1)} = -2.55$ para la latitud y $\beta_2^{(1)} = 1.91$ para la longitud. Estas variables son las coordenadas de un local, por lo cual nos indican el incremento (en el caso de la longitud) y la reducción (en el caso de la latitud) de la probabilidad de que una región específica se encuentre en cualquiera de los dos estados extremos (completamente cubierta o completamente deforestada). Podemos decir que se espera que en zonas de mayor latitud y menor longitud hayan más locales completamente cubiertos o completamente deforestados. Luego, los coeficientes asociados a γ complementan lo anterior, ya que este parámetro nos indica la probabilidad de que, dado que un local está en un estado extremo, este se encuentre completamente cubierto de árboles. Los coeficientes estimados para la latitud y la longitud son $\beta_1^{(2)} = -3.84$ y $\beta_2^{(2)} = 7.43$, el primero contribuye de forma negativa y el segundo de forma positiva, lo cual indica que a mayor latitud y menor longitud hay mayor probabilidad de encontrar un local completamente deforestado. Por otro lado, los coeficientes de regresión estimados para la de densidad poblacional y la diferencia de altitudes máxima y mínima en el local son $\beta_1^{(3)} = -0.68$ y $\beta_2^{(3)} = 0.67$ respectivamente. Dado que el

enlace utilizado fue el logit, se debe aplicar la función exponencial para obtener el incremento/decremento en los odds asociado a cada variable. En el caso de la densidad poblacional, esta fue transformada a logaritmo antes de entrar al modelo, por lo cual su coeficiente ya explica el porcentaje de cambio en la variable dependiente, lo cual se traduce en una reducción del 30 % de la proporción media de estar parcialmente cubierto de árboles por cada unidad de incremento de la densidad poblacional. Esto tiene sentido ya que a mayor cantidad de población se va necesitando de mayor espacio para infraestructura, reduciendo el número de áreas verdes. Con respecto a la otra variable, de diferencias de altitudes, se tiene que $e^{0.67} = 1.96$ lo cual se traduce en un incremento del 95 % de la proporción media de estar parcialmente cubierto de árboles por cada unidad de incremento en la diferencia de altitudes. Esto nos indica que en los terrenos más llanos tienden a haber menos arboles que en los terrenos más irregulares. Por otro lado, con respecto a los parámetros asociados a la dependencia espacial, a partir del $\delta = 0.34$ podemos calcular el rango efectivo $r* = \frac{2}{\delta}$, el cual nos indica que hasta una distancia de 5.8 agrados existe dependencia espacial entre los datos. Con respecto a σ^2 y φ son parámetros que indican la variabilidad de los datos, en particular el $\varphi = 16.29$ del modelo geoestadístico es mayor que el del modelo de referencia, lo cual indica que la variabilidad restante no espacial es menor.

5.5. Predicción

Para la predicción se seleccionaron aleatoriamente 1000 datos que no fueron utilizados en el modelamiento, con el objetivo de evaluar el ajuste del modelo en locales nuevos. Dado un nuevo local t, se intentará predecir la PCF con el conjunto de variables disponibles, parámetros calculados a posteriori y el efecto espacial (f_t) a partir de sus vecinos más cercanos. Para encontrar los vecinos más cercanos se utilizó la distancia euclideana, a través de un arbol k- dimensional para realizar la búsqueda de los vecinos. Por otro lado, se tiene la distribución condicional completa $f_t \mid y \sim N(B_t f_{N(t)}, F_t)$, donde $B_t = C_{t,N(t)} C_{N(t)}^{-1}$ y $F_t = C(t, t) - C_{t,N(t)} C_{N(t)}^{-1} C_{t,N(t)}^{T}$, siendo C la función de covarianza exponencial. El algoritmo utilizado para la predicción se encuentra en el Apendice B. En la Figura 5.13 se observan las predicciones y sus valores reales. El ajuste obtenido ha sido razonable dado que los datos se mantienen en la diagonal aproximadamente.

