APLICACIÓN DE TÉCNICAS GEO-ESTADÍSTICAS PARA ESTIMAR PREVALENCIA DE ROYA EN CAFÉ

Por
Abner J. Ortiz Camacho

Tesis sometida en cumplimiento parcial de los requerimientos para el grado de
MAESTRÍA EN CIENCIAS
en
MATEMÁTICAS (ESTADÍSTICA)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ
2014

Aprobada por:

Raúl E. Macchiavelli, Ph.D. Presidente, Comité Graduado	Fecha
Edgardo Lorenzo González, Ph.D. Miembro, Comité Graduado	Fecha
Olgamary Rivera Marrero, Ph.D. Miembro, Comité Graduado	Fecha
Jaime Acosta, Ph.D. Representante de Estudios Graduados	Fecha
Omar Colón Reyes, Ph.D. Director del Departamento	Fecha

Abstract of Thesis Presented in Partial Fulfillment of

the Requirements for the Degree of Master of Science

APPLICATION OF GEO-STATISTICAL TECHNIQUES TO ESTIMATE RUST PREVALENCE IN COFFEE

By: Abner J. Ortiz Camacho

Chair: Raúl E. Macchiavelli

Major Department: Department of Mathematical Sciences

Two geo-statistical generalized linear models were fitted in a lot of coffee with a high incidence of rust. The models assumed normal distribution and binomial distribution, using different correlation functions and empirical semivariograms. Comparing the values of the AIC, BIC, mean estimation, correlation parameters estimation, predicted incidence maps, residuals maps, and the sums of the square error, the normal exponential model is the best fit to data. This model has a smaller sum of squared errors than the binomial exponential model. Furthermore, the predicted incidence by this model is similar to the observed incidence of the disease. The normal exponential model provides a clearer picture to understand how the disease spreads in the region of study.

Resumen de Tesis Presentada como Requisito Parcial de los

Requerimientos para el Grado de Maestría en Ciencias

APLICACIÓN DE TÉCNICAS GEO-ESTADÍSTICAS PARA ESTIMAR PREVALENCIA DE ROYA EN CAFÉ

Por: Abner J. Ortiz Camacho

Consejero: Raúl E. Macchiavelli

Departamento: Departamento de Ciencias Matemáticas

En un lote de café con una alta variedad de incidencia de roya se ajustaron dos modelos lineales generalizados geo-estadísticos. En los modelos se asumieron distribución normal y distribución binomial, utilizando distintas funciones de correlación y semivariogramas empíricos. Se compararon los valores de AIC, BIC, las estimaciones de la media, las estimaciones de los parámetros de correlación, mapas de incidencia predicha, mapa de residuales y sumas del cuadrados de error. El modelo exponencial normal fue el que mejor ajustó los datos, ya que mostró una suma de cuadrado de error menor que el modelo exponencial binomial. Además, la incidencia predicha por el modelo exponencial normal es parecida a la incidencia de la enfermedad observada. El modelo exponencial normal proporciona un panorama más claro para entender y comprender cómo la enfermedad se dispersa en la región de estudio.

Dedicado a:

Mis padres: Angela Camacho Madera y Juan Alberto Ortiz Cordero, por su amor, cooperación y apoyo en todas las etapas de mi vida.

AGRADECIMIENTOS

- Al Dr. Raúl E. Macchiavelli por su guía y apoyo.
- A todos mis compañeros estudiantes graduados, por su amistad y aliento durante la realización de mi grado de maestría.
- A Mariela Medina Pérez por su apoyo incondicional.
- Al personal administrativo, por su disposición.

Índice

1. Introducción **1**

1.1. Justificación 1

1.2. Objetivos 4

2. Revisión de Literatura **5**

2.1. Terminología y notación 5

2.1.1. Correlación espacial 7

2.1.2. El efecto pepita 7

2.1.3. Tendencia espacial 8

2.2. Modelo geo-estadístico básico 9

2.3. Modelos lineales generalizados 12

2.3.1. Modelo lineal generalizado geo-estadístico 15

2.3.2. Modelo lineal logístico binomial para datos geo-estadísticos . . . 18

2.4. Función de semivariograma y función de correlación 19

2.5. Estimación en modelos mixtos normales 30

2.5.1. Estimador de máxima verosimilitud 30

2.5.2. Estimador de máxima verosimilitud restringida 32

2.5.3. Predicción espacial 34
2.5.4. Kriging 38
2.6. Estimación y predicción bayesiana para modelos lineales generalizados geoestadísticos 44

3. Metodología **54**
3.1. Datos 54
3.2. Aplicación de técnicas geo-estadístas en base a distribución Gaussiana y en base a distribución binomial 56

4. Resultados y discusión **60**
4.1. Análisis descriptivo 60
4.2. Comparación de Modelos 62

5. Conclusión y trabajos futuros **71**
5.1. Conclusión 71
5.2. Trabajos futuros 73

Capítulo 1

Introducción

1.1. Justificación

Puerto Rico fue uno de los principales exportadores de café del mundo durante los Siglos XVIII y XIX. Hoy día, hay 21 municipios que son productores de café con 9,500 agricultores. Este producto agrícola genera 25,000 empleos directos en zonas rurales centrales y representa $33.8 millones en el ingreso bruto agrícola según García [8].

Una enfermedad importante de los cafetales en Puerto Rico es la Roya del Café, causada por el hongo *Hemileia vastatrix*. Los síntomas típicos son hojas infectadas hasta provocar su caída prematura. Esto ocasiona la reducción en la producción del café y su rendimiento al afectar la fotosíntesis, provocando defoliación y reduciendo el vigor de la planta según Macchiavelli y Rodríguez [10]. Además, Monroig y Rodríguez [14] dicen que "si hay ataques por insectos, mala fertilización y condiciones de crecimiento deficientes, los cafetos estarán en un continuo estrés y desbalance lo que afectará negativamente la producción". Para identificar la enfermedad se busca en el envés de la

hoja si existen pústulas con esporulación profusa de color amarillo-anaranjado.

La incidencia o índice de la enfermedad según Madden et al. [11] se refiere o se define como "la proporción de plantas o partes de plantas (por ejemplo hojas, ramas, etc.) que expresan las enfermas entre el número total N de plantas o partes de plantas evaluadas". Por tanto, la incidencia de la enfermedad es una variable continua. Dependiendo de la unidad de la planta que se evalua por individuos enfermos puede haber varias escalas de incidencia, según Madden et al. [11].

"La geoestadística es la aplicación de la teoría de las variables regionalizadas a la estimación de los depósitos mineros con todas las aproximaciones que esto implica" según Matheron [12]. Para la realización de esta tesis se utilizaron modelos lineales generalizados geo-estadísticos para resumir las características esenciales de la incidencia de la roya en el café en la región de estudio. Se comparan varios modelos con distintas funciones de correlación. Entonces, se obtendrá el mejor modelo que ajuste los datos para establecer la relación entre la incidencia de la enfermedad y la ubicación en el área en donde se encuentra la planta de café. Incluso, este modelo ayudará en la visualización y comprensión del comportamiento de la incidencia de la roya en la región de estudio. El área de estudio seleccionada fue la Estación Experimental de Adjuntas, Puerto Rico. La muestras son proporciones binomiales basadas en 40 hojas del tercio central del árbol.

Se ajustaron dos modelos, uno asumiendo que la incidencia de la enfermedad se distribuye normal y otro el asumiendo que la incidencia de la enfermedad se distribuye binomial. En el caso de la distribución Gaussiana se ajustarón distintas funciones de correlaciones, a partir de la cual surgen distintos tipos de modelos bajo el mismo su-

puesto. Luego, se hace un análisis para determinar, qué función de correlación ajusta mejor utilizando los criterios: Akaike, Schwartz, análisis de residuales, y suma de cuadrado de error. En el caso de la distribución binomial el número de ensayos es de 40, ya que se evaluaron 40 hojas si tenian o no presencia de enfermedad. Además, se utiliza el conocimiento previo del análisis de la función de correlación. Tomando la mejor función de correlación, se ajusta el modelo a través de la estimación bayesiana. Se hacen varias estimaciones para identificar cuál distribución a priori es más adecuada para cada parámetro del modelo. Luego se compara el mejor modelo mixto normal con el modelo binomial. Esta comparación se hace a través de mapa de incidencias predichas, mapas de residuales y suma de cuadrado de error.

Una vez obtenido el mejor modelo, éste se puede utilizar para entender y comprender cómo la enfermedad se dispersa en el cafetal. Así se puede identificar los posibles focos de la roya y poder hacer un plan efectivo para tratar la enfermedad. Además, permite tener un mejor control de la enfermedad, de modo tal que así podemos predecir el efecto económico de la enfermedad.

Mantener un control en la incidencia de la enfermedad es de suma importancia. Los agrónomos utilizan distintas formulaciones de fungicidas para el manejo de la enfermedad, que han sido identificados como eficaces para reducir la enfermedad pero no la elimina por completo. Incluso según Macchiavelli y Rodríguez [10] no todas las plantas en la plantación se ven afectadas con la misma intensidad. Además, ellos informan que debido a la distribución irregular y el hecho de que la incidencia de la enfermedad varía con los años, la decisión de utilizar fungicidas en un programa de manejo integrado de la enfermedad debe basarse en los niveles de enfermedad en meses

específicos y en determinados lugares dentro de la plantación.

1.2. Objetivos

- Aplicar distintas funciones de correlación para estimar distintos modelos asumiendo distribución Gaussiana.
- Aplicar técnicas geo-estadísticas Bayesianas asumiendo distribución binomial.
- Hallar el mejor modelo que explique la incidencia de roya en el café en la región de estudio. (distribución binomial y distribución Gaussiana).
- Crear mapas de foco de enfermedad para los dos mejores modelos.
- Aplicar software R con la librería geoR y geoRglm para los métodos geo-estadísticos.

Capítulo 2

Revisión de Literatura

2.1. Terminología y notación

"Un *proceso estocástico* es una familia o colección de variables aleatorias, los miembros del cuál pueden ser identificados o localizados de acuerdo con alguna métrica", según Schabenberger y Gotway [18]. Estos autores también definen un *proceso espacial* como una colección de variables aleatorias que están indexados por algún conjunto $D \subset \mathbb{R}^d$ que contiene coordenadas en el espacio $\mathbf{x} = [x_1, x_2, \ldots, x_d]'$. Se dice que el proceso estocástico es un *campo aleatorio* usualmente cuando la dimensión d del conjunto índice del proceso estocástico es mayor que uno. Por lo general, Diggle y Ribeiro [6] denominan señal a la naturaleza de esta superficie, que es de interés científico, aunque la propia superficie no pueda ser medida directamente.

La estadística espacial es una rama de la estadística que modela la relación del fenómeno o medida con una región de estudio determinada que está espacialmente referenciada. Usualmente a la estadística espacial se la conoce como *geo-estadística*.

Estos métodos tienen una amplia variedad de aplicaciones. El formato básico para datos univariados geo-estadísticos, según Diggle y Ribeiro [6], es $(x_i, y_i) : i = 1, \ldots, n$, donde x_i es el lugar en el espacio que se realiza la muestra. Además, y_i es un valor escalar asociado a la ubicación x_i. Los autores denominan a y como la *variable respuesta* o *medida de un fenómeno en el espacio*. Ellos consideran una característica definitiva para la geo-estadística es que la variable respuesta o la medida del fenómeno que está, por lo menos, definida a través de toda la región de estudio. A consecuencia de esto, se asume que las ubicaciones x_i son deterministas o estocásticamente independientes del proceso que genera las medidas o valores y_i según se diseñe el plan de muestreo. Por lo tanto, cada y_i es una realización de la variable aleatoria Y_i. Entonces, la distribución de Y_i es dependiente del valor en la ubicación x_i del proceso estocástico en el espacio continuo subyacente, $S(x)$, el cual no se observa directamente. Diggle y Ribeiro [6] indican que existe una distinción entre las cantidades observadas de Y_i y las sin observar o proceso latente $S(x)$.

Igualmente Diggle y Ribeiro [6] dicen que en ocasiones, según la aplicación, los datos geo-estadísticos pueden tener más de un tipo de variables explicativas. Cuando hay más de una variable de medición o respuesta, se define una variable respuesta multivariada, $y_i = \{y_{i1}, \ldots, y_{id}\}$. En cambio, si hay más de una variable explicativa en el espacio, éstas pueden estar incluidas en los datos $\{d_k(x) : x \in A\}$; a veces también llamadas *covariables*. Ellos diferencian entre los dos tipos de variables, desde el punto de vista del modelo, es que el modelo para una variable respuesta multivariada requiere la especificación de un vector del proceso estocástico sobre la región de estudio. Mientras que las variables explicativas en el espacio se considerarán cantidades deterministas sin un modelo estocástico asociado. Una consecuencia de esto es que una variable

explicativa en el espacio, debe al menos en principio, estar disponible en cualquier ubicación x_i.

2.1.1. Correlación espacial

Dentro de la aplicación geo-estadística usualmente existe correlación espacial entre las observaciones, conforme a Lawson [9]. El autor explica que esta correlación es geográfica y se relaciona con la idea básica de las localizaciones. Mientras las variables respuestas estén más cercanas en el espacio, éstas tienden a tener valores similares; sin embargo, las localizaciones distantes tienden a tener valores distintos. Una definición más formal, sea $S(\mathbf{x})$ un atributo de S que se observa en el plano de localidades en el espacio $\mathbf{x} = [x, y]'$, entonces según Schabenberger y Gotway [18] la *correlación* o *autocorrelación* espacial se refiere a la correlación entre $S(x_i)$ y $S(x_j)$ para cualquier i, j en el espacio. En presencia de correlación espacial positiva, al un par de puntos x_i y x_j estar cercanos en el espacio sus valores $S(x_i)$ y $S(x_j)$ deben ser similares.

2.1.2. El efecto pepita

Schabenberger y Gotway [18] definen el término efecto pepita ("nugget") como la magnitud de discontinuidad en el origen. Igualmente, Diggle y Ribeiro [6] definen el término efecto pepita dentro del modelo como una medición de la varianza del error τ^2 o una equivalencia de la varianza condicional de cada valor medido Y_i, dado el valor de la señal subyacente $S(x_i)$. También dicen que cuando el diseño de muestreo especifica una sola medición en cada una de las distintas ubicaciones, el efecto pepita tiene una doble interpretación. Una interpretación del efecto pepita es como un error de medición. Otra

interpretación es como la variación espacial a una escala más pequeña que la distancia más pequeña entre dos puntos en el diseño de muestreo, o una combinación de estos dos efectos. Diggle y Ribeiro [6] mencionan que los dos componentes antes mencionados sólo pueden identificarse por separado si la varianza del error de la medida se conoce, o se puede estimar directamente tomando medidas repetidas en lugares coincidentes.

2.1.3. Tendencia espacial

En la geo-estadística el término *tendencia espacial* se utiliza cuando la esperanza de la variable respuesta varía en el espacio y se especifica en función de las coordenadas, conforme a Diggle y Ribeiro [6]. Incluso, se puede decir que cuando exista cualquier tipo de media variable hay *tendencia espacial*. Estos autores mencionan que en las aplicaciones se puede elegir modelar directamente a $\mu(x)$ como función de x. Usualmente esto se hace a través de un modelo de regresión polinomial utilizando potencias y productos cruzados en coordenadas cartesianas de x como variables explicativas. Es a esto a lo que Diggle y Ribeiro [6] le llama modelos de *superficies de tendencia*. Estos autores utilizan como ejemplo $\mu(x) = \alpha + d(x)\beta$ donde $d(x)$ es una propiedad científica relevante para la localización de x. Cuando sólo se muestrea en las mismas localizaciones, los valores de la variable explicativa $d(x)$, dan lugar a los datos geo-estadíticos (x_i, y_i). Ellos dicen que hay que tener en cuenta si es necesario considerar en tratar a $d(x)$ como otra variable estocástica para ser analizada conjuntamente con el proceso estocástico señal $S(x)$, en vez de una cantidad determinada.

2.2. Modelo geo-estadístico básico

En el formato básico del modelo geo-estadístico hay que incorporar dos elementos importantes: el proceso estocástico $\{S(x) : x \in D\}$, el cual típicamente se considera ser la realización parcial del proceso estocástico $\{S(x) : x \in \mathbb{R}^2\}$ en todo el plano y una distribución multivariada para la variable aleatoria $Y = (Y_1, \ldots, Y_n)$ condicional a $S(\cdot)$ según Diggle y Ribeiro [6]. A veces, Y_i se puede considerar como la versión ruidosa de $S(x_i)$ y se asume que Y_i es condicionalmente independiente dado $S(\cdot)$.

