



COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN
EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

PROGRAMA DE POSTGRADO EN SOCIOECONOMÍA, ESTADÍSTICA E
INFORMÁTICA ESTADÍSTICA

**PRUEBAS DE HIPÓTESIS PARA
PROCESOS GAUSIANOS ESPACIALES**

DAVID ISRAEL CELIS EUAN

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA
OBTENER EL GRADO DE:

DOCTOR EN CIENCIAS

MONTECILLO, TEXCOCO, EDO. DE MÉXICO

2017

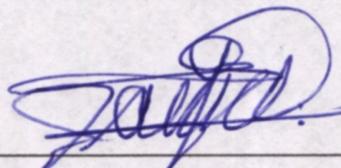
La presente tesis titulada: **Pruebas de hipótesis para procesos gaussianos espaciales**, realizada por: **David Israel Celis Euan**, bajo la dirección del Consejo Particular indicado ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

DOCTOR EN CIENCIAS

**SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA**

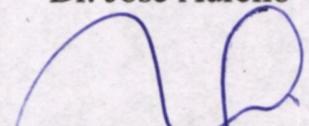
CONSEJO PARTICULAR

CONSEJERO



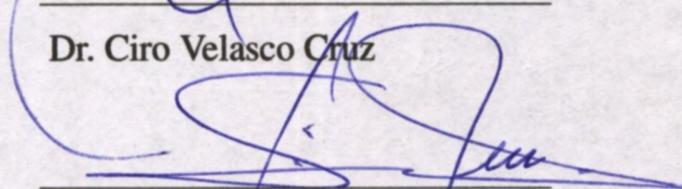
Dr. Jose Aurelio Villaseñor Alva

ASESOR



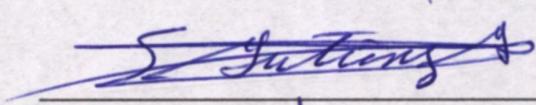
Dr. Ciro Velasco Cruz

ASESOR



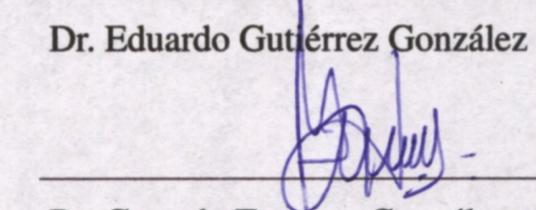
Dr. Gilberto Rendón Sánchez

ASESOR



Dr. Eduardo Gutiérrez González

ASESOR



Dr. Gerardo Terrazas González

Montecillo, Texcoco, Estado de México, Junio de 2017.

PRUEBAS DE HIPÓTESIS PARA PROCESOS GAUSIANOS ESPACIALES

DAVID ISRAEL CELIS EUAN
COLEGIO DE POSTGRADUADOS, 2017

RESUMEN

Los métodos estadísticos para el análisis de datos espaciales desempeñan un papel cada vez más importante. Con el paso de los años, estos métodos han evolucionado hasta convertirse en una disciplina independiente que continua creciendo y desarrollándose hasta producir un vocabulario propio. Es característico de la estadística espacial su inmensa diversidad metodológica. En parte, esto es debido a sus múltiples aplicaciones tales como en la geología, geografía, meteorología y otras áreas temáticas.

En este trabajo de investigación se tienen dos objetivos. El primer objetivo es proponer una prueba para probar la hipótesis de un proceso espacial gaussiano (o campo aleatorio gaussiano). Esta prueba ayuda a decidir si un conjunto de datos son una realización de un proceso espacial gaussiano con parámetros media y matriz de varianza-covarianza desconocidos, tomando en cuenta que esta matriz depende de un modelo de semivariograma. El segundo objetivo consiste en proponer una prueba de hipótesis para verificar si la media de los datos es constante. Esta prueba ayuda a decidir si los datos pueden ser usados para llevar a cabo Kriging Ordinario, ó Kriging Universal. Los parámetros de los modelos son estimados por los métodos de máxima verosimilitud y de máxima verosimilitud restringida.

En el primer caso se hizo una transformación de los datos usando un resultado de la distribución normal multivariada. A la transformación de los datos se le aplicó la prueba de Anderson Darling para decidir si tiene distribución normal estandar, lo cual implica normalidad multivariada en los datos no transformados.