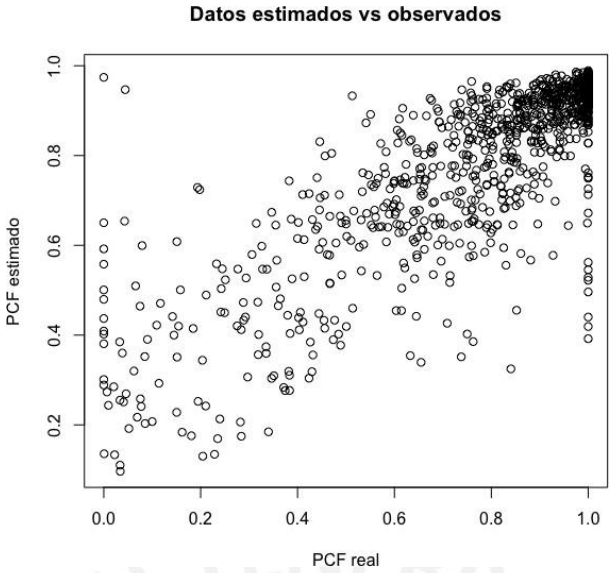

Figura 5.13: Validación del modelo geoestadístico con M=5 locales vecinos. PCF predicho vs PCF observado en 1000 locales

Capítulo 6

Conclusiones

6.1. Comentarios finales

En la presente tesis se ha desarrollado un modelo geoestadístico beta inflacionado en cero y uno, en particular para modelar la proporción de cobertura forestal en los bosques de Hiroshima.

Es importante resaltar que a la fecha no existen trabajos que usen los NNGP para variables que siguen una distribución beta inflacionada en ceros y unos. Esta es una de las contribuciones de este trabajo, debido a que el modelo geoestadístico propuesto requiere de una elevada capacidad computacional, debido a la cantidad de parámetros del modelo y la dependencia espacial entre las variables observadas. Como consecuencia un punto clave en la tesis fue usar procesos Gaussianos de vecinos más cercanos (NNGP) para estimar los efectos espaciales aleatorios. Esto se pudo verificar al ajustar el modelo propuesto para distintas cantidades de vecinos, entre los cuales la mayor diferencia fue el tiempo de ejecución, conforme los vecinos aumentaban el tiempo de ejecución aumentaba exponencialmente pero la precisión se mantenía.

Por otro lado para evaluar la eficiencia del modelo propuesto, se aplicó a una base de datos "grande" de cobertura forestal. Desde un punto de vista ecológico, es importante este tipo de estudios para evaluar qué variables pueden influir en la deforestación, así como predecir las regiones de deforestación. Esta información es importante para proponer diferentes políticas de cuidado del medio ambiente. A través de la inclusión de efectos espaciales aleatorios en el modelo se ha conseguido modelar la dependencia espacial. Esto se ha podido verificar al comparar un modelo de referencia (sin efectos espaciales) con el modelo propuesto, encontrándose un mejor ajuste con este último. Incluir este tipo de efectos, ayuda a ajustar adecuadamente la variable PCF, sobre todo cuando no hay covariables dificiles de recolectar que expliquen la dependencia espacial en la PCF.

6.2. Sugerencias para investigaciones futuras

Como ideas para futuras investigaciones se tiene lo siguiente:

- Incluir efectos espaciales en los otros parámetros del modelo y no solo en la media de la variable acotada en (0,1).
- Usar otras reparametrizaciones del modelo beta inflacionado en cero y uno para proponer nuevos modelos geoestadísticos.
- Comparar el modelo geoestadístico beta inflacionado en cero y uno usando NNGP con otros tipos de modelos geoestadísticos por ejemplo usando ecuaciones diferenciales parciales estocásticas.