La metodología de superficie de tendencia modela la función de la media

$$S(x) = \mathbf{X}(x)\beta + e(x), \qquad e(x) \sim (0, \Sigma(\theta))$$

con una sobreparametrización, incluyendo la función de coordenadas en el espacio $x = [x_i, y_i]'$. Utilizando como ejemplo un modelo lineal de superficie de tendencia

$$S(x) = \beta_0 + \beta_1 x_i + \beta_2 y_i + \epsilon_i, \qquad \epsilon_i \sim iid(0, \sigma^2).$$

Se puede decir que si $\mathrm{E}[S(x)] = \mu$ y $S(x)$ están correlacionados, este modelo estará incorrecto en algunas partes. En donde $\beta_0 + \beta_1 x_i + \beta_2 y_i$ no es el modelo para la media y los errores no son independientes e idénticamente distribuidos (iid). Se puede observar que cuando se realiza una sobreparametrización de la media, se observa que el modelo representa una variabilidad asociada a la estructura aleatoria en el espacio. El modelo

pretende hacerlo a través de

$$S(x) = 1\mu + e(x), \qquad e(x) \sim (0, \Sigma)$$

ó

$$S(x) = \mathbf{X}(x)\beta + \epsilon, \qquad \epsilon \sim (0, \sigma^2 I)$$

donde éstas sean representaciones *equivalentes* de la variabilidad en el espacio en un campo aleatorio. Schabenberger y Gotway [18] dicen que si $e(x)$ contiene un componente de suavidad a pequeña escala, entonces $\mathbf{X}(x)\beta$ adquiere el comportamiento del campo aleatorio local, pero ésta no es la media. En cambio si $e(x)$ contiene un efecto pepita, notamos que su varianza debe ser igual a σ^2. Si se fija la estructura de los efectos $\mathbf{X}(x)\beta$, se elimina toda la variación de suavidad a pequeña escala.

El modelo de superficie de tendencia se obtiene con la parametrización de $\mathbf{X}(x)\beta$ en términos del polinomio de superficie de tendencia en coordenadas d con grado p. En $\mathbb{R}^2$, escribimos

$$\begin{aligned} S(x) &= \mathbf{X}(x)\beta + \epsilon_i \\ \mathbf{X}(x)\beta &= \sum_{k=0}^{p}\sum_{m=0}^{p} \beta_{km} x_i^k y_i^m \qquad k+m \leqslant p \\ \epsilon &\sim (0, \sigma^2) \\ Cov[\epsilon_i, \epsilon_j] &= 0 \ \forall i \neq j. \end{aligned} \tag{2.1}$$

Schabenberger y Gotway [18] mencionan que la teoría de estimación y predicción es directa en los modelos con errores sin correlación. Los coeficientes de regresión espacial

se estiman por mínimos cuadrados ordinarios como

$$\hat{\beta}_{ols} = \Big((\mathbf{X}(x))'\mathbf{X}(x)\Big)^{-1}\mathbf{X}(x)'S(x),$$

y la estimación de la varianza de los residuales es

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}(S(x_i) - \hat{E}[S(x_i)])^2.$$

Se puede decir que el "mejor estimador lineal insesgado" (BLUE) es E$[S(x_i)]$ y el "mejor predictor lineal insesgado" (BLUP) de $S(x_i)$ son los mismos, por tanto el mejor estimador y predictor son

$$\hat{S}(x_0) = \hat{E}[S(x_0)] = x'(x_0)\hat{\beta}_{ols}.$$

El predictor del cuadrado medio del error para $S(x_0)$ basado en $\hat{S}(x_0)$ es

$$CME[S(x_0); \hat{S}(x_0)] = \sigma^2(1 + x'(x_0)\Big(\mathbf{X}(x)'\mathbf{X}(x)\Big)^{-1}x(x_0)),$$

donde se asume que el nuevo punto de los datos $S(x_0)$ no está correlacionado con los datos observados. Esto es consistente con la suposición de los errores no correlacionados del modelo (2.1). Cabe destacar que el número de coeficientes de regresión en un modelo de tendencia en la superficie aumenta rápido con el grado del polinomio, β es un vector de largo $(p+1)(p+2)/2$.

2.3. Modelos lineales generalizados

Modelos lineales generalizados (MLG) extienden los modelos de regresión ordinaria para incluir distribuciones de respuesta no normales y funciones de la media, según informa Agresti [1]. Para un modelo lineal generalizado se especifican tres componentes:

1. El componente aleatorio
2. El componente sistemático
3. La función de enlace

Nelder y Wedderburn introdujeron la clase de modelos lineales generalizados en 1972 según Agresti [1].

Los componentes de los modelos lineales generalizados

Agresti [1] define los componentes de los modelos lineales generalizados de la siguiente forma: los componentes aleatorios de un MLG constan de una variable respuesta Y con observaciones independientes $(y_1, \ldots, y_N)$ de una familia de distribución exponencial natural. Esta familia tiene la función de densidad de probabilidad o función de masa de la siguiente forma

$$f(y|\theta) = \exp[yb(\theta) + c(\theta) + d(y)] \tag{2.2}$$

La variable θ es llamada el *parámetro canónico*, según Faraway [7]. Varias distribuciones importantes en la familia exponencial son la normal, Poisson y binomial.

El *componente sistemático* de un MLG se refiere a un vector $(\eta_1, \ldots, \eta_N)$ de las variables explicativas a través de un modelo lineal. Sea x_{ij} denotar el valor del predictor j $(j = 1, 2, \ldots, p)$ para el sujeto i. Entonces

$$\eta_i = \sum_j \beta_j x_{ij}, \qquad i = 1, \ldots, N.$$

es combinación lineal de las "variables explicativas", esto se conoce como el *predictor lineal*. Por lo general, un $x_{ij} = 1$ para todos i, para el coeficiente de un intercepto (a menudo denotado por α) en el modelo.

El tercer componente de un MLG es la *función de enlace* que conecta los componentes aleatorios y sistemáticos. Sea $\mu_i = \mathrm{E}(Y_i), i = 1, \ldots, N$. El modelo del enlace μ_i a η_i por $\eta_i = g(\mu_i)$, donde la función de enlace g es una función monótona diferenciable. Por lo tanto, g es el enlace de $\mathrm{E}(Y_i)$ para las variables explicativas a través de la fórmula

$$g(\mu_i) = \sum_j \beta_j x_{ij}, \qquad i = 1, \ldots, N. \tag{2.3}$$

La función de enlace $g(\mu) = \mu$ se llama *enlace identidad*, tiene $\eta_i = \mu_i$. Esto especifica un modelo lineal de la media. Además, ésta es la función de enlace para la regresión ordinaria con distribución Gaussiana Y.

Modelo de regresión logística

Supongamos que la variable respuesta Y_i para $i = 1, \ldots, n_i$ se distribuye binomial $Bin(n_i, p_i)$ tal que

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i},$$

asumiendo que las Y_i son independientes. Los experimentos individuales que componen la variable respuesta Y_i están sujetas a los mismos predictores p $(x_1, \ldots, x_n)$. Utilizando los MLG podemos crear una relación lineal entre las covariables y p. Para empezar, definimos los miembros de la familia exponencial (2.2) especificando las funciones b, c y d.

$$\begin{aligned} f(y|p) &= \binom{n}{y} p^y (1-p)^{n-y} \\ &= \exp\left(y \log(p) + (n-y)\log(1-p) + \log\binom{n}{y} \right) \\ &= \exp\left(y \log\Big(\frac{p}{1-p}\Big) + n\log(1-p) + \log\binom{n}{y} \right) \end{aligned} \tag{2.4}$$

Entonces, notamos que en (2.2), $b(p) = \log(\frac{p}{1-p})$, $c(p) = n\log(1-p)$ y $d(y) = \log\binom{n}{y}$. Por consiguiente, escogemos una función de enlace, $g(\mu) = \log\frac{\mu}{1-\mu}$, la cual describe la media de la variable respuesta, $\mathrm{E}(Y) = \mu$. Esta enlaza las covariables a través del predictor lineal:

$$\eta = g(\mu).$$

En principio la función de enlace puede ser cualquier función monótona continua y diferenciable, pero ya existen opciones comunes y convenientes para MLG estándar.

Para el MLG binomial, sea p la probabilidad de éxitos y sea esto nuestro μ. Si definimos la variable respuesta como una proporción en vez de un conteo, esto requiere que $0 < p < 1$. Podemos ver que no se puede utilizar $\eta = p$ ya que no es apropiado, por que por lo general la variable respuesta de una binomial no tiene relación lineal entre p y x. Existen distintas funciones de enlace que aseguran que se cumpla la condición. Las más comunes son:

1. Logit: $\eta = log(\frac{p}{1-p})$.

2. Probit: $\eta = \Phi^{-1}(p)$ donde Φ es la distribución acumulada de la normal estándar.

3. log-log: $\eta = \log(-\log(1-p))$.

El *enlace canónico* tiene a g tal que $\eta = g(\mu) = \theta$, el parámetro canónico de la familia de distribución exponencial. Esto significa que $g(b'(\theta)) = \theta$. Por lo tanto, el enclace canónico de una binomial es el logit. Si se utiliza el enlace canónico, $\mathbf{X}^T Y$ es suficiente para β.

Ahora utilizando el enlace logit obtenemos el modelo regresión logística o modelo logit:

$$logit[p(x)] = \log \frac{p(x)}{1-p(x)} = \beta_0 + \beta x_1 + \ldots + \beta_q x_q \tag{2.5}$$

2.3.1. Modelo lineal generalizado geo-estadístico

Un *Modelo lineal generalizado geo-estadístico* (MLGG) se define como un modelo lineal generalizado mixto de una forma específicamente orientada a los datos geo-estadísticos, según Diggle y Ribeiro [6]. Por consiguiente, para estos autores lo primordial para el modelo es tener un proceso gaussiano estacionario $S(x)$. Incluso para que el proceso sea estacionario el valor esperado, la varianza de $S(x)$ es la misma para todo x, y la correlación entre $S(x)$ y $S(x')$ depende solamente de $u = \|x - x'\|$, donde u es la distancia euclídea entre x y x'. Entonces, un proceso estocástico $S(x)$ es un modelo gaussiano, si la distribución conjunta de $S(x_1), \ldots, S(x_n)$ es gaussiana multivariada $\forall n \in \mathbb{N}$ y un conjunto de ubicaciones x_i.

Por otra parte, Diggle y Ribeiro [6] dicen que en esta parte del modelo se sigue un modelo clásico lineal generalizado como se describe por McCullagh y Nelder en el

1989. Asimismo como ellos lo describen anteriormente, pero con $S(x)$ como un desplazamiento en el predictor lineal. Entonces, las variables respuestas $Y_i : i = 1, \ldots, n$ en ubicaciones $x_i : i = 1, \ldots, n$ son explícitamente condicionales en $S(\cdot)$. En efecto, éstas son variables aleatorias mutuamente independientes entre sí, cuya esperanza condicional, $\mu_i = \mathrm{E}[Y_i|S(\cdot)]$ se determinan como

$$g(\mu_i) = S(x_i) + \sum_{k=1}^{p} \beta_k d_k(x_i) \tag{2.6}$$

donde $g(\cdot)$ es una función conocida, llamada la función de enlace. Los $d_k(\cdot)$ son variables espaciales explicativas observadas en el espacio y los β_k son parámetros desconocidos de la regresión espacial. Los términos en el lado derecho de la ecuación (2.6) se conocen colectivamente como el predictor lineal del modelo.

El MLG se puede extender de diversas maneras para incorporar los datos dependientes. La forma que Diggle y Ribeiro [6] extienden el MLG para que funcione con respuestas dependientes, introduce los *efectos aleatorios* sin observar al predictor lineal. Luego, η_i modifica a

$$\eta_i = d_i'\beta + S_i$$

donde ahora $S = (S_i, \ldots, S_n)$ sigue una distribución multivariable de media cero. Los S_i se llaman *efectos aleatorios* o *variables latentes*. Lo antes mencionado provee un marco para los modelos de regresión de los datos continuos o discretos. En resumen, si un modelo tiene las características antes mencionada lo denominamos como un *modelo lineal generalizado mixto* (MLGM).

Ahora, se asume que los S_i son mutuamente independientes, para que el modelo

incorpore extra-variación, o el exceso de dispersión, con respecto a MLG clásico. Esto se hace para poder modelar los datos dependientes mediante un MLGM, pero se necesita especificar adecuadamente la dependencia de los S_i. Entonces en las aplicaciones lo usual es que S sea una variable aleatoria gaussiana multivarida con una estructura específica de covarianza estipulada por el contexto de la aplicación.

Por lo general, para aplicaciones geo-estadísticas no se puede confiar en ninguna forma de replicación independiente según indican Diggle y Ribeiro [6]. Por lo tanto, los autores sugieren que las variables respuestas observadas $y = (y_1, \ldots, y_n)$ deben ser consideradas como una sola realización de una variable aleatoria n-dimensional Y. Entonces, en este contexto se debe utilizar MLGM donde S es equivalente a $S = \{S(x_1) \ldots, S(x_n)\}$, los valores de un proceso de señal gaussiana subyacente en cada uno de los puntos de muestreo x_i. Ellos se refieren a un modelo de este tipo, un modelo lineal generalizado geo-estadístico o MLGG. Esta no es la única manera en que podemos adaptar el MLG clásico para el uso de aplicaciones geo-estadísticos según explican Diggle y Ribeiro [6].

Diggle y Ribeiro [6] informan que la estrategia para modelos lineales generalizados es más atractiva cuando las variables respuesta Y_i siguen una distribución en la familia exponencial, condicionada a los efectos aleatorios S en el caso de un modelo mixto, desde el diseño de muestreo. Por esta razón, dos de los MLG más utilizados son el modelo log-lineal de Poisson para la respuesta de recuentos y el modelo logístico lineal para datos binarios, o más en general con respuesta binomial.

2.3.2. Modelo lineal logístico binomial para datos geo-estadísticos

El *modelo lineal logístico binomial* es un MLG con función de enlace logit y una distribución condicional Y_i, donde Y_i es binomial. La forma más simple del modelo es cuando los Y_i representan ensayos Bernoulli independientemente condicionales con $\{Y_i = 1|S(\cdot)\} = p(x_i)$, donde

$$log[p(x_i)/\{1 - p(x_i)\}] = \alpha + S(x_i). \tag{2.7}$$

y $S(\cdot)$ son procesos Gaussianos estacionarios con media cero, varianza σ^2 y función de correlación $\rho(u)$. Diggle y Ribeiro [6] dicen que "el contenido de información en los datos generados a partir de este modelo es bastante limitado, al menos que la intensidad de los puntos de muestreo en el espacio sea más grande que la variación en el proceso de señal $S(\cdot)$". También asesoran que en el ajuste geo-estadístico el modelo binomial es mucho más útil cuando las respuestas binarias Y_i se sustituye por conteos binomiales condicionalmente con denominadores n_i grandes.

Diggle y Ribeiro [6] muestran que la diferencia entre el modelo binomial y modelo lineal Gaussiano es que la varianza condicional Y_i dado $S(x_i)$ no es un parámetro libre, pero está limitada a ser igual a la esperanza condicional de Y_i. En algunos casos podemos encontrar evidencia de variabilidad adicional en los datos. Ésta a menudo se conoce como varianza extra-binomial, la cual no está estructurada en el espacio. Estos autores sugieren que en este caso se haga una extensión natural al modelo, es decir, incluir el efecto pepita dentro del predictor lineal. La distribución condicional Y_i sigue

siendo modelada por una binomial, pero (2.7) se extiende a

$$log[p(x_i)/\{1-p(x_i)\}] = \alpha + S(x_i) + Z_i \tag{2.8}$$

donde $S(\cdot)$ es igual que anteriormente y los Z_i son mutuamente independientes $N(0, \tau^2)$.

En principio, esta extensión del modelo permite descomponerlo en dos componentes de la varianza pepita, que en general no se podían distinguir en el modelo Gaussiano lineal:

1. La variación binomial inducida por el sistema de muestreo, análogo a nuestra interpretación anterior del efecto pepita como error de medida.

2. Componente en el espacio sin correlacionar, análogamente a la interpretación alterna del efecto pepita como la variación en el espacio a escalas pequeñas.

2.4. Función de semivariograma y función de correlación

Schabenberger y Gotway [18] indican que un campo aleatorio es estacionario de segundo orden si tiene las siguientes condiciones; $\mathrm{E}[S(x)] = \mu$ y $Cov[S(x), S(x+\mathbf{h})] = C(\mathbf{h})$ en donde $S(x)$ es un campo aleatorio tal que $\{S(x) : x \in D \subset \mathbb{R}^d\}$. Por lo tanto, esto significa que un campo aleatorio estacionario de segundo orden tiene media constante y la covarianza entre dos atributos en diferentes localidades es una función de separación en el espacio $\mathbf{h}$. Esta función se llama *función de covarianza* y se denota $C(\mathbf{h})$. Si existe la función de covarianza $C(\mathbf{h})$ en un campo aleatorio estacionario de

segundo orden y $C(\mathbf{h})$ no depende de las coordenadas absolutas $Cov[S(x), S(x+\mathbf{0})] = Var[S(x)] = C(\mathbf{0})$. Por ende, la variabilidad en un campo aleatorio estacionario de segundo orden es la misma en todas partes de la región de estudio. En conclusión, un proceso espacial estacionario de segundo orden tiene media constante, varianza constante y su covarianza no depende del proceso espacial.