En el segundo caso se usó una modificación de la estadística de Wald multivariada, como estadística de prueba, y el valor crítico fue obtenido por el método de bootstrap paramétrico.

Se estudió el tamaño y potencia de las pruebas por medio de simulación de Monte Carlo. Se realizó una aplicación a datos de lluvia del estado de Paraná, Brasil.

Palabras clave: Estadística espacial, Campo Aleatorio Gaussiano, Semivariograma, Tamaño de la prueba, Potencia de la prueba.

HYPOTHESIS TESTS FOR SPATIAL GAUSSIAN PROCESSES

DAVID ISRAEL CELIS EUAN
COLEGIO DE POSTGRADUADOS, 2017

ABSTRACT

Statistical methods for the analysis of spatial data play an ever increasingly important role. Over the years, these methods have evolved into an independent discipline that continues to grow and develop into a vocabulary on its own. The immense methodological diversity is a characteristic of spatial statistics. In part, this is due to its multiple applications such as in geology, geography, meteorology and other subject areas.

This research has two objectives. The first objective is to propose a hypothesis test with which it can be tested the hypothesis of a Gaussian spatial process (or Gaussian random field). This test helps to decide if a dataset is a realization of a Gaussian spatial process with unknown mean and variance-covariance matrix parameters, taking into account that this matrix depends on a semivariogram model. The second objective is to propose a hypothesis test to verify if the mean of the data is constant. This test helps to decide if the data can be used to perform Ordinary Kriging, or Universal Kriging. The parameters of the models are estimated by the maximum likelihood and the restricted maximum likelihood methods.

In the first case a transformation of the data was made by using some theory of the multivariate normal distribution. Then, the Anderson Darling test was applied to the transformed data to decide if they have a standard normal distribution, which implies multivariate normality in the untransformed data.

In the second case, a modification of the multivariate Wald statistic was used as the statistic test, and the critical value was obtained by using the parametric bootstrap method.

The size and power of the tests were studied by means of Monte Carlo simulation. An application to rainfall data from the state of Paraná, Brazil, was performed.

Key words: Spatial Statistics, Gaussian Random Field, Semivariogram, Test Size, Test Power.

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología, por el apoyo económico recibido para la realización de mis estudios de doctorado.

Al Colegio de Postgraduados, Campus Montecillo, por haberme brindado la oportunidad de realizar mis estudios de doctorado en sus instalaciones.

Al Dr. José Aurelio Villaseñor Alva, por el apoyo y los consejos recibidos durante la realización de este trabajo de investigación.

Al Dr. Ciro Velasco Cruz, por los consejos recibidos durante la realización de este trabajo.

Al Dr. Gilberto Rendón Sanchez, por el apoyo recibido durante la realización de este trabajo.

Al Dr. Eduardo Gutierrez González, por la amistad y el apoyo recibido en la realización de este trabajo.

Al Dr. Gerardo Terrazas González, por el apoyo recibido durante la realización de este trabajo.

DEDICATORIA

Al Dios y Padre de nuestro Señor Jesucristo, por su compañía, por su ayuda, por las bendiciones en mi vida.

A la memoria de mi madre Sofía Euan, que en paz descansa, por los deseos de superación que infundió en mí.

A mi padre José C. Celis, por haberme enseñado a trabajar arduamente.

A mis hermanos, Martín, Manuel, Juan, Lizza Minely y Jacqueline, por el cariño de hermanos, por la amistad y la compañía que he recibido de parte de ellos.

A mi hermana Lizza Minely, por la amistad, por que ha sido y es un gran apoyo en tiempos difíciles.

A mis amigas Mehida V.M., Zenaida M.J. y Patricia E.D., por su amistad durante los años del doctorado.

CONTENIDO

1	Introducción	1
2	Marco Teórico	5
2.1	Procesos Estocásticos espaciales y Campos Aleatorios	5
2.2	Campo Aleatorio Gaussiano	8
2.3	Una representación en términos de un modelo	8
2.4	El Semivariograma	10
2.5	Modelos de Covarianza estacionaria y el Semivariograma	11
2.5.1	Validez del modelo	11
2.5.2	Algunas funciones de covarianza	12
2.6	Estimacion de Máxima Verosimilitud en CAG	15
2.7	Estimacion de Máxima Verosimilitud Restringida en CAG	16
2.8	Predicción espacial y kriging	17
2.8.1	Kriging Ordinario	18
2.8.2	Kriging Universal	18
3	Una prueba para la hipótesis de Campo Aleatorio Gaussiano	20
3.1	Proposición	20