Bibliografía

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Bayes, C. L. and Valdivieso, L. (2016). A beta inflated mean regression model for fractional response variables. *Journal of Applied Statistics*, 43(10):1814–1830.

Chakravarty, S., Ghosh, S., Suresh, C., Dey, A., and Shukla, G. (2012). Deforestation: causes, effects and control strategies. In *Global perspectives on sustainable forest management*. InTech.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

Fernández, R., Bayes, C. L., and Valdivieso, L. (2018). A beta-inflated mean regression model with mixed effects for fractional response variables. *Journal of Statistical Computation and Simulation*, pages 1–22.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC.

Ghosh, J. K., Delampady, M., and Samanta, T. (2007). *An introduction to Bayesian analysis: theory and methods*. Springer Science & Business Media.

Kalhori, L. and Mohhamadzadeh, M. (2017). Spatial beta regression model with random effect. *Journal of Statistical Research of Iran JSRI*, 13(2):215–230.

Koistinen, P. (2010). Monte carlo methods with an emphasis on bayesian computation. *Lecture notes*.

Lagos-Alvarez, B. M., Fustos-Toribio, R., Figueroa-Zuniga, J., and Mateu, J. (2017). Geostatistical mixed beta regression: a bayesian approach. *Stochastic Environmental Research and Risk Assessment*, 31(2):571–584.

Nishii, R. and Tanaka, S. (2013). Modeling and inference of forest coverage ratio using zero-one inflated distributions with spatial dependence. *Environmental and ecological statistics*, 20(2):315–336.

Ospina, R. and Ferrari, S. L. (2010). Inflated beta distributions. *Statistical Papers*, 51(1):111.

Ospina, R. and Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6):1609–1623.

Parker, A. J., Bandyopadhyay, D., and Slate, E. H. (2014). A spatial augmented beta regression model for periodontal proportion data. *Statistical Modelling*, 14(6):503–521.

Puyravaud, J.-P. (2003). Standardizing the calculation of the annual rate of deforestation. *Forest Ecology and Management*, 177(1-3):593–596.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