La estacionaridad es importante para los datos espaciales, según Schabenberger y Gotway [18]. Estos autores nos informan que en el caso que campo aleatorio no sea estacionario en el espacio, este se puede convertir en uno estacionario. Esto se hace similarmente a una serie de tiempo, para convertir una serie no estacionaria en una estacionaria es sólo la diferencia entre las series. Es decir que si $S(x)$ no es estacionario de segundo orden, los incrementos $S(x) - S(x+h)$ pueden ser estacionarios de segundo orden. Un proceso con la característica antes mencionada se dice que es *intrínsecamente estacionario*.

Schabenberger y Gotway [18] dan la siguiente definición; sea $\{S(x) : x \in D \subset \mathbb{R}^d\}$ un proceso espacial es intrínsecamente estacionario si $\mathrm{E}[S(x)] = \mu$ y

$$\frac{1}{2}Var[S(x) - S(x+\mathbf{h})] = \gamma(\mathbf{h}) \tag{2.9}$$

donde la función $\gamma(\mathbf{h})$ se llama *semivariograma* del proceso espacial. Además, Schabenberger y Pierce [19] dicen que semivariograma es una función básica que transmite información sobre la estructura espacial y el grado de continuidad de un campo aleatorio, donde $S(x)$ es nuestra variable de interés. Algunos autores como Diggle y Ribeiro [6], Chilès y Delfiner [3] definen a (2.9) como $2\gamma(x_i - x_j)$ y lo llaman como *variograma*. Chilès y Delfiner [3] reconocen que $2\gamma(x_i - x_j)$ también se le llama semivariograma.

No hay nada "establecido" por estar fuera el factor 2, para más claridad nos vamos a referir a γ como semivariograma y a 2γ como variograma. Schabenberger y Gotway [18] ilustran que para ver que un proceso estacionario de segundo orden sea también intrínsecamente estacionario, solo es suficiente examinar

$$\begin{aligned} Var[S(x) - S(x+\mathbf{h})] &= Var[S(x)] + Var[S(x+\mathbf{h})] - 2Cov[S(x), S(x+\mathbf{h})] \\ &= 2\{Var[S(x)] - 2C(\mathbf{h})\} \\ &= 2\{C(\mathbf{0}) - C(\mathbf{h})\} = 2\gamma(\mathbf{h}). \end{aligned} \tag{2.10}$$

Por otra parte, intrínsecamente estacionario, no implica estacionario de segundo orden. Entonces, según (2.10) existe una relación entre $\gamma(\mathbf{h})$ y $C(\mathbf{h})$, la cual es $\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}) = \sigma^2\{1-\rho(\mathbf{h})\}$ donde $\sigma^2 = Var\{S(x)\}$ y $\rho(\mathbf{h}) = Corr\{S(x), S(x+\mathbf{h})\}$. Por ende, todo método estadístico para campos aleatorios estacionarios de segundo orden puede hacerse en términos de funciones de semivariograma o funciones de correlaciones. Incluso, Schabenberger y Gotway [18] dicen que los parámetros del proceso estocástico bajo estudio, se pueden visualizar como reparametrizaciones de estructura de segundo orden del proceso, por tanto son equivalentes. Pero Diggle y Ribeiro [6] advierten que si el proceso es intrínsecamente estacionario no necesariamente es estacionario de segundo orden, hay que tener cuidado al calcular $\rho(\mathbf{h})$ como $\gamma(\mathbf{h})/\sigma^2$. También, si el proceso es intrínsecamente estacionario pero no estacionario de segundo orden, $\rho(\mathbf{h})$ no es un parámetro del proceso, por lo tanto hay que trabajar con $\gamma(\mathbf{h})$.

Schabenberger y Gotway [18] mencionan que la estructura de segundo momento de un campo aleatorio estacionario de segundo orden es la función de separación $\mathbf{h}$.

Además, los autores dicen que la función de correlación puede depender de la dirección. Definen que cuando hay ausencia de la dependencia de la dirección, es decir, cuando la función de correlación o el semivariogama dependen sólo de la distancia absoluta entre los puntos, la función se conoce como *isotrópica*. Entonces si el campo aleatorio es estacionario de segundo orden con función de correlación isotrópica, $\rho(\mathbf{h}) = \rho^*(\|\mathbf{h}\|)$, donde $\|\mathbf{h}\|$ es la norma Euclideana del vector $\mathbf{h}$ ("distancia"),

$$(x + \mathbf{h}) - x = \mathbf{h} = \sqrt{h_1^2 + h_2^2}.$$

Similarmente si el semivariograma de un proceso intrínsecamente estacionario es isotrópica, entonces $\gamma(\mathbf{h}) = \gamma^*(\|\mathbf{h}\|)$

Cuando la media de un proceso estacionario es constate, Diggle y Ribeiro [6] dicen que el semivariograma puede ser definido como $\gamma(\mathbf{h}) = \frac{1}{2}\mathrm{E}[\{S(x) - S(x-\mathbf{h})\}^2]$. Ahora, suponemos que $(x_i, x_j) : i = 1, \ldots, n$ son generados por un proceso estacionario

$$Y_i = S(x_i) + Z_i$$

donde los Z_i son mutuamente independientes, idénticamente distribuidas con media cero y varianza τ^2. Los autores definen el semivariograma de un proceso observado, $\gamma_Y(\mathbf{h})$ como

$$\gamma_Y(h_{ij}) = \frac{1}{2}E[(Y_i - Y_j)^2]$$

donde $h_{ij} = \|x_i - x_j\|$. Esto resulta que

$$\gamma_Y(\mathbf{h}) = \tau^2 + \sigma^2\{1 - \rho(\mathbf{h})\}. \tag{2.11}$$

Diggle y Ribeiro [6] señalan que $\rho(\mathbf{h})$ es una función monótona decreciente con $\rho(0) = 1$ y $\rho(\mathbf{h}) \to 0$ cuando $u \to \infty$. La ecuación (2.11) resume perfectamente las cualidades esenciales de un modelo geo-estadístico clásico.

Además, estos autores mencionan que el semivariograma es una función monótona creciente [6]. Ellos también detallan la función de semivariograma en un campo aleatorio estacionario de segundo orden que tiene las siguientes características. La varianza del *efecto pepita* corresponde al intercepto τ^2. La asíntota, $\tau^2 + \sigma^2$, corresponde a la varianza de las observaciones del proceso Y, es decir, $\text{Var}[S(x)] = \gamma(0)$ (la magnitud de discontinuidad del origen en el semivariograma), esto sucede por la propiedad o ecuación (2.10). A esto se le llama *meseta* ("sill"), lo cual es la suma de la varianza del efecto pepita y la varianza de la señal, σ^2. En la manera que el semivariograma crece desde su intercepto a la asíntota es determinado por su función de correlación $\rho(\mathbf{h})$. Una de las características más importantes de $\rho(\mathbf{h})$ es su comportamiento cerca de $h = 0$, que se refiere a la suavidad analítica del proceso de señal subyacente. Otra característica importante es la rapidez con que $\rho(\mathbf{h})$ se aproxima a cero con el aumento de h, que refleja la extensión física de la correlación espacial del proceso. A la distancia h que el semivariograma alcanza la meseta se le conoce como el *alcance* ("range") del semivariograma, es decir, cuando $\rho(h) = 0$ para h mayor que algún valor finito. Además, si $\rho(\mathbf{h})$ sólo tiende a cero asintóticamente cuando h crece, entonces el alcance no está definido. Según la convención geo-estadística definimos la *zona de influencia* ("practical range") como la distancia h_0 para la cual $\rho(h_0) = 0.05$, luego $\gamma_Y(h_0) = \tau^2 + 0.95\sigma^2$, es decir, el segmento de la distancia a la cual el semivariograma alcanza el 95 % de la meseta.

Para Schabenberger y Pierce [19] los semivariogramas no alcanzan la meseta con frecuencia. Esto se podría deber a:

- El proceso espacial no es estacionario, por ejemplo la media de $S(x)$ no es constante a través del dominio.
- Para un proceso espacial intrínsecamente estacionario. La hipótesis intrínseca estipula que un semivariograma tiene que satisfacer:

$$2\frac{\gamma(u)}{\|\mathbf{h}\|^2} \to 0 \; si \; \|\mathbf{h}\| \to 0.$$

- Un proceso espacial estacionario de segundo orden, la distancia mayor del semivariograma estimado es menor que el alcance del proceso. La distancia en que el semivariograma alcance la meseta aún no se ha observado.

Según Diggle y Ribeiro [6] la varianza del efecto pepita es un parámetro importante para la predicción espacial, en el contexto de (2.11) equivale al intercepto de $\gamma_Y(\mathbf{h})$. Además, ellos dicen que valor de τ^2, no afecta el grado en que se predice la superficie $\hat{S}(x)$, esto está para todos los datos observados Y_i (En particular, si ajustamos $\tau^2 = 0$ a la predicción espacial para interpolar los datos). La decisión si se establece $\tau^2 = 0$, o se estima un valor en la posición de τ^2, es también importante cuando se elige un modelo con una estructura conocida.

El semivariograma de un campo aleatorio bajo las condiciones de estacionaridad intrínseca o estacionariedad de segundo orden son parámetros importantes en los métodos de geo-estadística según Diggle y Ribeiro [6]. En donde ambos requieren que

$\mathrm{E}[S(x)] = \mu$. También Diggle y Ribeiro [6] dicen que los parámetros se utilizan para el análisis de datos en el espacio. Igualmente ambos son muy importantes para los métodos de *kriging*, el cual se utiliza para la predicción espacial (Schabenberger y Pierce [19]). También Schabenberger y Gotway [18] indican que el semivariograma no es sólo una herramienta descriptiva, ni un dispositivo para derivar la estructura de dependencia en el espacio en un campo aleatorio ni para construir la matriz de varianza-covarianza de $S(x)$, que es necesaria para el modelo basado en inferencia estadística. Es una herramienta estructural a través de la cual las propiedades de segundo orden de un proceso espacial pueden ser estudiadas y transmite mucha información sobre el comportamiento de un campo aleatorio.

Sea $\gamma(\mathbf{h})$ un semivariograma isotrópico de un campo aleatorio estacionario de segundo orden o intrínsecamente estacionario. Según Schabenberger y Pierce [19] se tienen las siguiente propiedades:

- Si $\gamma(\mathbf{h})$ es válido en $\mathbb{R}^d$, entonces es también válido en $\mathbb{R}^s, s < d$. Si $C(\mathbf{h})$ es válido en $\mathbb{R}^d$, entonces es también válido en $\mathbb{R}^s, s < d$.

- Si $\gamma_1(\mathbf{h})$ y $\gamma_2(\mathbf{h})$ son semivariogramas válidos, entonces $a\gamma_1(\mathbf{h}) + b\gamma_2(\mathbf{h}),\ a, b \geq 0$, es un semivariograma válido.

- Si $\gamma(\mathbf{h})$ es un semivariograma válido, entonces tiene la propiedad de uniformidad, esto es, $\gamma(\mathbf{h}) = \gamma(-\mathbf{h})$. También pasa por el origen, $\gamma(0) = 0$, como $\mathrm{Var}[S(x) - S(x - \mathbf{h})] = \mathrm{Var}[0] = 0$.

- Un semivariograma válido $\gamma(\mathbf{h})$ es condicionalmente negativo definido, esto es,

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i, a_j\gamma(\mathbf{x}_i - \mathbf{x}_j) \leq 0$$

  para cualesquiera números reales $a_1, \ldots, a_m$, tal que $\sum_{i=1}^{m} a_i = 0$ y un número finito de localidades espaciales.

- Una condición necesaria para $\gamma(\mathbf{h})$ sea un semivariograma válido es que $2\gamma(\mathbf{h})$ crezca más lento que $\|\mathbf{h}\|^2$. Esto usualmente se conoce como la hipótesis de intrínseca, es decir, la hipótesis intrínseca estipula que un semivariograma tiene que satisfacer:

$$2\frac{\gamma(u)}{\|\mathbf{h}\|^2} \to 0 \; si \;\; \|\mathbf{h}\| \to 0.$$

Como se mencionó anteriormente, existe una relación entre el semivariograma y la función de correlación, la cual es

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}) = (\sigma^2)1 - \rho(\mathbf{h}).$$

Esta relación existe, si el proceso es estacionario de segundo orden. Entonces, se puede calcular la función de correlación $\rho(\mathbf{h})$ como $\gamma(\mathbf{h})/\sigma^2$. En cambio, si el proceso es intrínsecamente estacionario pero no es estacionario de segundo orden, hay que tener cuidado cuando se hace esto ya que la función de correlación no es un un parámetro del proceso. Las familias más utilizadas en las aplicaciones para calcular el semivariograma son: Matérn, exponencial, entre otras:

La familia Matérn

Schabenberger y Gotway [18] dicen que Matérn (1986) construyó una clase flexible de la función de correlación,

$$\rho(h) = \sigma^2 \frac{1}{\Gamma(\kappa)} \left(\frac{\phi h}{2} \right)^{\kappa} 2K_{\kappa}(\phi h) \qquad \kappa > 0, \phi > 0, \tag{2.12}$$

donde K_{κ} es la función modificada Bessel de segundo tipo para orden $\kappa > 0$. Entonces, $\phi > 0$ es un *parámetro de escala* con las dimensiones de la distancia y gobierna el alcance de la dependencia espacial. κ lo llamamos *orden* y se conoce como el *parámetro de forma* que determina la suavidad analítica del proceso espacial subyacente $S(x)$ y esta suavidad aumenta según κ aumenta. Además, σ^2 es la varianza del proceso espacial. Diggle y Ribeiro [6], Schabenberger y Gotway [18] establecen que $\rho(h)$ dada por (2.12) es válida en $\mathbb{R}^d$. Además, ellos señalan que existe una relación entre la zona de influencia y el parámetro de escala ϕ. Ambos dependen del valor de κ y la zona de influencia es asímismo una función de κ.

Diggle y Ribeiro [6] hacen referencia a que el parámetro ϕ y κ en (2.12) no son ortogonales, en el siguiente sentido. Por ejemplo, si la estructura de correlación es Matérn con parámetro ϕ y κ, entonces la mejor aproximación que ajusta con orden $\kappa^* \neq \kappa$ y también de $\phi^* \neq \phi$. Otra manera de decirlo es que los parámetros de escala que corresponden a correlaciones Matérn de diferentes órdenes, no son directamente comparables.

Schabenberger y Gotway [18] destacan que uno de los modelos isotrópicos más conocidos con la familia Matérn se presenta cuando $\kappa \to \infty$. Este modelo de correlación

límite es conocido como el *modelo gaussiano* y la ecuación (2.12) toma una forma simple

$$\rho(u) = \sigma^2 \exp\{-\phi u^2\} = \sigma^2 \exp\{-3\frac{u^2}{\alpha^2}\} \tag{2.13}$$

La segunda parametrización es común en las aplicaciones geo-estadísticas, donde α es la zona de influencia, la distancia a la que las correlaciones se han reducido $\rho \approx 0.05$ o menos.

También Schabenberger y Gotway [18] mencionan que otro modelo isotrópico conocido con la familia Matérn es cuando $\kappa = 1/2$. Este modelo de correlación es conocido como el *modelo exponencial* y la ecuación (2.12) toma una forma simple

$$\rho(u) = \sigma^2 \exp\{-\phi u\} = \sigma^2 \exp\{-3\frac{u}{\alpha}\} \tag{2.14}$$

También la segunda parametrización es común en geo-estadística, donde de nuevo α es la zona de influencia.

La familia exponencial

Diggle y Ribeiro [6] definen esta familia por la función de correlación:

$$\rho(h) = \sigma^2 \exp\{-(u/\phi)^\kappa\}. \tag{2.15}$$

Los autores informan que la familia exponencial tiene un parámetro de escala $\phi > 0$ y otro de forma κ. El parámetro de forma κ está limitado por $0 < \kappa \leq 2$. Además, dicen que estos dos parámetros generan las funciones de correlación, las cuales son monótonas decrecientes en h. La relación entre la zona de influencia y el parámetro ϕ,

dependerá del valor de κ. Incluso señalan que sin embargo, la familia es menos flexible que la Matérn en el sentido de que el proceso Gaussiano subyacente $S(x)$ es contínuo en cuadrado medio y no diferenciable cuando $0 < \kappa < 2$, aunque sí uno infinitamente diferenciable cuando $\kappa = 2$, valor máximo legítimo.