CONTENIDO

3.2	La familia Matérn de funciones de covarianza	21
3.3	Estimación de los parámetros del semivariograma Matérn	21
3.4	Construcción de una prueba para probar H_0	23
3.5	Estudio de simulación	24
3.5.1	Algoritmo para probar H_0	24
3.5.2	Simulación	25
3.6	Discusión de resultados	26
3.7	Una aplicación	40
4	Una prueba para la hipótesis de media constante de un Campo Aleatorio Gaussiano	42
4.1	Un modelo particular para $X(s)$	42
4.2	Prueba de hipótesis lineal para efectos fijos	44
4.3	Una prueba para probar Kriging Ordinario	44
4.4	Un estudio de simulación	45
4.4.1	Bootstrap paramétrico para obtener F_α	45
4.4.2	Simulación	46
4.5	Discusión de resultados	47
4.6	Una aplicación	57
5	CONCLUSIONES	60
	REFERENCIAS	61
	APÉNDICES	66
A.	Conceptos Teóricos	66

CONTENIDO

A1. Bootstrap Paramétrico	66
A2. La prueba de Anderson Darling	66
A3. Una breve revisión acerca de campos aleatorios no gaussianos	67
Apéndice B: Códigos en R	72

LISTA DE TABLAS

3.1	Tamaño estimado de la prueba PCAG. Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$ con parámetros $\beta_0 = 3$, $\beta_1 = 0$, $\beta_2 = 0$, $\sigma^2 = 4$, $\phi = 2$, $\tau^2 = 1$ y $\kappa = 1.1$. $H_0 : \mathbf{Z}(\mathbf{s}) \sim N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \kappa)$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula con el semivariograma Matérn. Los modelos de semivariograma están definidos en la subsección 2.5.2.	27
3.2	Tamaño estimado de la prueba PCAG. Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$ con parámetros $\beta_0 = 3$, $\beta_1 = 3$, $\beta_2 = 3$, $\sigma^2 = 4$, $\phi = 2$, $\tau^2 = 1$ y $\kappa = 1.1$. $H_0 : \mathbf{Z}(\mathbf{s}) \sim N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \kappa)$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula con el semivariograma Matérn. Los modelos de semivariograma están definidos en la subsección 2.5.2.	29
3.3	Tamaño estimado de la prueba PCAG. Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$ con parámetros $\beta_0 = 3$, $\beta_1 = 0$, $\beta_2 = 0$, $\sigma^2 = 4$, $\phi = 0.5$, $\tau^2 = 1$ y $\kappa = 1.1$. $H_0 : \mathbf{Z}(\mathbf{s}) \sim N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \kappa)$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula con el semivariograma Matérn. Los modelos de semivariograma están definidos en la subsección 2.5.2.	31
3.4	Tamaño estimado de la prueba PCAG. Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$ con parámetros $\beta_0 = 3$, $\beta_1 = 3$, $\beta_2 = 3$, $\sigma^2 = 4$, $\phi = 0.5$, $\tau^2 = 1$ y $\kappa = 1.1$. $H_0 : \mathbf{Z}(\mathbf{s}) \sim N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \kappa)$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula con el semivariograma Matérn. Los modelos de semivariograma están definidos en la subsección 2.5.2.	33