Apéndice A: Cadenas del estudio de simulación

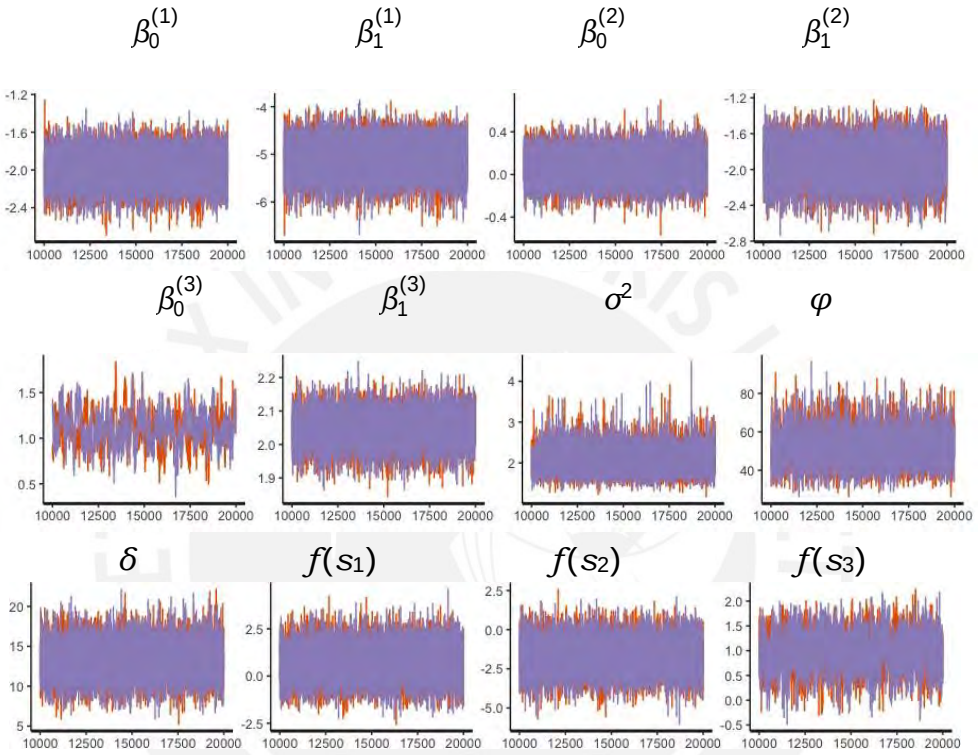

Figura 1: Convergencia de las cadenas de los parámetros y algunos efectos espaciales para M=3 locales vecinos

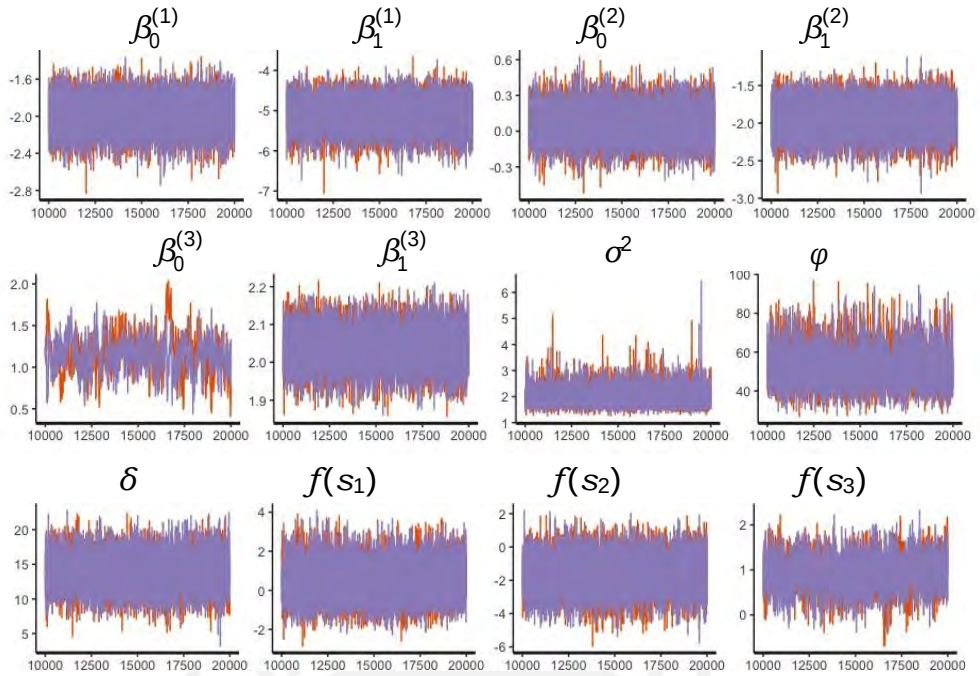

Figura 2: Convergencia de las cadenas de los parámetros y algunos efectos espaciales para M=7 locales vecinos