Un caso extremo es cuando $\kappa = 2$, el cual equivale a un caso límite de una función de correlación Matérn como $\kappa \to \infty$, conforme a Diggle y Ribeiro [6]. Ellos dicen que este caso puede generar la estructura de covarianza "ill-conditioned". Para Diggle y Ribeiro [6], un proceso $S(x)$ con la propiedad teórica antes mencionada, es una función de correlación con realización en un intervalo arbitrariamente pequeño. Este intervalo será continuo, el cual determina la realización en toda la línea real. Se puede decir que para la mayoría de las aplicaciones, esto es considerado irrealista.

Otras familias

En la geo-estadística clásica existen una variedad de familias. Dentro de las más utilizadas se encuentra la *familia esférica*, la cual tiene función de correlación

$$\rho(h) = \sigma^2 \left(1 - \frac{3h}{2\alpha} + \frac{1}{2}\left(\frac{h}{\alpha}\right)^3\right). \tag{2.16}$$

Diggle y Ribeiro [6] notan que una diferencia cualitativa entre ésta y las familias descritas anteriormente es que tiene un alcance finito. Comparando con la clase Matérn de dos parámetros, se dice que la familia esférica carece de flexibilidad. Schabenberger y Gotway [18] dicen que la función de correlación de esta familia se comporta lineal o casi linealmente cerca del origen. Además, ellos señalan que el modelo de correlación esférico es exactamente cero a la distancia $h = \alpha$, porque estos modelos tienen un

verdadero alcance y usualmente exhiben la meseta en $h = \alpha$.

2.5. Estimación en modelos mixtos normales

2.5.1. Estimador de máxima verosimilitud

Diggle y Ribeiro [6] comentan que la estimación de máxima verosimilitud es un método estadístico ampliamente aceptado y conocido, que tiene propiedades óptimas para muestras grandes. Casella y Berger [2] definen el estimador de máxima verosimilitud (EMV) de la siguiente forma. Sea $X_1, \ldots, X_n$ muestras iid de una población con función de densidad de probabilidad o función de masa de probabilidad $f(x|\theta_1, \ldots, \theta_k)$. Entonces la función de verosimilitud se define por

$$L(\theta|X) = L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x|\theta_1, \ldots, \theta_k) \tag{2.17}$$

Si la función de verosimilitud es diferenciable en θ_i, los posibles estimadores para el estimador de máxima verosimilitud son los valores de $(\theta_1, \ldots, \theta_k)$ tal que

$$\frac{\partial}{\partial \theta_i} L(\theta|x) = 0, \qquad i = 1, \ldots, k \tag{2.18}$$

Note que la solución (2.18) son sólo los posibles estimadores de máxima versimilitud. Por conocimiento previo se sabe que esta condición no es suficiente, porque los puntos encontrados cuando la primera derivada es cero pueden ser máximos o mínimos locales, máximos o mínimos globales, o puntos de inflección.

Inicialmente para poder estimar estos parámetros de un campo aleatorio espacial

se debe conocer la distribución espacial. Schabenberger y Gotway [18] dicen que el estimador de máxima verosimilitud para modelos espaciales se ha desarrollado solamente para el caso del campo aleatorio Gaussiano (Mardia and Marshall, 1984). Entonces, sea $S = [S(x_1), \cdots, S(x_n)]'$ el vector de observaciones y se supone que $S(x) \sim N(\mathbf{X}(x)\beta, \Sigma(\theta))$. Schabenberger y Gotway [18] mencionan que el EMV estima los parámetros de la media y covarianza al mismo tiempo. Además que la característica antes mencionada puede ser dificultada por el perfil de β, pero lo importante es que el estimador de máxima verosimilitud tiene una solución simultánea al problema de miminización de dos veces el negativo de logaritmo de verosimilitud gaussiano:

$$\varphi(\mu; \theta; S(x)) = \ln\{|\Sigma(\theta)|\} + n\ln\{2\pi\} + (S(x) - \mathbf{X}(x)\beta)'\Sigma(\theta)^{-1}(S(x) - \mathbf{X}(x)\beta). \quad (2.19)$$

Los autores señalan que si $\mathbf{X}$ es una matriz de rango k, entonces el problema de optimización envuelve $k + q$ parámetros, donde q es el número de parámetros de la función de correlación [18]. Incluso, explican que esto se debe a que los elementos de Σ son usualmente funciones no lineales de los elementos de θ. Además, indican que el proceso de optimación es usualmente iterativo. Se empieza de un valor $[\theta^{(0)}, \beta^{(0)}]$, se calcula recursivamente de acuerdo a técnicas no lineales de optimización. Las técnicas más comunes son Newton-Raphson y QuasiNewton.

Diggle y Ribeiro [6] mencionan que el estimador de máxima verosimilitud se distribuye asintóticamente normal, insesgado y totalmente eficiente bajo condiciones de regularidad estándar. Luego, mencionan que dentro del contexto geo-estadístico, la implementación de estimación de máxima verosimilitud sólo es directa cuando los datos

son generados por un modelo Gaussiano. Además, los detalles prácticos de la optimización puede depender de la familia en particular bajo consideración. Incluso, es importante tener en cuenta las diferentes parametrizaciones de $\Sigma(\hat{\theta})$, ya que pueden afectar a la convergencia de la optimización numérica.

Diggle y Ribeiro [6] enfatizan en particular la parametrización estándar Matérn y las familias exponenciales con potencias, las cuales conducen a una interpretación natural con parámetro de escala ϕ y un parámetro de forma κ, pero ambos parámetros no son ortogonales en el sentido estadístico (Los estimadores de máxima verosimilitud por ϕ y κ tienden a estar fuertemente correlacionados). Como respuesta a esto se necesita considerar sólo una pequeña cantidad de valores candidatos para κ, correspondientes a una diferencia cualitativa de suavidad del proceso de la señal.

Además ellos enfatizan otra característica del estimador de máxima verosimilitud, la cual es la habilidad de estimar la matriz de varianza-covarianza para los parámetros estimados, sólo basándose en las observaciones o matriz de información esperada. La matriz de información esperada es igual a $0.5\mathbf{H}$, donde $\mathbf{H}$ es la matriz Hessian (segunda derivada) de la ecuación (2.19). Los errores estándar del estimador de máxima verosimilitud se obtienen con los elementos de la diagonal de $2\mathbf{H}^{-1}$ o $2\mathrm{E}[\mathbf{H}]^{-1}$.

2.5.2. Estimador de máxima verosimilitud restringida

La estimación de máxima verosimilitud restringida (EMVR) fue introducida por Patterson y Thompson en el contexto de la estimación de componentes de la varianza en los experimentos diseñados según Schabenberger y Gotway [18], También, ellos mencionan que en las aplicaciones, el EMVR es preferido sobre el estimador de máxima

verosimilitud por que el estimador de máxima verosimilitud exhibe mayor sesgo para los estimadores de parámetros de correlación, mientras que EMVR reduce el sesgo. Ellos dicen que esto se debe a que el estimador de máxima verosimilitud falla en tomar en consideración el número de parámetros para la media en la estimación de los parámetros de correlación.

Para que un modelo espacial $S(x) \sim N(\mathbf{X}(x)\beta, \Sigma(\theta))$. Entonces, Schabenberger y Gotway [18] dicen que la idea principal del EMVR es estimar los parámetros de varianza y covarianza por medio de maximizar la verosimilitud de $KS(x)$, en vez de maximizar la verosimilitud de $S(x)$. Donde K es una matriz $((n-k)\times n)$ elegida talque: $\mathrm{E}[KS(x)] = 0$, con rango $[K] = n-k$. Además, estos autores mencionan que K se llama una matriz de contraste de error, debido a la propiedad antes mencionada [18]. Incluso dicen que la función de la matriz K es "eliminar la media" y explican que de ahí sale su nombre de máxima verosimilitud restringida. Incluso ellos destacan que la estimación de EMVR se desarrolla sólo para el caso de una función lineal, de lo contrario, no está claro cómo se construye la matriz de contraste de error K.

Schabenberger y Gotway [18] dicen que la diferencia entre la estimación de EMVR y la estimación de EMV se basa en el manejo de β. Ellos toman en cuenta la estimación de máxima verosimiltud para el vector $KS(x)$ de $(n-k)$, obtenemos el dos veces negativo logaritmo de verosimilitud de $KS(x)$ y notamos que β no se encuentra en la función.

$$\begin{aligned}\varphi_R(\theta; KS(x)) = &\ln\{|K\Sigma(\theta)K'|\} + (n-k)\ln\{2\pi\} \\ &+ S(x)'K'(K\Sigma(\theta)K')^{-1}KS(x).\end{aligned} \tag{2.20}$$

Otra observación que hacen Schabenberger y Gotway [18] es que la ecuación (2.20)

sólo es función de θ. Ahora queremos escribir $\varphi_R(\theta; KS(x))$ en términos de $\hat{\beta}$. Los autores comentan que se debe tener en mente que "no hay un estimador EMVR para β" [18] . Minimizando la ecuación (2.20) se obtiene $\hat{\theta}_{emvr}$. Entonces, el término $\hat{\beta}_{emvr}$ es un estimador de mínimos cuadrados generalizados evaluado en $\hat{\theta}_{emvr}$ y se calcula como

$$\hat{\beta}_{emvr} = (\mathbf{X}'\Sigma(\hat{\theta}_{emvr})^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma(\hat{\theta}_{emvr})^{-1}S(x). \tag{2.21}$$

Esta diferencia entre el EMV y la estimación EMVR es importante, porque el logaritmo de verosimilitud (2.19) se puede utilizar para probar hipótesis sobre μ y θ con una prueba de la razón de verosimilitud. Las comparaciones de la razón de verosimilitud basado en (2.20) sólo tienen sentido cuando se relacionan con los parámetros de correlación en θ y los modelos tienen la misma estructura media. Este punto reviste importancia cuando la media se modela con una estructura de regresión más general, por ejemplo, $\mu = \mathbf{X}'(x)\beta$. Para Diggle y Ribeiro [6] el ERMV es ampliamente recomendado para los modelos geo-estadísticos, suele ser más sensible que la EMV con el modelo elegido para $\mu(x)$.

2.5.3. Predicción espacial

Diggle y Ribeiro [6] dicen que el problema principal de la predicción es utilizar los datos que hay disponibles para predecir la data no observada del proceso señal de $S(\cdot)$. Para ellos, otra manera de explicar el objetivo para la predicción es predecir la variable $T = T(S)$, donde S es el conjunto completo de todos lo valores de $S(x)$ cuando x varía sobre todo el espacio de interés.

Diggle y Ribeiro [6] utilizan un ejemplo general de este tipo de problema. Dicen que

la mayor dificultad se encuentra a la hora de predecir el valor de la señal $T = S(x)$ en una localización arbitraria de x. Cuando se utilizan datos observados $Y = (Y_1, \ldots, Y_n)$, donde cada Y_i representa una posible versión ruidosa de cada $S(x_i)$. Los autores dicen que otros objetivos comunes son la de predicción T que incluye la integral de $S(x)$ sobre la sub región A. También, aunque un poco más trabajoso, una función no lineal tal como el máximo de $S(x)$ o con un conjunto de localizaciones para el cual $S(x)$ excede cualquier valor ya dado.

Predicción de mínimo error cuadrático medio

Diggle y Ribeiro [6] definen un predictor puntual de la siguiente forma. Sea Y un vector de variables aleatorias observadas, y sea T otra variable aleatoria la cual se utiliza para predecir el valor observado de Y. Entonces un *predictor puntual* para T es cualquier función de Y, la que se escribe de forma $\hat{T} = t(Y)$. Asimismo la *predicción del error cuadrático medio* de $\hat{T}$ es

$$ECM(\hat{T}) = E[(\hat{T} - T)^2], \tag{2.22}$$

donde la esperanza es con respecto a la distribución conjunta de T y $\hat{T}$ o equivalentemente, la distribución conjunta de T y Y. Para ellos, la forma general del predictor puntual que minimiza $ECM(\hat{T})$ es dada por el siguiente resultado conocido.

Teorema 2.5.1. *$ECM(\hat{T})$ toma su valor mínimo cuando $\hat{T} = E(T|Y)$.*

Diggle y Ribeiro [6] instruyen diciendo que una predicción puntual proporciona un resumen conveniente, pero una respuesta completa al problema de predicción es la distribución condicional de T dado Y. Incluso que la media de esta distribución

condicional es simplemente un resumen de muchos otros resúmenes que podríamos haber usado. Según el Teorema 2.5.1 el error cuadrático medio de $\hat{T}$ es:

$$E[(T - \hat{T})^2] = E_Y[Var(T|Y)]. \tag{2.23}$$

Se llama a $Var(T|Y)$ el *predictor de varianza*. Según Diggle y Ribeiro [6], el valor de la predicción de varianza en los valores observados de Y, estima el valor obtenido del error cuadrático medio de $\hat{T}$.

Además, note que si T y Y son variables aleatorias independientes, entonces $\mathrm{E}[(T - \hat{T})^2] \leqslant Var(T)$. Esto surge del hecho que $Var(T) = \mathrm{E}[(T - \mathrm{E}[T])^2]$ es el error cuadrático medio trivial para el predictor $\tilde{T} = \mathrm{E}[T]$, el cual ignora los datos de Y. Informalmente, la diferencia entre la varianza marginal de $Var(T)$ y la varianza condicional $Var(T|Y)$ brinda un resumen de variable respuesta qué tan útil son los datos de Y para predecir T según Diggle y Ribeiro [6].

Predicción de mínimo error cuadrático medio para un modelo gaussiano estacionario

Diggle y Ribeiro [6], en esta parte se debe asumir que nuestros datos $Y = (Y_1, \ldots, Y_n)$ son generados por un modelo Gaussiano estacionario. Sea $S = (S(x_1), \ldots, S(x_n))$ los valores sin observar de la señal en las localidades $x_1, \ldots, x_n$ en la muestra. S es una Gaussiana multivariada con un vector de media $\mu\mathbf{1}$ donde $\mathbf{1}$ es un vector de 1 y de la matriz de varianza $\sigma^2 R$, en donde R es una matriz n por n con elementos $r_{ij} = \rho(\|x_i - x_j\|)$.

Similarmente, Y es una Gaussiana multivariada con un vector de media $\mu\mathbf{1}$

$$\sigma^2 V = \sigma^2(R + \nu^2 I) = \sigma^2 R + \tau^2 I, \tag{2.24}$$

donde I es la matriz identidad.

Ahora, supongamos que nuestro objetivo es predecir los valores de la señal en un localidad arbitraria, por ende nuestro objetivo de predicción es $T = S(x)$. Entonces, (T, Y) es también una Gaussiana multivariante y obtenemos la predicción de mínimo error cuadrático medio $\hat{T}$ usando el siguiente resultado estándar en la distribución Gaussiana multivariante.

Teorema 2.5.2. *Sea $X = (X_1, X_2)$ una Gaussiana multivariante conjunta, con un vector de media $\mu = (\mu_1, \mu_2)$ y una matriz de covarianza*

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

esto es, $X \sim MVN(\mu, \Sigma)$. Entonces, la distribución condicional de X_1 dada X_2 es también una Gaussiana multivariada, $X_1|X_2 \sim MVN(\mu_{1|2}, \Sigma_{1|2})$ donde

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$$

y

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Diggle y Ribeiro [6] dicen que para aplicar el teorema 2.5.2 a nuestro problema de predicción, note que (T, Y) es una Gaussiana multivariada con un vector de media $\mu\mathbf{1}$

y una matriz de varianza

$$\begin{bmatrix} \sigma^2 & \sigma^2 r' \\ \sigma^2 r & \sigma^2 V \end{bmatrix}$$

en donde r es un vector con elementos $r_i = \rho(\|x - x_i\|) : i = 1, \ldots, n$ y V es dado por (2.24). Entonces, el Teorema 2.5.2 con $X_1 = T$ y $X_2 = Y$ da como resultado que el predictor de mínimo error cuadrático medio para $T = S(x)$ es

$$\hat{T} = \mu + r' V^{-1}(Y - \mu \mathbf{1}) \tag{2.25}$$

con una predicción de la varianza

$$Var(T|Y) = \sigma^2(1 - r' V^{-1} r). \tag{2.26}$$

Ellos mencionan que debemos tener en cuenta que en el marco especial de la distribución Gaussiana multivariante, la varianza condicional no depende de Y, por lo tanto el predictor de mínimo error cuadrático medio obtenido es igual a la varianza de predicción.

2.5.4. Kriging

Schabenberger y Pierce [19] indican que las técnicas clásicas de Kriging son los métodos para la predicción de $S(x_0)$ basados en la combinación de suposiciones sobre el modelo en el espacio con requisitos sobre el predictor $p(S; x_0)$. Además, estos autores mencionan que el nombre se reconoce por la influencia de D.G. Krige en el 1951. El conjunto habitual de requisitos según Schabenberger y Pierce [19] son:

1. $p(S; x_0)$ es una combinación lineal de los valores observados $S(x_1), \cdots, S(x_n)$.