3.5	Tamaño estimado de la prueba PCAG. Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$ con parámetros $\beta_0 = 3, \beta_1 = 0, \beta_2 = 0, \sigma^2 = 4, \phi = 0.01, \tau^2 = 1$ y $\kappa = 1.1$. $H_0 : \mathbf{Z}(\mathbf{s}) \sim N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \kappa)$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula con el semivariograma Matérn. Los modelos de semivariograma están definidos en la subsección 2.5.2.	35
3.6	Tamaño estimado de la prueba PCAG. Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$ con parámetros $\beta_0 = 3, \beta_1 = 3, \beta_2 = 3, \sigma^2 = 4, \phi = 0.01, \tau^2 = 1$ y $\kappa = 1.1$. $H_0 : \mathbf{Z}(\mathbf{s}) \sim N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \kappa)$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula con el semivariograma Matérn. Los modelos de semivariograma están definidos en la subsección 2.5.2.	37
3.7	Potencia estimada de la prueba PCAG. Estudio de simulación de campos aleatorios no gaussianos, en una malla irregular de $(0, 50) \times (0, 50)$. $H_1 : \mathbf{Z}(\mathbf{s}) \approx N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula usando el semivariograma Matérn.	39
4.1	Tamaño estimado de la prueba para Kriging Ordinario. Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$, con $\beta_0 = 3, \beta_1 = 0, \beta_2 = 0, \sigma^2 = 4, \phi = 2$ y $\tau^2 = 1$. $H_0 : \beta_1 = \beta_2 = 0, \beta_0 > 0$. Los distintos modelos de semivariograma están definidos en la subsección 2.5.2.	48
4.2	Tamaño estimado de la prueba de Kriging Ordinario. Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$, con $\beta_0 = 3, \beta_1 = 0, \beta_2 = 0, \sigma^2 = 4, \phi = 0.5$ y $\tau^2 = 1$. $H_0 : \beta_1 = \beta_2 = 0, \beta_0 > 0$. Los distintos modelos de semivariograma están definidos en la subsección 2.5.2.	50
4.3	Tamaño estimado de la prueba de Kriging Ordinario. Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$, con $\beta_0 = 3, \beta_1 = 0, \beta_2 = 0, \sigma^2 = 4, \phi = 0.01$ y $\tau^2 = 1$. $H_0 : \beta_1 = \beta_2 = 0, \beta_0 > 0$. Los distintos modelos de semivariograma están definidos en la subsección 2.5.2.	52
4.4	Potencia estimada de la prueba de Kriging Ordinario. Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$, con $\beta_0 = 3, \beta_1 = 3, \beta_2 = 3, \sigma^2 = 4, \phi = 0.01$ y $\tau^2 = 1$. $H_1 : \beta_p \neq 0$ para algún $p = 1, 2$. Los distintos modelos de semivariograma están definidos en la subsección 2.5.2.	54

4.5 Potencia estimada de la prueba de Kriging Ordinario. Estudio de simulación en una malla irregular de $(0,50) \times (0,50)$, con $\beta_0 = 3$, $\beta_1 = 0.25$, $\beta_2 = 0.25$, $\sigma^2 = 4$, $\phi = 0.01$ y $\tau^2 = 1$. $H_1 : \beta_p \neq 0$ para algún $p = 1,2$. Los distintos modelos de semivariograma están definidos en la subsección 2.5.2. 56

LISTA DE FIGURAS

3.1	Mapa del estado de Paraná con 143 estaciones.	41
4.1	Semivariograma empírico, de tipo Cressie-Hawkins, ajustado a los datos de precipitación del estado de Paraná, Brasil. El semivariograma de la izquierda fue ajustado usando una tendencia constante y el de la derecha, usando una tendencia de primer orden.	58
4.2	Semivariograma teórico (exponencial), ajustado al semivariograma empírico. El semivariograma de la izquierda fue ajustado usando una tendencia constante y el de la derecha, usando una tendencia de primer orden.	58

Capítulo 1

Introducción

El término estadística espacial se utiliza para describir una amplia gama de modelos estadísticos y métodos destinados al análisis de datos espacialmente referenciados (Diggle y Ribeiro, 2007). Estos métodos han tenido un rápido aumento en popularidad debido a la demanda en una amplia gama de campos de la ciencia. Estos incluyen, entre otros, la biología, la economía espacial, el procesamiento de imágenes, las ciencias ambientales y de la tierra, la ecología, la geografía, la epidemiología, la agronomía, el área forestal y la prospección de minerales (Gaetan y Guyon, 2010).

Dado que los datos espaciales surgen en varios campos y aplicaciones, existen varios tipos de datos espaciales, estructuras y escenarios. Una clasificación, invariablemente, sería o bien un tanto amplia o excesivamente detallada (Schabenberger y Gotway, 2005). Cressie (1993) clasifica los datos espaciales, por la naturaleza de su dominio espacial, en tres tipos: datos geoestadísticos, datos en látilce y datos en patrones de punto. En este trabajo el interés está orientado hacia los datos geoestadísticos que se definen a continuación.