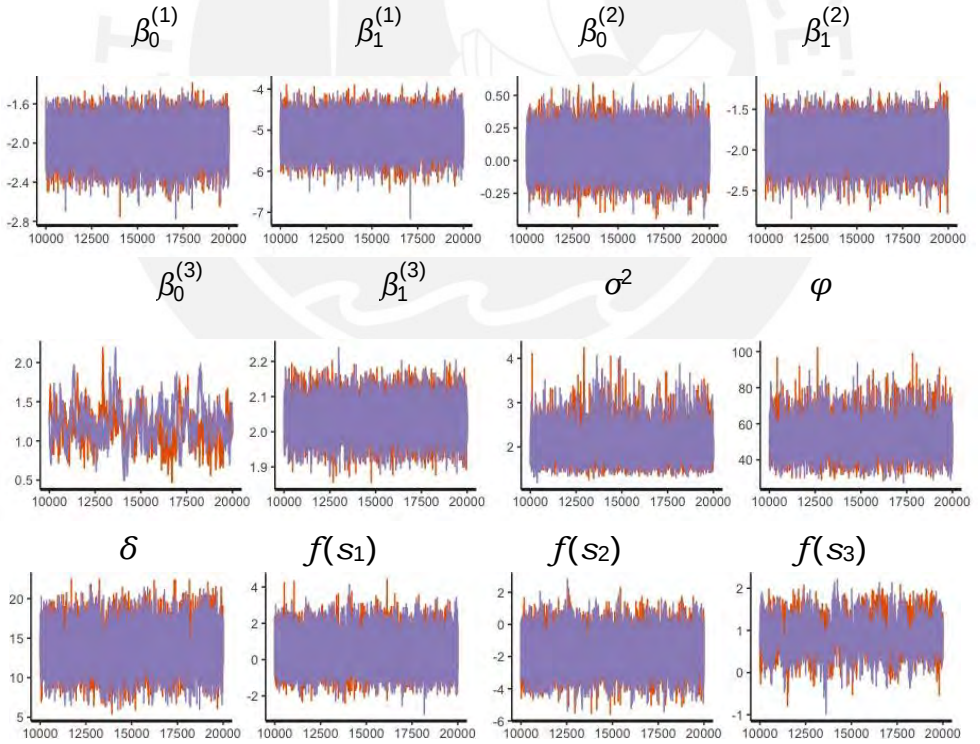

Figura 3: Convergencia de las cadenas de los parámetros y algunos efectos espaciales para M=10 locales vecinos

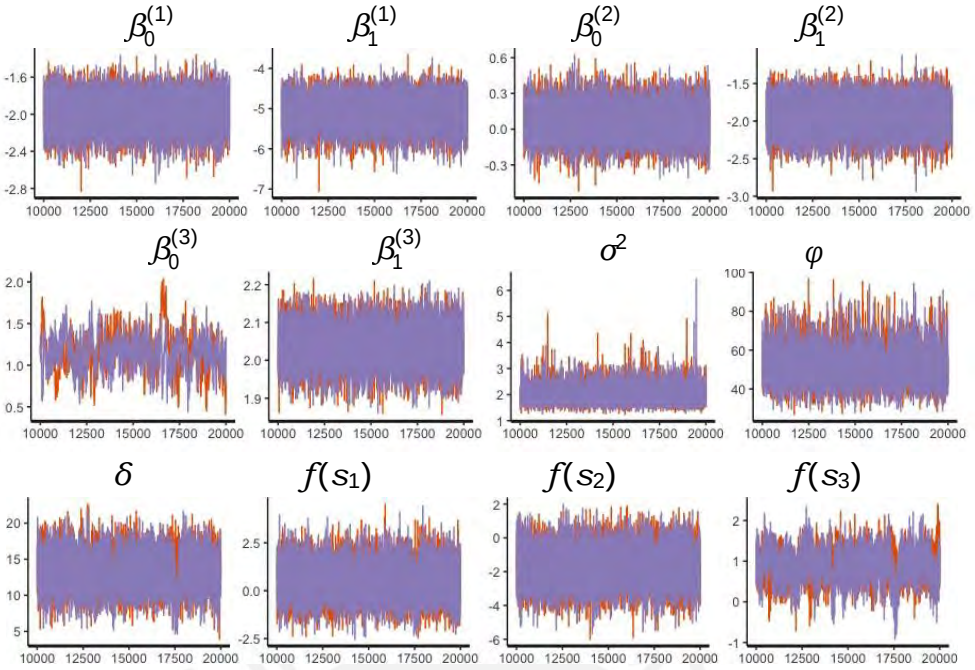

Figura 4: Convergencia de las cadenas de los parámetros y algunos efectos espaciales para M=15 locales vecinos