2. $p(S; x_0)$ es insesgado en el sentido de que $E[p(S; x_0)] = E[S(x_0)]$.

3. $p(S; x_0)$ minimiza el predictor de error cuadrático medio

El requisito (i) establece que los predictores tiene la forma general.

$$p(S; x_0) = \sum_{i=1}^{n} \lambda(x_i) S(x_i),$$

donde $\lambda(x_i)$ es el peso asociado con la observación en la localidad x_i. En relación a otros pesos, $\lambda(x_i)$ determina cuando la observación $S(x_i)$ contribuye al valor de predicción en la localidad x_0. Schabenberger y Pierce [19] dicen que los $\lambda(x_i)$ son llamados *pesos de predicción*, o *pesos kriging*. Esta propiedad antes mencionada no implica que los pesos kriging son positivos. Para satisfacer los requerimiento 2 y 3, los pesos son escogidos para minimizar

$$E[\{S(x_0) - \sum_{i=1}^{n} \lambda(x_i) S(x_i)\}^2]$$

sujeto a ciertas restricciones que garantizan la insesgabilidad. Schabenberger y Gotway [18] enfatizan que estas restricciones dependen de las suposiciones del modelo. Los métodos básicos de kriging son simple, ordinario y universal.

$$S(x) = \mu(x) + \delta(x)$$

Los métodos de kriging se distinguen por la suposiciones de la estructura de la media en el modelo espacial. El kriging simple $\mu(x)$ es conocido, mientras que para los kriging

ordinario y universal $\mu(x) = \mu$, μ desconocido. En cambio, la diferencia entre el kriging ordinario la media es constante mientras que para el kriging universal la media es variable $\mu(x) = \mathbf{x}'(x)\beta$, β desconcido. Pero para todos los kriging básicos se supone que los parámetros de correlación son conocidos y donde $\delta(x)$ es estacionario de segundo orden o intrínsecamente estacionario.

Kriging simple

En la terminología geoestadística tradicional, la construcción de la superficie $\hat{S}(x)$, donde $\hat{T} = \hat{S}(x)$ es dada por (2.25), la cual es llamada como *kriging simple*. La solución del problema de minimización de los requerimientos 2 y 3, si $\mu(x)$ es conocida se llama el predictor kriging simple (Matheron, 1971)

$$p_{KS}(S; x_0) = \mu(x_o) + \mathbf{c}'\Sigma^{-1}(S(x) - \mu(x)). \tag{2.27}$$

Note que $\mu(x_0)$ es un escalar, $S(x)$ y $\mu(x)$ son vectores. El predictor kriging simple es insesgado por que $\mathrm{E}[p_{KS}(S; x_0)] = \mu(x_0) = \mathrm{E}[S(x_0)]$. El mínimo error cuadrático medio de un predictor kriging insesgado se conoce como *varianza kriging* o *error kriging*. Schabenberger y Pierce [19] establecen que la varianza kriging para un predictor kriging simple es

$$\sigma^2_{KS}(x_0) = \sigma^2 - \mathbf{c}'\Sigma^{-1}\mathbf{c} \tag{2.28}$$

donde σ^2 es la varianza del campo aleatoria en la ubicación x_0. Se asume que el campo aleatorio es estacionario de segundo orden, por lo tanto, $Var[S(x)] = Var[S(x_0)] = \sigma^2$ y la función de correlación existe. Otro supuesto es que la media es conocida por todo el campo aleatorio.

Kriging ordinario y kriging universal

Los kriging ordinario y kriging universal tienen en común que la media del campo aleatorio es desconocida y se expresa por un modelo lineal según Schabenberger y Pierce [19]. El caso general es $\mu(x) = \mathbf{x}'(x)\beta$ donde la media es una regresión lineal con variables $\mathbf{x}'(x)$. Usualmente las $\mathbf{x}'(x)$ son las coordenadas espaciales.

Para el kriging ordinario se asume que la media del campo aleatorio es constante en todas sus localidades espaciales, pero desconocida. Por lo tanto, se remplaza $\mathbf{x}'(x)$ por 1 y β por μ, la media desconocida. El predictor kriging ordinario es

$$p_{KO}(S; x_0) = \sum_{i=1}^{n} \lambda_{KO}(x_i) S(x_i)$$

el cual minimiza el predictor de error cuadrático medio sujeto a restricciones de insesgabilidad. Estas restricciones se notan en $\mathrm{E}[p_{KO}(S; x_0)] = \mathrm{E}\Big[\sum_{i=1}^{n} \lambda_{KO}(x_i) S(x_i)\Big] = \sum_{i=1}^{n} \lambda_{KO}(x_i)\mu$, el cual debe ser igual a μ para que $p_{KO}(S; x_0)$ sea insesgado. Como consecuencia $\sum \lambda(x_i) = 1$ para cualquier ubicación de destino x_i.

Si la media de un campo aleatorio es $\mu(x) = \mathbf{x}'(x)\beta$, no es suficiente requerir que los pesos kriging sumen a uno. En cambio se necesita

$$E\Big[\sum_{i=1}^{n} \lambda_{KU}(x_i) S(x_i)\Big] = \sum_{i=1}^{n} \lambda_{KU}(x_i)\mathbf{x}'(x)\beta = \mathbf{x}'(x_0)\beta.$$

Schabenberger y Pierce [19] escriben el modelo de kriging universal como $S(x) = \mathbf{X}(x)\beta + \delta(x)$ y el predictor como

$$p_{KU}(S; x_0) = \sum_{i=1}^{n} \lambda_{KO}(x_i) S(x_i) = \lambda' \mathbf{X}(x)\beta,$$

donde λ es el vector de pesos kriging universal. Para que $p_{KU}(S; x_0)$ sea insesgago se necesita que $\lambda'\mathbf{X} = \mathbf{x}'(x_0)$ según Schabenberger y Pierce [19]. Entonces, la minimización de

$$E\Big[\big(S(x_0) - \sum_{i=1}^{n} \lambda_{KO}(x_i)S(x_i)\big)^2\Big] \text{ sujeta a } \lambda'\mathbf{1} = 1$$

para encontrar los pesos kriging ordinario y la minimización de

$$E\Big[\big(S(x_0) - \sum_{i=1}^{n} \lambda_{KO}(x_i)S(x_i)\big)^2\Big] \text{ sujeta a } \lambda'\mathbf{X} = \mathbf{x}'(x_0)$$

para derivar los pesos kriging universales, esto es un problema de optimización restringido. Schabenberger y Pierce [19] resuelven el problema como un problema de minimización no restringida usando uno (kriging ordinario) o varios (krining universal) multiplicadores de Lagrange. Estos autores prefieren expresar los predictores de la siguiente forma

$$p_{KU}(S; x_0) = \mathbf{x}'(x_0)\hat{\beta} + \mathbf{c}'\Sigma^{-1}\big(S(x) - \mathbf{X}\hat{\beta}\big), \tag{2.29}$$

donde $\hat{\beta}$ es el estimador de mínimos cuadrados generalizados

$$\hat{\beta} = \big(\mathbf{X}'\Sigma^{-1}\mathbf{X}\big)^{-1}\mathbf{X}'\Sigma^{-1}S(x).$$

Un caso especial de (2.29) es cuando $\mathbf{x} = 1$, el predictor krigning ordinario es

$$p_{KO}(S; x_0) = \hat{\mu} + \mathbf{c}'\Sigma^{-1}\big(S(x) - \mathbf{1}\hat{\mu}\big)$$

Donde, $\hat{\mu}$ es el estimador de mínimos cuadrados generalizados de la media,

$$\left(\mathbf{1}'\Sigma^{-1}\mathbf{1}\right)^{-1}\mathbf{1}'\Sigma^{-1}S(x) = \frac{\mathbf{1}'\Sigma^{-1}S(x)}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}.$$

Schabenberger y Pierce [19] comentan que los predictores kringing son obtenidos sólo si los estimadores de mínimos cuadrados generalizados ($\hat{\beta}o\hat{\mu}$) son subtituidos.

Las varianzas kriging se calculan como

$$\sigma^2_{KU} = \sigma^2 - \mathbf{c}'\Sigma^{-1}\mathbf{c} + \left(\mathbf{x}(x_0) - \mathbf{X}'\Sigma^{-1}\mathbf{c}\right)'\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{x}(x_0) - \mathbf{X}'\Sigma^{-1}\mathbf{c}\right)$$

$$\sigma^2_{KO} = \sigma^2 - \mathbf{c}'\Sigma^{-1}\mathbf{c} + \left(1 - \mathbf{1}'\Sigma^{-1}\mathbf{c}\right)^2/\mathbf{1}'\Sigma^{-1}\mathbf{1}$$

Diggle y Ribeiro [6], Pyrcz y Deutsch [15] mencionan que el método de pesos kriging es uno de los métodos "declustering". Se debe a que los métodos dependen de la ponderación de los datos en la muestra, para explicar la referencia espacial. Este es un aspecto distintivo del predictor kriging en comparación con otros métodos de interpolación como ponderación al cuadrado inverso de la distancia. Diggle y Ribeiro [6] describen el efecto *enmascaramiento* ("masking") cuando dos ubicaciones de la muestra y la ubicación de destino son colineales, o cercana; cuanto más cerca están dos ubicaciones muestrales se da un peso grande, positivo, mientras que el punto más distante de la ubicación enmascarada le asigna un peso negativo. En general, ubicaciones enmascarada pueden ser pesos positivos, cero o negativos; dependiendo del modelo de correlación asumido. El peso bajo a puntos individuales tiene sentido intuitivo en este contexto. Esto es una consecuencia que asume la estructura de la correlación en el espa-

cio de los datos; es que dos puntos cercanos en el espacio transmitan más información que dos puntos aislados.

Otra cosa que se puede notar es que cuando la zona de influencia decrece, en general, la correlación correspondiente es más débil entre $S(x)$ y los Y_i, la suma de los pesos decrece. Cuando ϕ se acerca a cero, los pesos también se acercan a cero y $\widehat{S}(x) \approx \mu = 0$. Ya que $S(x)$ y Y son independientes, el valor observado Y no es de ayuda para predecir $S(x)$, según Diggle y Ribeiro [6]. Otra observación que hacen estos autores es que cuando el valor de τ^2 incrementa, los pesos de predicción se extienden progresivamente sobre las Y_i y el peso total decrece. Para un valor τ^2 grande, el ruido en los datos domina la señal, implicando que $S(x)$ y Y son aproximadamente independientes, todo los pesos se acercan a cero y $\hat{S}(x) \approx \mu = 0$ para toda ubicación x. Esto sugiere que cualquier conocimiento contextual relativo a la suavidad de la señal subyacente debe ser una consideración en la elección de una función de correlación para aplicaciones particulares, según Diggle y Ribeiro [6].

2.6. Estimación y predicción bayesiana para modelos lineales generalizados geoestadísticos

Diggle y Ribeiro [6] mencionan que el modelo geo-estadístico habitualmente se define a través de dos sub-modelos. Un sub-modelo para un proceso espacial sin observar $\{S(x) : x \in \mathbb{R}^2\}$, conocido como la señal. El otro sub-modelo sería los datos $Y = (Y_1, \cdots, Y_n)$ condicionales a $S(\cdot)$. Se utiliza θ para todos los parámetros descono-

cidos. Una notación para la especificación del modelo es

$$[Y, S|\theta] = [S|\theta][Y|S, \theta], \tag{2.30}$$

donde S es el proceso señal completo, $\{S(x) : x \in \mathbb{R}^2\}$. La notación de corchete, $[\cdot]$, significa "la distribución de" una variable aleatoria o variable encerrada dentro de los corchetes con una línea vertical que se denota condicional.

No existe una diferencia formal entre el proceso señal sin observar S y los parámetros del modelo θ ya que ambas son variables aleatorias sin observar para la predicción Bayesiana. Por esta razón, Diggle y Ribeiro [6] comienzan con una distribución conjunta especificada para tres entidades aleatorias: los datos, Y, la señal, S, y los parámetros del modelo, θ. Esto es una extensión de dos niveles de (2.30) a una de tres niveles

$$[Y, S, \theta] = [\theta][S|\theta][Y|S, \theta], \tag{2.31}$$

en donde $[\theta]$ es la distribución a priori para θ. Lawson [9] define la distribución a priori como una distribución asignada a un parámetro θ antes de visualizar los datos. Diggle y Ribeiro [6] mencionan que la distribución a priori debe reflejar el conocimiento científico acerca de los posibles valores de θ antes de la recolección e inspección de los datos. Cuando la distribución a priori tiene un conocimiento científico, se puede interpretar como una distribución a priori que provee "datos" adicionales para el problema, estos datos ayudan a mejorar la estimación o la identificación de los parámetros. Cuando la distribución a priori no refleja un conocimiento previo, usualmente, se la conoce como una distribución a priori no informativa.

Lawson [9] define la distribución a posteriori como el producto de la función de verosimilitud y de la distribución a priori. Esta distribución describe el comportamiento de los parámetros luego que se observan los datos y se da el supuesto a priori. Diggle y Ribeiro [6] dicen que la distribución a posteriori para S es la distribución condicional $[S|Y]$. Esta distribución condicional se obtiene por la aplicación de teorema de Bayes en la especificación del modelo, a partir de (2.31) en lugar de (2.30). Esto conduce al resultado

$$[S|Y] = \int [S|Y, \theta][\theta|Y]d\theta, \tag{2.32}$$

donde la distribución a posteriori es un promedio ponderado de la distribución predictiva "clásica" [6]. Estos pesos reflejan la incertidumbre a posteriori acerca de los valores en los parámetros del modelo θ. De la misma manera que en la predicción "clásica", la distribución predictiva para cualquier objetivo T es función de S, donde la transformación de S a T es determinística. Diggle y Ribeiro [6] indican que en las aplicaciones se simula la muestra y se calcula con un valor de muestra correspondiente a la distribución predictiva de T. También, la estimación de los parámetros Bayesianos se hacen a través del teorema de Bayes para producir una distribución a posteriori para θ, en la cual se combina la función de verosimilitud en (2.17) lo llama $L(\theta; y)$ con una distribución a priori $\pi(\theta)$. Esta distribución a posteriori tiene densidad

$$p(\theta|y) = \frac{l(\theta; y)\pi(\theta)}{\int L(\theta; y)\pi(\theta)d\theta} \tag{2.33}$$

Entonces, las inferencias sobre θ se pueden expresar como enunciados de probabilidad derivadas de la a posteriori según Diggle y Ribeiro [6].

Como se expuso anteriormente la implementación del método de máxima verosi-

militud en base a la inferencia para los modelos lineales generalizados geo-estadísticos se ve obstaculizada por la necesidad de evaluar integrales de alta dimensión. Para la inferencia bayesiana, Diggle y Ribeiro [6] presentan la manera usual alrededor de esta dificultad, los métodos de Monte Carlo, en particular, de la Cadena de Markov Monte Carlo (CMMC), para generar muestras de la distribución a posteriori o de predicción deseada.

Cadena Markov Monte Carlo

La Cadena Markov Monte Carlo (CMMC) es ampliamente utilizada en la inferencia bayesiana. Diggle y Ribeiro [6] indican que lo atractivo de CMMC es que proporcionan una forma de eludir el cálculo analítico y numérico bayesiano mediante la generación de muestras de la distribución a posteriori. También comentan que la CMMC logra esto mediante la simulación de la cadena de Markov construida de tal manera que la distribución de equilibrio de la cadena es la a posteriori requerida, o la distribución predictiva bayesiana. Además, mencionan que es posible definir una forma general de construir la cadena para que cumpla con el requisito básico antes mencionado.

Diggle y Ribeiro [6] denotan a θ como el conjunto de parámetros que definen la estructura de covarianza del modelo, y por β los parámetros de regresión, con $S(\cdot)$, determina la esperanza condicional de Y. La parametrización de ellos asume que $\mathrm{E}[S(x)] = 0$, por ende, β siempre incluye el intercepto. Además, describen a S como el vector de valores de $S(x)$ en ubicaciones de datos x_i, a Y como el vector de mediciones Y_i correspondiente, y S^* el vector de valores de $S(x)$ en los lugares de predicción x. Incluso, advierten que se tenga en cuenta que en las aplicaciones, los lugares de predicción pue-

den o no incluir las ubicaciones de datos x_i. Por lo tanto, se supone que S^* y S son distintos. Sin embargo, los algoritmos de muestreo de la distribución predictiva de S^* generan automáticamente muestras de la distribución predictiva de S. Por lo tanto, si se requieren predicciones en las localidades de los datos, simplemente combinamos los valores muestreados de S^* y S. Para la estimación de parámetros, necesitamos generar muestras de la distribución a posteriori $[\theta, \beta|Y]$. Para la predicción, también se requiere muestras de la distribución a posteriori $[S^*|Y]$.