Se denota un proceso espacial (o campo aleatorio) en d dimensiones como $\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$, donde $\mathbf{Z}(\mathbf{s})$ es un vector aleatorio que denota los atributos que se observan. La ubicación en la que se observa \mathbf{Z} es \mathbf{s} , un vector de coordenadas de dimensión d . D es un conjunto fijo de $\mathbf{s} \in \mathbb{R}^d$ que contiene un rectángulo d -dimensional de volumen positivo. Debido a la continuidad de D , los datos geoestadísticos también son conocidos como datos espaciales con variación continua. En este documento se va a trabajar con procesos espaciales en un espacio bidimensional, $d = 2$, y coordenadas cartesianas $\mathbf{s} = [x, y]'$.

Del mismo modo que la distribución gaussiana univariada es la distribución usada para muchos

1. Introducción

métodos estadísticos clásicos, el análisis espacial se basa en campos aleatorios gaussianos. El mejor predictor lineal insesgado para el atributo $Z(s_0)$ en una localización observada s_0 , en general, es mejor en esta clase restringida de predictores. En un campo aleatorio gaussiano, no estacionariedad de segundo orden implica estacionariedad estricta. En un campo aleatorio no gaussiano, esta implicación no se mantiene.

En estadística espacial se debe de hacer la distinción entre el tipo de datos (conectado a las características del dominio D) y las propiedades distribucionales del atributo que se está estudiando. El hecho de que el dominio D sea continuo, es decir, los datos son geoestadísticos, no tiene relación con la naturaleza del atributo con respecto a que si son discretos o continuos. Se puede observar la presencia o ausencia de una enfermedad en un dominio espacialmente continuo. El hecho de que D sea discreto, no impide al atributo $\mathbf{Z}(\mathbf{s})$ en \mathbf{s} de seguir la ley de Gauss. Tampoco debe interpretarse la continuidad en D como condición para tener un campo aleatorio gaussiano (Schabenberger y Gotway, 2005).

Dada la importancia que tiene la distribución gaussiana en el análisis de datos espaciales, se hace necesario desarrollar metodología estadística que permita determinar, mediante pruebas de hipótesis, si un conjunto de datos espaciales tienen distribución Gaussiana Multivariada; es decir, determinar si los datos son una realización de un Campo Aleatorio Gaussiano.

En la literatura estadística existen varios métodos estadísticos que se usan para probar si un conjunto de datos tiene distribución gaussiana multivariada. Algunas referencias son Villaseñor y Gonzalez (2009), Székely y Rizzo (2005), Hwu *et al.* (2002), Mecklin y Mundfrom (2005), Srivastava y Mudholkar (2003), Thode (2002), Henze (2002), Royston (1982), Royston (1983) y las pruebas propuestas por Mardia (1970); sin embargo, no existen pruebas de hipótesis para la contraparte espacial. Uno de los objetivos de este trabajo es proponer un método para determinar si un conjunto de datos espaciales constituyen una realización de un Campo Aleatorio Gaussiano. Mediante la técnica de pruebas de hipótesis se prueba la hipótesis nula de un Campo Aleatorio Gaussiano. En la hipótesis alternativa se consideran varios tipos de Campos Aleatorios no Gaussianos. Se estudian las propiedades de tamaño y potencia de la prueba propuesta mediante un estudio de Monte Carlo. Los resultados muestran que el tamaño estimado se aproxima al tamaño nominal con un tamaño de muestra igual o mayor a 200, y la potencia es alta a partir de tamaños de muestra iguales o mayores a 100.

Suponiendo que los datos espaciales tienen distribución gaussiana multivariada, se puede utilizar la teoría basada en la distribución gaussiana multivariada. En el CAG se supone que:

$$\mathbf{Z}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{e}(\mathbf{s}), \mathbf{e}(\mathbf{s}) \sim N^n(\mathbf{0}, \Sigma(\boldsymbol{\theta})) \quad (1.1)$$

1. Introducción

En donde $\mathbf{X}(\mathbf{s})\boldsymbol{\beta}$ es conocida como la tendencia de escala grande, que es la media del proceso, y $\mathbf{e}(\mathbf{s})$ es un término aleatorio que representa la variación de $\mathbf{Z}(\mathbf{s})$, que puede ser modelada a través de un modelo de semivariograma con vector de parámetros $\boldsymbol{\theta}$.