Apéndice B: Código de Aplicación

```
library(rstan)
library(fields)
library(spNNGP)
library(loo)
library(MBA)
source("NNmatrix.R")

inv_logit <- function(x){
  exp(x) / (1 + exp(x))
}

#estandarización por rango
range_scale <- function(x){(x-min(x))/(max(x)-min(x))}

#lectura de datos
datos = read.csv('hiroshima.csv')
names(datos) <- c('Y','X','F', 'N', 'R', 'R2')

N=3000
#Se toma una muestra de los datos
set.seed(199118)
indsamp <- sample(1:dim(datos)[1],N)
datos.samp = datos[indsamp,]
mean(datos.samp$F)

#variables estandarizadas
datos.samp$s_N=scale(log(datos.samp$N+1),center=TRUE,scale=TRUE)
datos.samp$s_R=scale(datos.samp$R,center=TRUE,scale=TRUE)
datos.samp$r_X=range_scale(datos.samp$X)
datos.samp$r_Y=range_scale(datos.samp$Y)

#ploteo de las coordenadas
coords <- cbind(datos.samp$X, datos.samp$Y)
plot(coords)

## Inicio del NNGP
## Construcción de la matriz de vecinos
M = 5   # numero de puntos de los que depende
NN.matrix <- NNMatrix(coords = coords, n.neighbors = M, n.omp.threads = 2)
```

```
str(NN.matrix)


par(mfrow=c(1,1))
Check_Neighbors(NN.matrix$coords.ord, n.neighbors = M, NN.matrix, ind = 100)

## Inferencia bayesiana ----
P = 1                 # number of regression coefficients
uB = rep(0, P + 1)    # mean vector in the Gaussian prior of beta
VB = diag(P + 1)*1000 # covariance matrix in the Gaussian prior of beta

ss = 3 * sqrt(2)      # scale parameter in the normal prior of sigma
st = 3 * sqrt(0.1)    # scale parameter in the normal prior of tau
ap = 3; bp = 0.5      # shape and rate parameters in the Gamma prior of phi

options(mc.cores = parallel::detectCores())

#datos para correr en stan
data <- list(N = nrow(datos.samp), M = M, P = P,
             y = datos.samp$F[NN.matrix$ord],
             x = datos.samp$s_N[NN.matrix$ord],
             z=datos.samp$s_R[NN.matrix$ord],
             rx=datos.samp$r_X[NN.matrix$ord],
             ry=datos.samp$r_Y[NN.matrix$ord],# sorted Y and X
             NN_ind = NN.matrix$NN_ind,
             NN_dist = NN.matrix$NN_dist,
             NN_distM = NN.matrix$NN_distM,
             uB = uB, VB = VB, ss = ss, ap = ap, bp = bp)


# Parámetros
parameters <- c("coef_alpha","coef_gamma","coef_mu",
 "sigmasq", "phi0", "phi", "w")

start.time <- Sys.time()
# Modelo Stan
samples_w <- stan (
  file = "nngp_latent_beta_inf_v4.stan",
  data = data,
  pars = parameters,
  iter = 10000,
  chains = 2,
  thin = 1,
  seed = 123,
  verbose =TRUE,
  control = list(max_treedepth = 15)
)
```

```
end.time <- Sys.time()
end.time

#Código en Stan

/* Latent NNGP model: implementación tomada de
https://mc-stan.org/users/documentation/case-studies/nngp.html*/
functions{
    real nngp_w_lpdf(vector w, real sigmasq, real phi, matrix NN_dist,
                     matrix NN_distM, int[,] NN_ind, int N, int M){