Estimación

S^* es irrelevante para la inferencia de los parámetros del modelo. Diggle y Ribeiro [6] dicen que un solo ciclo del algoritmo CMMC implica muestrear primero de $[S|\theta, \beta, Y]$, luego de $[\theta|S]$, y finalmente de $[\beta|S, Y]$. Además, la segunda etapa del ciclo puede descomponerse en una secuencia de muestras de las distribuciones condicionales univariadas $[S_i|S_{-i}, \theta, \beta, Y]$, donde S_{-i} denota el vector S con su i-ésimo elemento removido. En general, comenzando desde los valores iniciales de θ, β, S, y repitiendo el ciclo suficientes veces, eventualmente se estarían generando muestras de $[\theta, \beta, S|Y]$. Simplemente se requiere hacer caso omiso a los valores muestreados de S de la a posteriori $[\theta, \beta|Y]$ requerida.

Antes de empezar con el algoritmo, se presenta en una forma detallada cada una de las distribuciones condicionales según Diggle y Ribeiro [6] muestran. Para empezar, el teorema de Bayes implica inmediatamente que

$$[\theta|S] \propto [S|\theta][\theta], \tag{2.34}$$

y esto $[\beta|S,Y] \propto [Y|\beta,S][\beta]$. La independencia condicional del modelo lineal generalizado mixto implica que

$$p(Y|\beta,S) = \prod_{j=1}^{n} p(Y_j|\beta_j,S_j) \tag{2.35}$$

de donde se sigue que

$$p(\beta|S,Y) \propto \{\prod_{j=1}^{n} p(Y_j|\beta_j,S_j)\}\pi(\beta). \tag{2.36}$$

Por último, el teorema de Bayes en conjunto con la estructura de independencia condicional del modelo presenta $p(S_i|S_{-i},\theta,\beta,Y) \propto p(Y|S,\beta)p(S_i|S_{-i},\theta)$ y(2.35), después muestra

$$p(S_i|S_{-i},\theta,\beta,Y) = \{\prod_{j=1}^{n} p(Y_j|\beta_j,S_j)\}p(S_i|S_{-i},\theta). \tag{2.37}$$

Esto se debe a que $S(\cdot)$ es un proceso gaussiano, la distribución condicional $[S|\theta]$ en la ecuación (2.34) es una gaussiana multivariada, y $p(S_i|S_{-i},\theta)$ en la ecuación (2.37) es una densidad gaussiana univariada. Diggle y Ribeiro [6] dicen que esto facilita el acercamiento alterno de actualización por bloque de los valores de S en conjunto. En principio, se puede especificar las distribuciones a priori para θ y β en las ecuaciones (2.34) y (2.36). Se debe tener en cuenta también que $p(Y_j|S_j,\beta) = p(y;\mu_j)$, donde $\mu_j = h^{-1}\{d_j'\beta + S(x_j)\}$ y $h(\cdot)$ es la función de enlace del modelo lineal generalizado.

Diggle y Ribeiro [6] presentan el siguiente algoritmo. El propósito general del algoritmo es describir de forma más explícita de cada paso utiliza una versión de una clase de métodos conocidos como algoritmos de Metropolis-Hastings, desarrollas por Metropolis et al. y Hastings según Diggle y Ribeiro [6]. Estos algoritmos envuelven una

actualización de la muestra propuesta en que la actualización de la muestra se acepta o rechaza con una cierta probabilidad que se elige para garantizar la convergencia de la cadena a la distribución de equilibrio.

- Paso 0. Se eligen los valores iniciales de θ, β y S. Los valores iniciales de θ y β deben ser compatibles con sus respectivas distribuciones a priori. Para obtener valores iniciales de S se iguala cada Y_i a su esperanza condicional μ_i dado β y $S(x_i)$, y se resuelve para $S_i = S(x_i)$.

- Paso 1. Se actualizan todos los componentes del vector de parámetros θ:

  1. Se elige un nuevo valor propuesto de θ' mediante el muestreo uniforme de los parámetros de la función de correlación especificados por su distribución a priori.

  2. Se acepta θ' con probabilidad $\triangle(\theta, \theta') = \text{mín}\{\frac{p(S|\theta')}{p(S|\theta)}, 1\}$ de lo contrario se rechaza θ' y θ no se cambia.

- Paso 2. Se actualizan los valors de S:

  1. Se elige un nuevo valor propuesto de S_i' para el i-ésimo componente de S de la densidad de probabilidad condicional univariada Gaussiana $p(S_i'|S_{-i}, \theta)$, donde S_{-i} denota S con su i-ésimo elemento removido;

  2. Se acepta S_i', con probabilidad $\triangle(S_i, S_i') = \text{mín}\{\frac{p(y_i|S_i', \beta)}{p(y_i|S_i, \beta)}, 1\}$ de lo contrario se rechaza S_i' y S_i no se cambia;

  3. Se repiten los pasos 1 y 2 para todos $i = 1, \ldots, n$.

- Paso 3. Se actualizan todos los elementos del parámetro de regresión β:

1. Se elige un nuevo valor propuesto de β' a partir de la densidad condicional $p(\beta'|\beta)$;

2. Se acepta β' con probabilidad

$$\triangle(\beta, \beta') = \text{mín}\{\frac{\prod_{j=1}^{n} p(y_j|s_j, \beta')p(\beta|\beta')}{\prod_{j=1}^{n} p(y_j s_j, \beta)p(\beta'|\beta)}, 1\},$$

de lo contrario se rechaza β' y β no se cambia.

En este contexto Diggle y Ribeiro [6] comentan que las densidades condicionales $p(S_i|S_{-i})$ en el paso 2, y $p(\beta'|\beta)$ en el paso 3 se llaman kernels (núcleos) de la transición. Añaden que cualquier kernel brinda un algoritmo válido, pero la elección puede tener un gran impacto en la eficiencia computacional. Incluso que se tenga en cuenta que en el paso 2, el kernel de la transición es la distribución condicional modelada de S_i dado S_{-i}, que parece una opción natural. Mientras que en el paso 3 el kernel de la transición $p(\beta'|\beta)$ es esencialmente arbitrario.

Los Pasos 1-3 se repiten hasta que la cadena haya llegado a su distribución de equilibrio, luego del llamado "burn-in" del algoritmo. Además durante los pasos 1-3, el ciclo obtiene una muestra de la distribución a posteriori, $[\theta, S, \beta|Y]$, la que puede ser analizada con los métodos de kriging utilizando las propiedades de la muestra empírica como aproximaciones a las propiedades correspondientes de la distribución a posteriori. Diggle y Ribeiro [6] dicen que en principio estas muestras se pueden hacer arbitrariamente mediante el aumento de la longitud de la ejecución de la simulación. Sin embargo, a diferencia de otros métodos Monte Carlo, el algoritmo CMMC genera muestras dependientes, a menudo en la práctica la dependencia es fuerte, y la regla simple de duplicar el tamaño de la simulación para reducir a la mitad la varianza

Monte Carlo no se aplica. La manera habitual de mostrar la distribución a posteriori obtenida a partir del algoritmo CMMC es con un histograma o una estimación de la densidad con suavizados no paramétricas basada en los valores muestreados luego que el algoritmo haya convergido.

Predicción

Para la predicción de las propiedades de la señal, $S(\cdot)$, se generan muestras de la distribución condicional $[(S, S^*)|Y] = [S|Y][S^*|S, Y]$, según Diggle y Ribeiro [6] mencionan.

- Paso 4. Extraer una muestra aleatoria de la distribución gaussiana multivariada $[S^*|Y, \theta, \beta, S]$, donde (θ, S, β) son los valores generados en los pasos 1 a 3.

Sin embargo, nuestro modelo implica que S^* es condicionalmente independiente tanto de Y y β, dada S, y el paso 4, por tanto, se reduce a la simulación directa de la distribución Gaussiana $[S^*|S, \theta]$. Específicamente,

$$[S^*|S, \theta] \sim MVN(\Sigma_{12}^T \Sigma_{11}^{-1} S, \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}), \tag{2.38}$$

donde $\Sigma_{11} = Var(S)$, $\Sigma_{12} = Cov(S, S^*)$ y $\Sigma_{22} = Var(S^*)$. Tenga en cuenta que si se reduce la muestra CMMC, el paso 4 sólo es necesario cuando el valor de la muestra correspondiente de S se almacena para uso futuro.

Para la predicción de cualquier variable de interés de $T = T(S^*)$ sigue inmediatamente, el cálculo de $T_j = T(S^*_{(j)}) : j = 1, \ldots, m$ para dar una muestra de tamaño m de distribución predictiva $[T|Y]$, según se requiere. Aquí, $S^*_{(j)}$ denota el j-ésimo muestra

simulada de la distribución predictiva del vector S^*. Para predicción puntual, se puede aproximar el predictor (mínimo error cuadrático medio), $\mathrm{E}[T(S^*)|y]$, por la media de la muestra, $\hat{T} = m^{-1}\sum_{j=1}^{M} T(S^*_{(j)})$. Sin embargo, por lo general será preferible examinar el conjunto de la distribución de predicción, como se menciona anteriormente, en el contexto de la distribución a posteriori de parámetros del modelo.

Diggle y Ribeiro [6] dicen que siempre que sea posible, es conveniente sustituir el muestreo Monte Carlo por evaluación directa. Por ejemplo, si es posible calcular $\mathrm{E}[T(S^*|S_{(j)}]$ directamente, usaríamos la aproximación

$$E[T(S^*)|Y] \approx m^{-1}\sum_{j=1}^{m} E[T(S^*)|S_{(j)}],$$

reduciendo de este modo el error Monte Carlo debido a la simulación.

Capítulo 3

Metodología

3.1. Datos

En el 2000 en la Estación Experimental de Adjuntas, Puerto Rico, se condujo un estudio sobre la roya en el café por Macchiavelli y Rodríguez [10]. En donde se seleccionó un lote de café de aproximadamente siete años, de la variedad Caturra. Este lote es típico de las zonas productoras de café, además el lote es de laderas marcadas con hileras de árboles paralelos a la pendiente que contiene una gran variabilidad en la incidencia de la enfermedad (roya). La distancia de siembra para el lote Caturra fue de 1.83 x 1.22m.

La incidencia de la roya en cada árbol es la proporción entre hojas enfermas y el total de hojas muestreadas en el tercio central del árbol. Por ende, la respuesta de interés (índice de la enfermedad), δ_i, para cada árbol i:

$$\delta_i = \frac{número\ de\ hojas\ infectadas\ del\ tercio\ central\ del\ árbol\ i}{número\ total\ de\ hojas\ del\ tercio\ central\ del\ árbol\ i}.$$

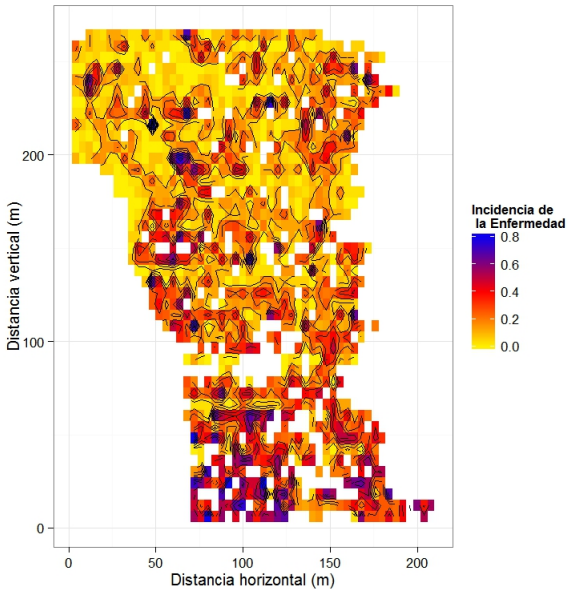

Figura 3.1: Mapa de incidencias observada de la enfermedad

En el estudio original de Macchiavelli y Rodríguez [10], se utilizó el procedimiento descrito por Hashizume et al. en 1975. El método se basa en cortar las hojas y determinar tanto las hojas infectadas como las no infectadas. Incluso, en el estudio se cortaron 40 hojas del tercio central de cada árbol y se evaluó la presencia de la enfermedad (el número de hojas por árbol fue determinado por el estudio original de Macchiavelli y Rodríguez [10]). Además, se muestrearon todos los 1,269 árboles del lote Caturra, es decir, por cada árbol del lote se evaluaron 40 hojas. Por lo tanto, como todos los árboles se evaluaron, la "verdadera" incidencia de la enfermedad de cada árbol en la parcela es conocida según Macchiavelli y Rodríguez [10].

3.2. Aplicación de técnicas geo-estadístas en base a distribución Gaussiana y en base a distribución binomial

Los datos de café en la Estación Experimental de Adjuntas, Puerto Rico, son proporciones binomiales basadas en muestras con n = 40. A los datos se les hizo un análisis geo-estadístico descriptivo. Este análisis se hizo a través de la librería geoR diseñada por Ribeiro y Diggle [17] en el software R [16]. Inicialmente, se convirtieron los datos a tipo "geodata" usando la función as.geodata. Entonces, se utiliza la función summary para obtener el análisis descriptivo y la función plot para visualizar los datos. Una de las salidas de la función plot es la Figura 3.1, el resto de la salida es un histograma de los valores observados y dos gráficos de los valores observados versus cada una de las coordenadas. El análisis conjunto con las gráficas produce un mejor entendimiento sobre nuestra variable respuesta (incidencia de la enfermedad).

Luego, se ajusta un semivariograma empírico con y sin tendencia espacial utilizando la función variog de la libreria geoR [17]. Además se hace una visualización del semivariograma empírico en la cual se puede obtener las posibles estimaciones iniciales a los posibles parámetros de la función de correlación. Incluso, brinda información sobre la posible función de correlación adecuada para el modelo. La función variog ajusta por el método de mínimos cuadrados ordinarios y calcula el semivariograma empírico utilizando los residuales.

Entonces, se ajustó el modelo lineal geo-estadístico basado en distribución normal.

Para hacer el ajuste se utiliza la función likfit. A la función se especifica la función de correlación con los respectivos valores iniciales de los parámetros de la función de correlación específicada y la tendencia espacial. Los métodos de estimación que utiliza la función son máxima verosimilitud y máxima verosimilitud restringida. La función likfit devuelve la estimación de los parámetros del modelo lineal geo-estadístico y distintos criterios de comparación de modelos (Akaike, Schwarz).

Se repite el uso de likfit con distintas funciones de correlación para obtener los distintos resultados, compararlos y así obtener la función de correlación apropiada. Finalmente, se utiliza la función krige.conv para predecir y obtener los mapas de incidencia predicha. Para crear los mapas de residuales y los mapas de la incidencia predicha de la enfermedad se uso la librería ggplot2 [20], en la cual se usó la función qplot.

A continuación, se ajustó el modelo lineal generalizado geo-estadístico basado en distribución binomial. Ahora usando la función binom.krige.bayes en la librería geoRglm [4], se estimaron los parámetros de correlación y la incidencia predicha de la enfermedad. A la función binom.krige.bayes hay que asignarle los componentes del modelo, las distribuciones a priori y los parámetros del algoritmo de CMMC. Esto se hace a través de las funciones model.glm.control, prior.glm.control y mcmc.control. En la función model.glm.control se provee la función de correlación y la tendencia lineal, en la función prior.glm.control se provee las distribuciones a priori y en la función mcmc.control se asignan los parámetros del algoritmo de CMMC. En el ajuste del modelo lineal generalizado geo-estadístico basado en distribución binomial se usó la misma función de correlación apropiada para el mejor ajuste del modelo lineal geo-estadístico basado

en distribución normal, el mismo efecto pepita estimado del mejor ajuste del modelo lineal geo-estadístico basado en distribución normal y la misma tendencia lineal. Al igual que en el modelo lineal geo-estadístico basado en distribución normal, el mapa de residuales y el mapa de la incidencia predicha de la enfermedad se obtuvieron con la función qplot.

Finalmente se hace un análisis comparativo de los resultados de los modelos lineales geo-estadísticos basados en la distribución normal y modelo lineal generalizado geo-estadístico basado en la distribución binomial. Esto se hace comparando sus estimaciones de los parámetros de correlación, mapas de incidencia predicha, mapa de residuales y sumas del cuadrados de error. Además, con los resultados verificamos qué modelo predice mejor los foco de incidencia de la enfermedad.