Basandose en el modelo (1.1), el investigador puede enfocarse en validar modelos para el semivariograma, o se puede dedicar a determinar modelos para la media del proceso. En [Guzmán et al. \(2015\)](#) se puede encontrar atención a algunas pruebas de hipótesis para determinar un modelo paramétrico de semivariograma para $\mathbf{e}(\mathbf{s})$ ([Barry, 1996](#), [Clark y Allingham, 2011](#), [Oliver y Webster, 2014](#)).

Por otro lado, al enfocarse a estudiar y comprender la media, los parámetros del semivariograma son considerados como parámetros de ruido. En este caso, la media $\mathbf{X}(\mathbf{s})\boldsymbol{\beta}$ puede tomar diversas formas que van de ser constante a ser un polinomio finito de grado k . Para el caso de media constante, el modelo resultante es conocido como Kriging Ordinario; el caso de un polinomio con $k \geq 1$ es conocido como Kriging Universal.

La palabra Kriging se refiere a un método de predicción espacial basado en la minimización del error cuadrático medio espacial que por lo general depende de las propiedades de segundo orden del proceso $\mathbf{Z}(\mathbf{s})$ ([Cressie, 1993](#)). Es claro que para poder predecir, primero se debe de estimar los parámetros del modelo; i.e., estimar $\boldsymbol{\theta}$ y $\boldsymbol{\beta}$. Dependiendo del modelo de semivariograma considerado y del polinomio para $\mathbf{X}(\mathbf{s})\boldsymbol{\beta}$, se tendrá un número determinado de parámetros a estimar. Por ejemplo, para un modelo de semivariograma Matern con pepita y una media constante se tienen que estimar cinco parámetros. El modelo se vuelve más complejo conforme el número de parámetros a estimar crece. El trabajo computacional también se complica conforme el número de parámetros aumenta.

Antes de que el investigador se dedique a aplicar algún tipo de Kriging a sus datos, sería conveniente que tuviese a su alcance alguna metodología que le ayude a decidir si la media es constante; esto le ahorraría tiempo en su búsqueda entre Kriging Ordinario y Universal. Este es uno de los objetivos de este trabajo. Mediante la técnica de pruebas de hipótesis se prueba la hipótesis nula de Kriging Ordinario. En la hipótesis alternativa se considera un modelo de Kriging Universal donde $\mathbf{X}(\mathbf{s})\boldsymbol{\beta}$ es un modelo de superficie de respuesta de primer grado en las coordenadas de \mathbf{s} . Se estudian las propiedades de tamaño y potencia de la prueba mediante un estudio de Monte Carlo. Los resultados muestran que el tamaño estimado se aproxima al tamaño nominal con un tamaño de muestra igual o mayor a 100, y la potencia es alta a partir de muestras iguales o mayores a 100.

Este trabajo esta organizado de la siguiente forma. En el capítulo dos se presenta el marco teórico

1. Introducción

de la Geoestadística. En el capítulo tres se plantea la prueba de hipótesis para un campo aleatorio gaussiano. En el capítulo cuatro se describe la prueba para la media de un campo aleatorio gaussiano. Por último en el capítulo cinco, se presentan algunas conclusiones

Capítulo 2

Marco Teórico

2.1 Procesos Estocásticos espaciales y Campos Aleatorios

Un proceso estocástico es una colección de variables aleatorias indizadas de acuerdo a alguna métrica. Por ejemplo, una serie de tiempo $Y(t)$, $t = t_1, \dots, t_n$ esta indizado por los puntos en el tiempo en los cuales es observada la serie. De manera similar, un proceso espacial es una colección de variables aleatorias que estan indizadas por algun conjunto $D \subset \mathbb{R}^d$ que contiene coordenadas espaciales $\mathbf{s} = [s_1, \dots, s_d]'$. Para un proceso en el plano, $d = 2$, las coordenadas de latitud y longitud se identifican como $\mathbf{s} = [x, y]'$. Si la dimension, d , del conjunto índice es mayor que uno, el proceso estocástico es llamado *Campo Aleatorio*. Para mayor precisión, se denota un proceso espacial (o campo aleatorio) en d dimensiones como,