        vector[N] V;
        vector[N] I_Aw = w;
        int dim;
        int h;

        for (i in 2:N) {

            matrix[ i < (M + 1)? (i - 1) : M, i < (M + 1)? (i - 1): M]
            iNNdistM;
            matrix[ i < (M + 1)? (i - 1) : M, i < (M + 1)? (i - 1): M]
            iNNCholL;
            vector[ i < (M + 1)? (i - 1) : M] iNNcorr;
            vector[ i < (M + 1)? (i - 1) : M] v;
            row_vector[i < (M + 1)? (i - 1) : M] v2;
            dim = (i < (M + 1))? (i - 1) : M;
            if(dim == 1){iNNdistM[1, 1] = 1;}
            else{
                h = 0;
                for (j in 1:(dim - 1)){
                    for (k in (j + 1):dim){
                        h = h + 1;
                        iNNdistM[j, k] = exp(- phi * NN_distM[(i - 1), h]);
                        iNNdistM[k, j] = iNNdistM[j, k];
                    }
                }
                for(j in 1:dim){
                    iNNdistM[j, j] = 1;
                }
            }

            iNNCholL = cholesky_decompose(iNNdistM);
            iNNcorr = to_vector(exp(- phi * NN_dist[(i - 1), 1:dim]));

            v = mdivide_left_tri_low(iNNCholL, iNNcorr);

            V[i] = 1 - dot_self(v);
```

```
            v2 = mdivide_right_tri_low(v', iNNCholL);

            I_Aw[i] = I_Aw[i] - v2 * w[NN_ind[(i - 1), 1:dim]];

        }
        V[1] = 1;
        return - 0.5 * ( 1 / sigmasq * dot_product(I_Aw, (I_Aw ./ V)) +
                        sum(log(V)) + N *  log(sigmasq));
    }


}


data {
  int<lower=1> N;
  int<lower=1> M;
  int<lower=1> P;

  int NN_ind[N - 1, M];
  matrix[N - 1, M] NN_dist;
  matrix[N - 1, (M * (M - 1) / 2)] NN_distM;
  vector[P + 1] uB;
  matrix[P + 1, P + 1] VB;
  real ss;
  real ap;
  real bp;

  vector[N] x;
  vector[N] z;
  vector[N] rx;
  vector[N] ry;
  vector<lower=0,upper=1>[N] y;
}

transformed data {
    int<lower=0, upper=1> is_discrete[N];
    int<lower=-1, upper=1> y_discrete[N];

    for (i in 1:N) {
      if (y[i] == 0) {
        is_discrete[i] = 1;
        y_discrete[i] = 0;
      } else if (y[i] == 1) {
        is_discrete[i] = 1;
        y_discrete[i] = 1;
      } else {
        is_discrete[i] = 0;
```

```
        // hack to ensure that throws error if passed to bernoulli_lpmf
        y_discrete[i] = -1;
      }
    }

}

parameters{

vector[3] coef_alpha;
vector[3] coef_gamma;
vector[3] coef_mu;
real<lower = 0> sigma;
real<lower = 0> phi0;
real<lower = 0> phi;
vector[N] w;
}

transformed parameters {

  real sigmasq = square(sigma);

  vector<lower=0>[N] alpha;
  vector<lower=0>[N] gamma;
  vector<lower=0,upper=1>[N] mu;
  vector<lower=0>[N] phi_2;
  vector<lower=0>[N] p;
  vector<lower=0>[N] q;

  for (i in 1:N) {
   alpha[i] = inv_logit(coef_alpha[1] +