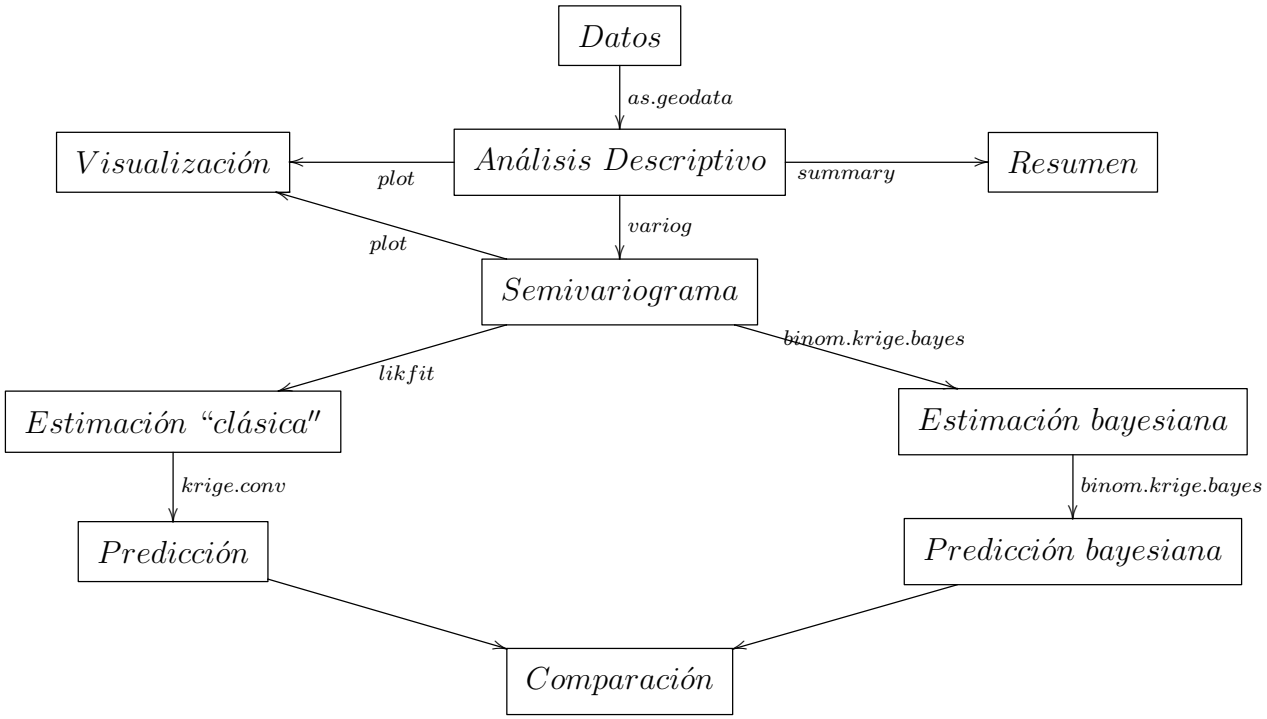

Figura 3.2: Diagrama de flujo.

Capítulo 4

Resultados y discusión

4.1. Análisis descriptivo

En el lote Caturra la distancia de dos árboles cualesquiera varía entre 4.0m y 321.7m. La incidencia de la enfermedad tiene media de 0.1939, mediana de 0.15 y desviación estándar de 0.1744.

Analizando la Figura 4.1, se puede observar asimetría ya que no hay ningún valor atípico ("outlier") obvio. Ahora, note que en la Figura 3.1, se puede distinguir un crecimiento de incidencia de la enfermedad del extremo norte hasta el sur de la región de estudio, el cual muestra una posible tendencia espacial de Norte a Sur. Esto sugiere que puede ser apropiado hacer un modelo de superficie de tendencia. Entonces para poder analizar la posible tendencia espacial más a fondo, en la Figura 4.2 se crean dos gráficos de los valores de los datos versus sus respectivas coordenadas (un gráfico para la coordenada x y otro para la coordenada y). Ambos gráficos se añade un ajuste polinomial lineal local. Este ajuste muestra una posible tendencia espacial lineal. Esto

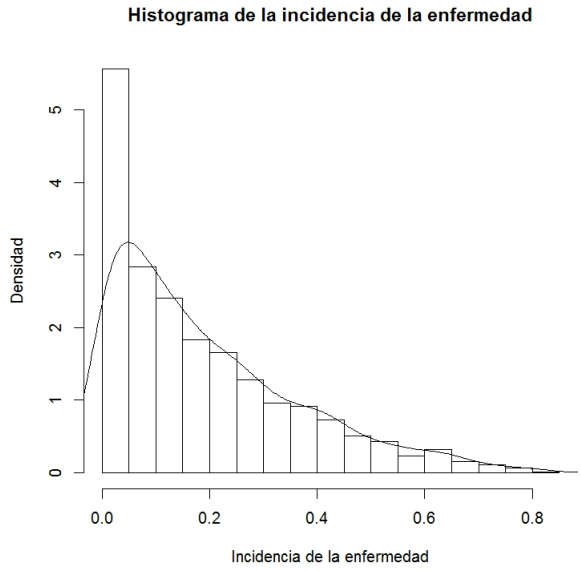

Figura 4.1: Histograma de la incidencia de la enfermedad observada.

sugiere la necesidad de incluir en el modelo una media variable en el espacio, o tal vez diferente comportamiento cualitativo en las diferentes sub-regiones.

Después se ajustan dos semivariogramas empíricos [17] a los datos de la incidencia de la enfermedad. Uno de los semivariogramas empírico [17] se ajustó con media constante y el otro con media variable. En el semivariograma empírico con media variable, se ajustó con una tendencia lineal. Los semivariogramas empíricos se visualizan en la Figura 4.3. Note que la forma del semivariograma empírico con media variable tiene la forma típica de un semivariograma teórico. Esto afirma un poco más el supuesto de ajustar un modelo de superficie de tendencia. En los semivariogramas empíricos obtenemos valores iniciales para los parámetros de la función de correlación. Para el semivariograma empírico con media constante, se eligió un alcance inicial de 250m y una meseta inicial de 0.06. En cambio, para el semivariograma empírico con media

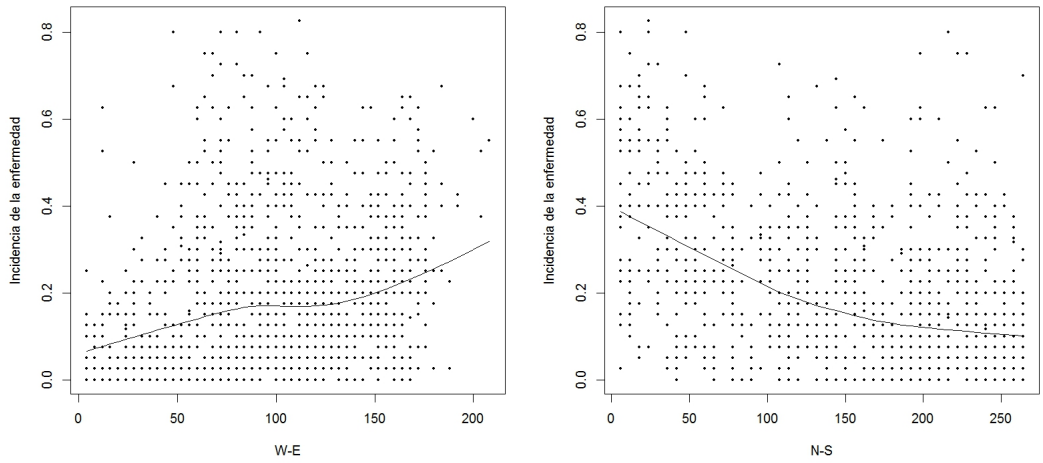

Figura 4.2: Incidencia de la enfermedad versus las coordenadas

variable, se eligió un alcance inicial de 200m y una meseta inicial de 0.03. Otra cosa que se observa es que se fijó la distancia máxima de los semivariogramas empíricos, ya que a la distancia mayor que 250m hay pocos datos y su correlación espacial es casi inexistente.

4.2. Comparación de Modelos

Ajuste suponiendo distribución normal

Se comienza ajustando varios modelos paramétricos asumiendo que la incidencia de la enfermedad se distribuye normal. Utilizando distintas funciones de correlación (exponencial, esférica y Matérn), se ajustaron los modelos exponencial, gaussiano, esférico y Matérn con $\kappa = 1.5, 2.5$. Para poder ver si es necesario una tendencia espacial, se ajustaron los modelos anteriormente mencionados bajo el supuesto de media constante

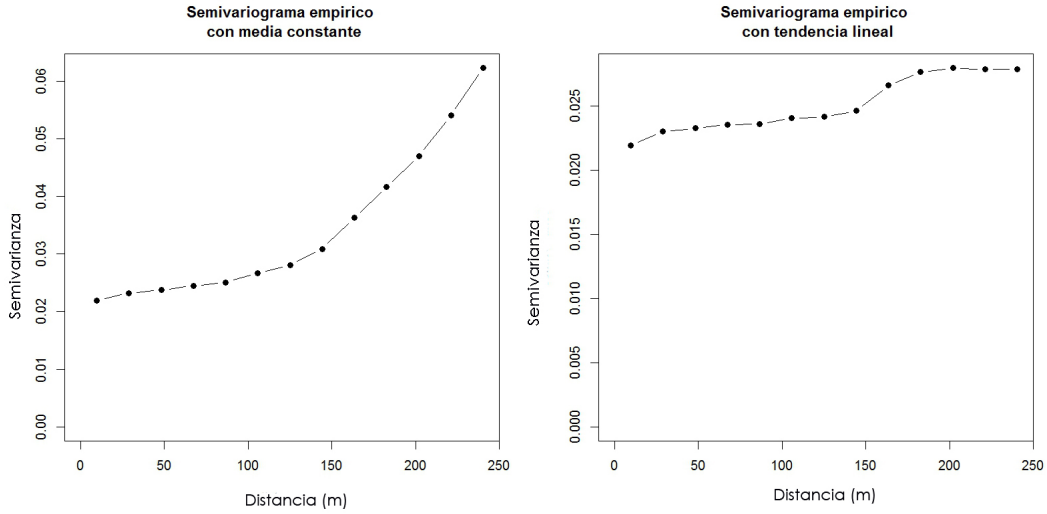

Figura 4.3: Semivariograma con media constante y semivariograma con tendencia lineal

y media variable.

Primero se considera con el supuesto de media constante. Se ajustaron los modelos con el método de máxima verosimilitud utilizando los valores iniciales obtenidos del semivariograma empírico y se obtuvieron los valores de AIC, BIC y log de máxima verosimilitud para los distintos modelos. En el caso del modelo gaussiano se obtuvo un error en la función de geoR (sistema es singular, el sistema singular no tiene solución, ya que el determinate de la matriz es cero). Los valores de AIC, BIC y log de máxima verosimilitud se visualizan en la parte superior del cuadro 4.1. También en el panel izquierdo en la Figura 4.4 se muestra el semivariograma ajustado a cada uno de los distintos modelos. La Figura 4.4 sugiere que los modelos Matérn $\kappa = 2.5$ y Matérn $\kappa = 1.5$ no tienen un buen ajuste. En cambio, los modelos Esférico y Exponencial son comparables. Suponiendo que la media es constante, cuando analizo la parte superior del cuadro 4.1 no se puede concluir cuál modelo es más apropiado debido a que su AIC

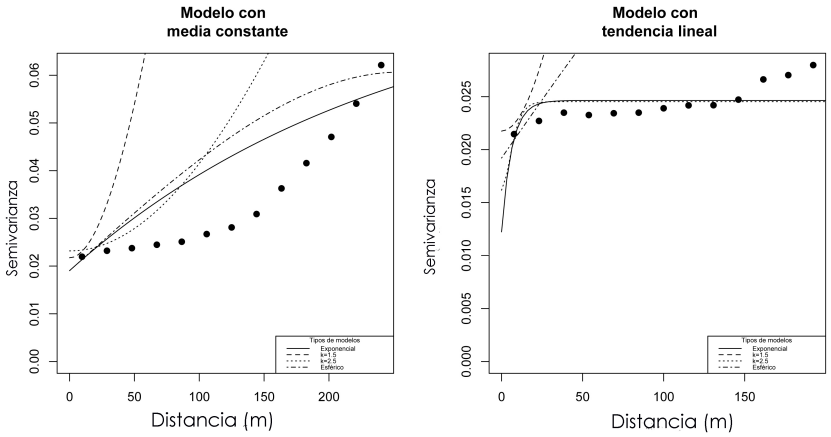

Figura 4.4: Ajuste visual de los modelos a los semivariogramas empíricos

y BIC son equivalentes.

Usando las mismas funciones de correlación, pero considerando la media es variable, se ajustaron los mismos modelos utilizando el método de máxima verosimilitud para obtener los valores de AIC, BIC y log de máxima verosimilitud. El panel derecho en la Figura 4.4 muestra el semivariograma ajustado a cada uno de los distintos modelos. La Figura 4.4 sugiere que los modelos Matérn ($\kappa = 1.5$) y Esférico no tienen un buen ajuste. Entonces, al analizar el cuadro 4.1 los modelos Matérn $\kappa = 2.5$ y Exponencial son comparables según sus AIC y BIC. En base al supuesto de que la media es variable, aún no se puede concluir cuál modelo es más apropiado según la visualización, AIC y BIC debido a que la diferencia no es notable.

Finalmente, se comparan los modelos paramétricos con media constante y variable. La estimación de los modelos paramétricos se hizo a través del método de máxima verosimilitud para poder comprar los ajustes. Entonces, analizando el cuadro 4.1 se observa que los modelos con media variable tienen valores de AIC y BIC menores que

Modelos con media constante			
Modelo	logL	AIC	BIC
Sin correlación en el espacio	415.9	-827.9	-817.6
Exponencial	601.9	-1196	-1175
Matérn $\kappa = 1.5$	581.9	-1156	-1135
Matérn $\kappa = 2.5$	570.9	-1134	-1113
Esférico	601.8	-1196	-1175
Modelos con tendencia lineal			
Modelo	logL	AIC	BIC
Sin correlación en el espacio	553.8	-1100	-1079
Exponencial	618.8	-1226	-1195
Matérn $\kappa = 1.5$	583.1	-1154	-1123
Matérn $\kappa = 2.5$	616.0	-1220	-1189
Esférico	603.9	-1196	-1165

Cuadro 4.1: Valores de AIC, BIC y log máxima verosimilitud para las a distintas función de correlación

los modelos de media constante. Se puede concluir que existe variabilidad asociada a la posición en el espacio; por lo tanto, es necesario añadir una tendencia espacial.

Hay que determinar por lo tanto qué modelo es mejor bajo el supuesto de media variable, ya que no hay una diferencia notable en los valores de AIC, BIC y log máxima verosimilitud entre el modelo exponencial y modelo Matérn con $\kappa = 2.5$. Nuevamente, se ajustan el modelo exponencial y el modelo Matérn con $\kappa = 2.5$ utilizando el método de máxima verosimilitud restingida. Los valores de AIC, BIC y log máxima verosimilitud restingida para modelo Matérn con $\kappa = 2.5$ son -1152, -1121 y 582.1 respectivamente. En cambio, los valores de AIC, BIC y log máxima verosimilitud restingida para el modelo exponencial son -1206, -1175 y 609 respectivamente. Después, se hace un análisis de residuales, el cual se visualiza en la Figura 4.5. Nota que los valores de AIC, BIC para el modelo exponencial son menores que para el modelo Matérn con

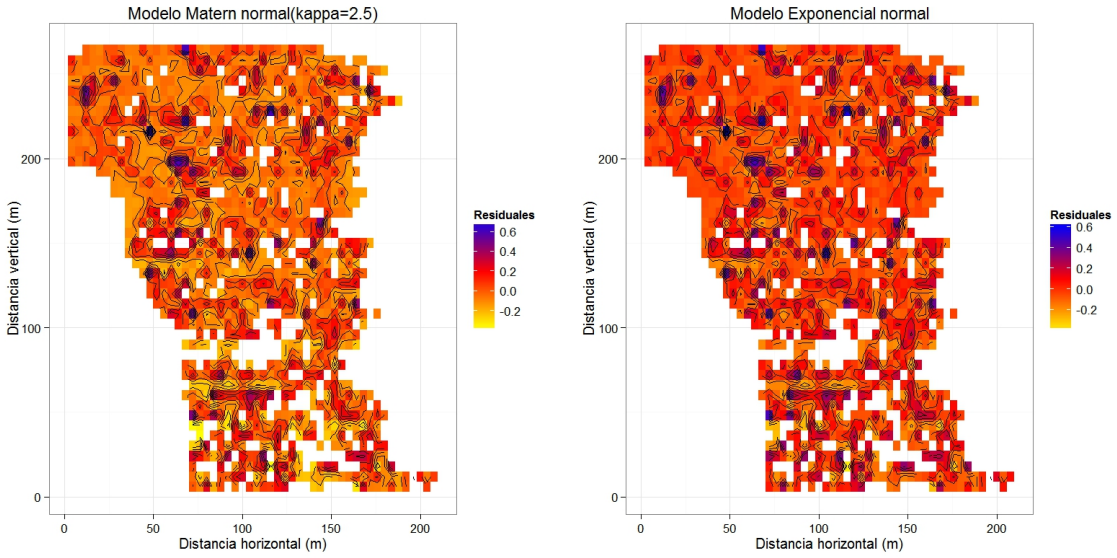

Figura 4.5: Mapa de residuales para el modelo exponencial normal y para el modelo Matérn ($\kappa = 2.5$) normal

$\kappa = 2.5$. Además, los residuales del modelo exponencial están más dispersos que los del modelo Matérn con $\kappa = 2.5$. En conclusión el modelo exponencial parece ajustar mejor que el modelo Matérn con $\kappa = 2.5$.