   coef_alpha[2]*rx[i] + coef_alpha[3]*ry[i]);

   gamma[i] = inv_logit(coef_gamma[1] +
   coef_gamma[2]*rx[i] + coef_gamma[3]*ry[i]);

   mu[i] = inv_logit(coef_mu[1] +
   coef_mu[2] * x[i] +coef_mu[3] * z[i] + w[i]);
   phi_2[i]=phi0 ;
    p[i] = mu[i] * phi0;
    q[i] = phi0 - mu[i] * phi0;
  }

}

model{

  //prioris de los parametros
```

```
  coef_alpha ~ normal(0,1000000);
    coef_gamma ~ normal(0,1000000);
    coef_mu ~ normal(0,1000000);
    phi ~ cauchy(0,5);
    psi ~ gamma(ap, bp);
    sigma ~ normal(0, ss)

  w ~ nngp_w(sigmasq, phi, NN_dist, NN_distM, NN_ind, N, M);

  is_discrete ~ bernoulli(alpha);
  for (i in 1:N) {
  if (is_discrete[i] == 1) {
    y_discrete[i] ~ bernoulli(gamma[i]);
  } else {
    y[i] ~ beta(p[i], q[i]);
  }
}
}

###PREDICCION
#estimados del modelo

psi.est=mean(phi[1][[1]])
sigmasq.est=mean(sigmasq[1][[1]])


datos.samp.2 = datos[-indsamp,]
N.2=1000
indsamp.2 =sample(1:dim(datos.samp.2)[1],N.2)
datos.samp.3=datos.samp.2[indsamp.2,]
dim(datos.samp.3)


datos.samp.3$s_N=scale(log(datos.samp.3$N+1),center=TRUE,scale=TRUE)
datos.samp.3$s_R=scale(datos.samp.3$R,center=TRUE,scale=TRUE)

datos.samp.3$r_X=range_scale(datos.samp.3$X)
datos.samp.3$r_Y=range_scale(datos.samp.3$Y)

coords <- cbind(datos.samp$X, datos.samp$Y)

plot(coords)

M = 10   # Number of Nearest Neighbors numero de puntos de los que depende

datos.samp.3[datos.samp.3$x.coord==Xsi & datos.samp.3$y.coord==Ysi,]

Ey=numeric()
```

```
for (i in 1:dim(datos.samp.3)[1]){
  #agrega nueva coordenada
  coords.aux=rbind(coords,c(datos.samp.3$X[i],datos.samp.3$Y[i]) )

  res=nn2(coords.aux,k=M+1)

  vecinos=coords.aux[res$nn.idx[3001,2:(M+1)],]
  s0.vecinos=rbind(vecinos,coords.aux[3001,,drop=FALSE])

  dist.aux = as.matrix(dist(s0.vecinos))
  #distancia de la observacion a todos los dem?s vecinos
  dist.s0.vecinos= as.matrix(dist.aux[M+1,1:M,drop=FALSE])
  dist.vecinos=as.matrix(dist(vecinos))

  C.s0.vecinos=sigmasq.est*exp(-psi.est*dist.s0.vecinos)
  C.vecinos=sigmasq.est*exp(-psi.est*dist.vecinos)

  C.s0.vecinos
  C.vecinos

  F.s0=sigmasq.est-C.s0.vecinos%*%solve(C.vecinos)%*% t(C.s0.vecinos)
  B.s0=C.s0.vecinos%*%solve(C.vecinos)
  pre=numeric()

  for (j in 1:M){
    Xsi=coords.aux[res$nn.idx[3001,j+1],][1]
    Ysi=coords.aux[res$nn.idx[3001,j+1],][2]
    pre[j]=datos.est[datos.est$x.coord==Xsi & datos.est$y.coord==Ysi,]$w.est
  }


  ws=rnorm(10000,B.s0%*%pre,F.s0)
  w.s0=mean(ws)

  alpha.est <- inv_logit(a_b0.est + a_b1.est*datos.samp.3$r_X[i] +
    a_b2.est*datos.samp.3$r_Y[i])
  gamma.est <- inv_logit(g_b0.est + g_b1.est*datos.samp.3$r_X[i] +
   g_b2.est*datos.samp.3$r_Y[i])
  mu.est <- inv_logit(m_b0.est + m_b1.est * datos.samp.3$s_N[i] +
   m_b2.est*datos.samp.3$s_R[i] + w.s0)
  Ey[i] = alpha.est*gamma.est + (1-alpha.est)*mu.est
}
```