Ajuste suponiendo distribución binomial

Después, se ajustó un modelo lineal generalizado geo-estadístico asumiendo que la incidencia de la enfermedad se distribuye Binomial. El modelo se ajustó a través de estimación bayesiana. Entonces se especificaron las distribuciones a priori para los parámetros β, σ^2 y ϕ. La distribución a priori que se fijó para σ^2 es uniforme y para β es no informativa. En cambio, para ϕ se utilizaron distintas distribuciones a priori (uniforme, recíproca y recíproca al cuadrado). La selección de la distribución a priori para ϕ se basó en la tasa de aceptación de β y ϕ en conjunto. La distribución a priori

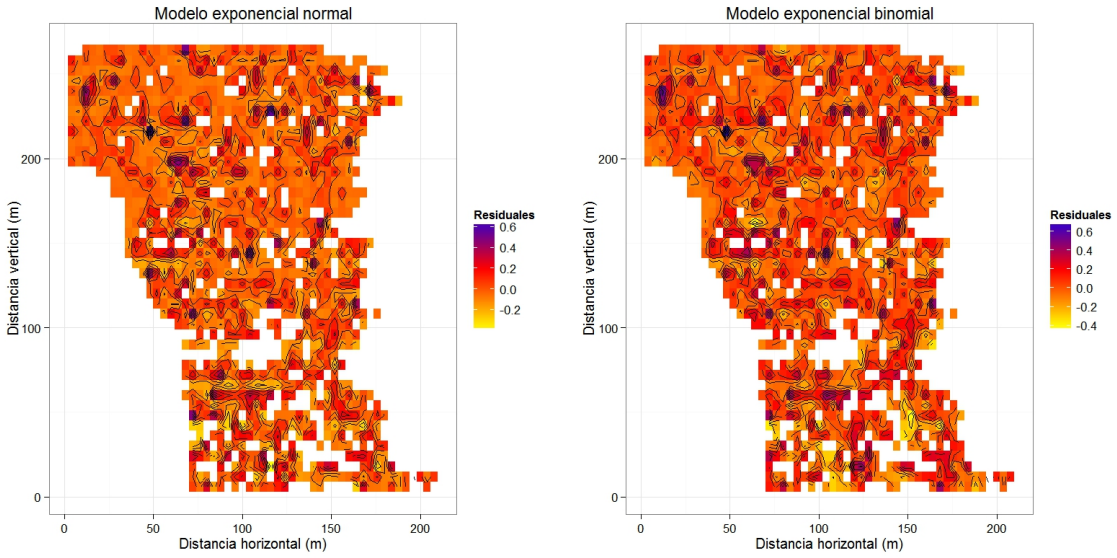

Figura 4.6: Mapa de residuales para el modelo exponencial normal y para el modelo exponencial binomial

con mejor tasa de aceptación es la recíproca al cuadrado. Luego, se fijó el valor de $\tau^2 = 0.02$ usando la estimación de τ^2 en el modelo exponencial normal. Se ajustó el modelo con las distribuciones a priori antes mencionadas, el valor de τ^2 y una función de correlación exponencial.

Finalmente, se comparan el modelo exponencial normal y el modelo exponencial binomial. En el modelo exponencial normal la estimación de máxima verosimilitud restingida para los parámetros de correlación son $\hat{\sigma}^2 = 0.0517$, $\hat{\phi} = 199.997$ y $\hat{\tau}^2 = 0.0189$. Mientras que la estimación de máxima verosimilitud restringida para los parámetros de la media son $\hat{\beta}_0 = 0.3730$, $\hat{\beta}_1 = 0.0004$ y $\hat{\beta}_2 = -0.0012$. En cambio, para el modelo exponencial binomial la estimación de máxima verosimilitud restingida para los parámetros de correlación son $\hat{\sigma}^2 = 20$, $\hat{\phi} = 290$ y $\hat{\tau}^2 = 0.02$. Entretanto, la estimación de máxima verosimilitud restingida para los parámetros de la media son $\hat{\beta}_0 = -0.7183$,

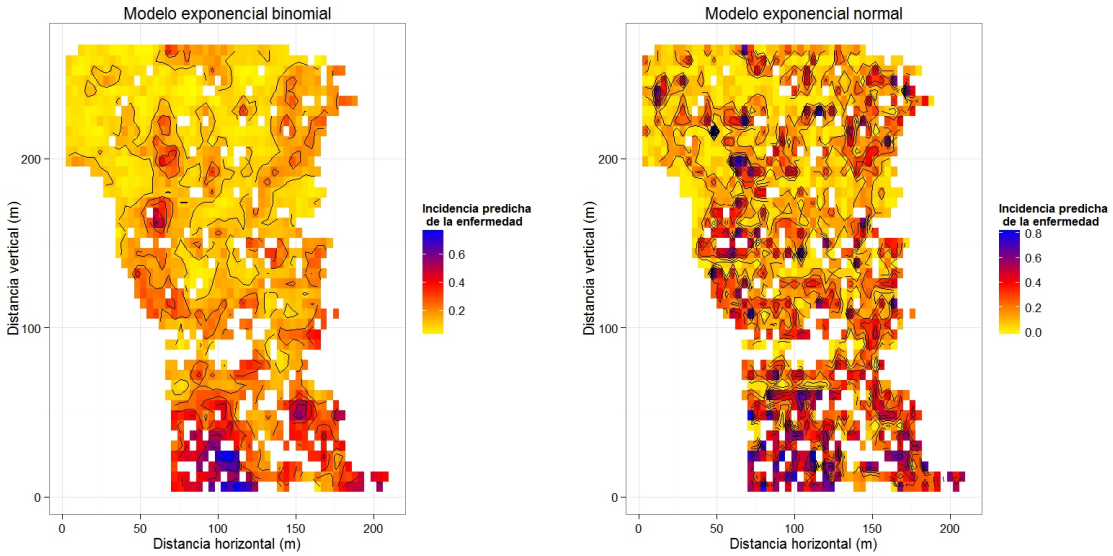

Figura 4.7: Mapa de incidencia predicha de la enfermedad para el modelo exponencial normal y para el modelo exponencial binomial

$\hat{\beta}_1 = 0.0035$ y $\hat{\beta}_2 = -0.0066$. Las estimaciones de los modelos no se pueden comparar directamente por que están en distintas escalas, excepto $\hat{\phi}$.

Después, se hizo un mapa de residuales para ambos modelos, el mapa se visualiza en la Figura 4.6. Note que ambos mapas son parecidos en la región norte. Pero, note que los residuales para el modelo exponencial normal son más dispersos que los residuales del modelo exponencial binomial. Luego, se visualiza el mapa de incidencia predicha de la enfermedad en la Figura 4.7 para el modelo exponencial normal y para el modelo exponencial binomial. Ambos mapas de incidencia predicha de la enfermedad se hicieron utilizando el método de kringing universal. Entonces, analizando la Figura 4.7 se observa que la incidencia predicha de la enfermedad para el modelo exponencial normal tiene una forma parecida al mapa de incidencia de la enfermedad observada, es decir, que cuando visualizamos ambos mapas no muestra mucha diferencia entre ellos.

Modelo	$\hat{\phi}$	$\sum_i^n (y_i - \hat{y}_i)^2$
Exponencial normal	199.997	21.2696
Exponencial binomial	290	25.0971

Cuadro 4.2: Zona de influencia y suma de cuadrado de error

Además, cuando se comparan la suma de cuadrados de error de ambos modelos en el cuadro 4.2, la del modelo exponencial normal es menor que la del modelo exponencial binomial. También se hace un mapa de incidencia predicha de la enfermedad en una región de 265m $\times$ 265m, se utilizó un kriging universal en el mapa, donde se visualiza en la Figura 4.8. Se observa que el mapa para el modelo exponencial normal es más suave y en los puntos donde se encuentran los datos observados existe una mayor influencia de esto. La influencia se debe a que la estimación bayesiana toma menos consideración de los datos observados en comparación de los datos observados a la estimación "clasica". En conclusión, considerando lo antes mencionado selecciono el modelo exponencial normal como el mejor modelo que ajusta los datos.

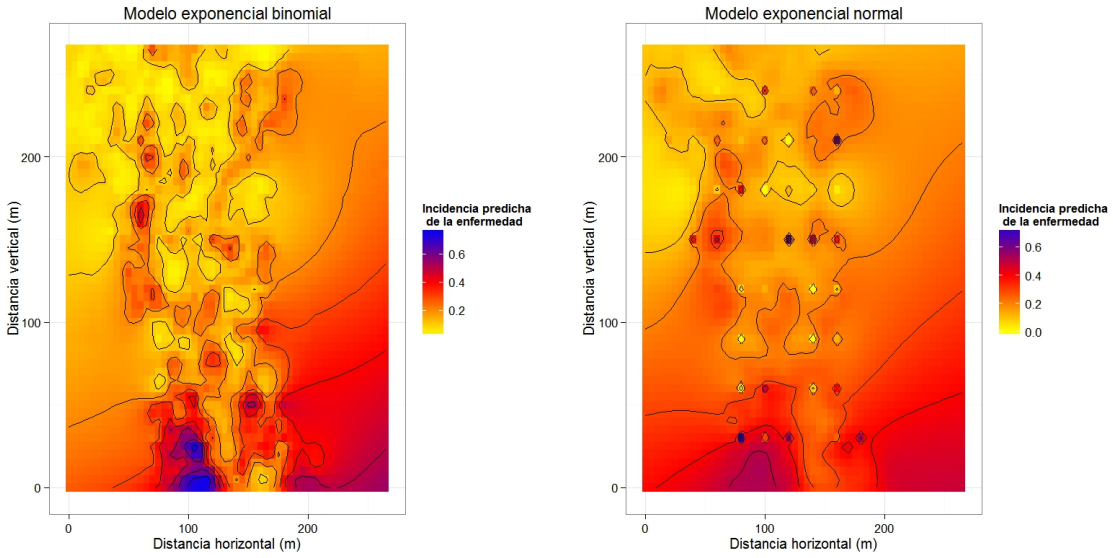

Figura 4.8: Mapa de incidencia predicha de la enfermedad para el modelo exponencial normal y para el modelo exponencial binomial para región 265m $\times$ 265m.

Capítulo 5

Conclusión y trabajos futuros

5.1. Conclusión

En la Estación Experimental de Adjuntas, Puerto Rico se obtuvieron los datos que permitieron aplicar técnicas geo-estadísticas para establecer la relación entre la incidencia de la enfermedad y la ubicación en el área en donde se encuentra la planta de café. Esta relación permite entender y comprender cómo la enfermedad se dispersa en la región de estudio. También, se identificó los focos de la roya en la región de estudio. Se compararon distintos modelos para identificar que modelo ajusta mejor los datos, el cual brindará mejor comprensión del comportamiento de la incidencia de la roya en la región de estudio.

Se ajustó un modelo asumiendo distribución Gaussiana. A este modelo se le ajustaron distintas funciones de correlaciones: la exponencial, esférica y Matérn. Además, se ajustó el modelo con media constante y media variable. Entonces, se compararon las funciones de correlaciones a través de los criterios de Akaike, Schwartz y análisis

de residuales. Obteniendo que la función de correlación exponencial es la mejor que describe la estructura de correlación de los datos, ya que los valores de AIC y BIC son menores. Luego se ajustó otro modelo asumiendo distribución binomial, utilizando la función de correlación exponencial. Este modelo se ajustó utilizando la estimación bayesiana, en donde se fijaron distribuciones a priori para σ^2 es uniforme, para β es no informativa y para ϕ es recíproca al cuadrado. Además, se fijó el valor de $\tau^2 = 0.02$ usando la estimación de τ^2 en el modelo exponencial normal.

Después se compararon los residuales, mapa de incidencia predicha de la enfermedad, suma de cuadrado de error. En conclusión, comparando los mapas de residuales, mapas de incidencia predicha y la suma de cuadrado de error, el modelo exponencial normal es el mejor modelo que ajusta los datos. Para interpretar las estimaciones de los parámetros de correlación en el modelo exponencial normal, si se fija la distancia vertical, por cada metro que aumenta la distancia horizontal la incidencia de la enfermedad aumenta 0.0004 en promedio. Igualmente, si se fija la distancia horizontal por cada metro que aumenta la distancia vertical la incidencia de la enfermedad disminuye 0.0012 en promedio. Otra dato importante es que $\hat{\phi} = 199.9$ el cual implica que los árboles que se encuentran a más de 199.9m de distancia tienen poca o casi nada de correlación con respecto a su incidencia de la enfermedad. Además, se nota una alta incidencia predicha de la enfermedad en la región sur, donde se visualiza en la Figura 4.8.

5.2. Trabajos futuros

Durante esta tesis, surgieron algunas ideas de trabajos que pueden ser realizados en un futuro. Algunas de las ideas que se pueden realizar en un futuro son:

- Añadir otras funciones de correlación a los modelos estudiados.
- La utilización de modelos Gaussianos transformados.
- Analizar desde el punto de vista agronómico y económico los resultados de los mapas de la inicidencia predicha de la enfermedad.

Bibliografía

[1] Agresti, Alan, *Categorical Data Analysis*, Second Ed. John Wiley & Sons, Hoboken, New Jersey, 2002.

[2] Casella, George y Berger, Roger L., *Statistical Inference*, Second Ed. Cengage Learning, Stamford, Connecticut, 2001.

[3] Chilès, J.P. y Delfiner, P., *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons, New York, 1999.

[4] Christensen, O.F. y Ribeiro Jr., P.J., *geoRglm: A package for generalised linear spatial models*, R-NEWS, Vol 2, No 2, 26-28. (2002).

[5] Cressie, N.A.C., *Statistics for Spatial Data*, Revised Ed. John Wiley & Sons, New York, 1993.

[6] Diggle, Peter J. y Ribeiro, Paulo Jr., *Model-based Geostatistics*, Springer Science & Business Media, New York, 2007.

[7] Faraway, Julian J, *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC, Boca Raton, Florida, 2006.

[8] García, Héctor A, *Café en Puerto Rico*. Recuperado el 22 de febrero de 2013 de http://www.proyectosalonhogar.com/El_Cafe/Indice.htm

[9] Lawson, Andrew B., *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*, Second Ed. Chapman & Hall/CRC Interdisciplinary Statistics, Boca Raton, Florida, 2013.

[10] Macchiavelli, Raúl E. y Rodríguez, Rocío del P., *Methods for efficient estimation of rust incidence in coffee plantations*, J. Agric. Univ. P.R. 84 (1-2): 65-78 (2000).

[11] Madden, Laurence V., Hughes, Gareth y van den Bosch, Frank *The Study of Plant Disease Epidemics*, The American Phytopathological Society, St. Paul, Minnesota, 2008.

[12] Matheron, Georges *La teoría de las variables regionalizadas y sus aplicaciones*, Los Cuadernos del Centro de Morfología Matemática de Fontainebleau, Fascículo 5,1970.

[13] McCullagh, P., y Nelder, J. A., *Generalized linear models*, Second Ed. Chapman & Hall, London, 1989.

[14] Monroig Inglés, Miguel F. y Rodríguez, Rocío del P. , *Manejo de la roya del cafeto*, Recuperado el 22 de febrero de 2013, de http://academic.uprm.edu/mmonroig/id22.htm

[15] Pyrcz, M. J., y Deutsch, C. V., *Declustering and Debiasing*, Centre for Computational Geostatistics, University of Alberta, Edmonton, Alberta, Canada

[16] R Development Core Team, *R: A language and environment for statistical computing* R Foundation for Statistical Computing, Vienna, Austria, 2005. *http://www.r-project.org.*

[17] Ribeiro Jr. y Diggle, P.J. *geoR: A package for geostatistical analysis*, R-NEWS Vol 1, No 2. ISSN 1609-3631, 2001

[18] Schabenberger, O. y Gotway, C.A., *Statistical Methods for Spatial Data Analysis*, Chapman & Hall/CRC, Boca Raton, Florida, 2005.

[19] Schabenberger, O. y Pierce, F.J., *Contemporary Statistical Models for the Plant and Soil Science*, Taylor & Francis, Boca Raton, Florida, 2002.

[20] Wickham, H., *ggplot2: elegant graphics for data analysis*, Springer New York, 2009. *http://had.co.nz/ggplot2/book*