$$\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$$

En donde, \mathbf{Z} denota los atributos que se observan en el sitio \mathbf{s} , por ejemplo, el rendimiento, la concentración, o el número de muertes súbitas en infantes. El dominio D es un conjunto continuo, fijo. Por continuo nos referimos a que $\mathbf{Z}(\mathbf{s})$ puede ser observado en cualquier punto dentro de D , i.e., entre dos ubicaciones muestrales \mathbf{s}_i y \mathbf{s}_j se puede colocar, teóricamente, un número infinito de otras muestras. Por fijo queremos decir que los puntos en D no son estocásticos. Debido a la continuidad en D , los datos en $\mathbf{Z}(\mathbf{s})$ son llamados *Datos Geoestadísticos* aunque también se conocen como *Datos Espaciales con Variación Continua*. Es importante asociar la continuidad con el dominio, no con el atributo que se mide. Si el atributo \mathbf{Z} es continuo o discreto, no influye en si los datos son geoestadísticos o no (Schabenberger y Gotway, 2005).

2.1. Procesos Estocásticos espaciales y Campos Aleatorios

Definición 2.1 Un campo aleatorio $\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ es estacionario en sentido estricto (o fuerte) si su distribución espacial es invariante bajo la traslación de las coordenadas, i.e.,

$$\begin{aligned} &Pr(\mathbf{Z}(\mathbf{s}_1) < z_1, \mathbf{Z}(\mathbf{s}_2) < z_2, \dots, \mathbf{Z}(\mathbf{s}_k) < z_k) = \\ &Pr(\mathbf{Z}(\mathbf{s}_1 + \mathbf{h}) < z_1, \mathbf{Z}(\mathbf{s}_2 + \mathbf{h}) < z_2, \dots, \mathbf{Z}(\mathbf{s}_k + \mathbf{h}) < z_k) \end{aligned}$$

para toda k y \mathbf{h} .

Definición 2.2 Un campo aleatorio $\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ es estacionario de segundo orden (débil) si su media es constante y la covarianza entre atributos en diferentes ubicaciones es solamente una función de su separación espacial (vector de retraso o lag-vector) \mathbf{h} , esto es,

$$\begin{aligned} E[\mathbf{Z}(\mathbf{s})] &= \boldsymbol{\mu} \\ Cov[\mathbf{Z}(\mathbf{s}), \mathbf{Z}(\mathbf{s} + \mathbf{h})] &= C(\mathbf{h}) \end{aligned}$$

Donde la función $C(\mathbf{h})$ es llamada función de covarianza del proceso espacial. Además si $C(\mathbf{h})$ es una función solamente de $\|\mathbf{h}\|$, entonces se dice que $C(\cdot)$ es isotrópica, donde $\|\mathbf{h}\|$ es la norma euclídeana del vector de retraso.

La propiedad de estacionariedad refleja la falta de importancia de coordenadas absolutas. La función $C(\mathbf{h})$ juega un papel importante en el modelado estadístico de datos espaciales. Por ejemplo, la covarianza de las observaciones espaciadas con dos días de diferencia en una serie de tiempo estacionaria de segundo orden será la misma, sin importar si el primer día es un lunes o un viernes. Estacionariedad estricta implica estacionariedad de segundo orden pero el inverso no es cierto por la misma razón por la que no podemos inferir la distribución de una variable aleatoria a partir de conocer solamente su media y su varianza.

La existencia de la función de covarianza $C(\mathbf{h})$ en un campo aleatorio estacionario de segundo orden tiene consecuencias importantes. Dado que $C(\mathbf{h})$ no depende de coordenadas absolutas y $Cov[\mathbf{Z}(\mathbf{s}), \mathbf{Z}(\mathbf{s} + \mathbf{0})] = Var[\mathbf{Z}(\mathbf{s})] = C(\mathbf{0})$, se deduce que la variabilidad de un campo aleatorio estacionario de segundo orden es la misma en todas partes. Un proceso espacial estacionario de segundo orden tiene media constante, varianza constante, y una función de covarianza que no depende de coordenadas absolutas. Tal proceso es el equivalente espacial de una muestra aleatoria en estadística clásica en la que las observaciones tienen la misma media y la misma dispersión (pero sin estar correlacionados).

La función de covarianza $C(\mathbf{h})$ de un campo aleatorio estacionario de segundo orden tiene las siguientes propiedades:

2.1. Procesos Estocásticos espaciales y Campos Aleatorios

- 1) $C(\mathbf{0}) \geq 0$;
- 2) $C(\mathbf{h}) = C(-\mathbf{h})$, i.e. C es una función par;
- 3) $C(\mathbf{0}) \geq |C(-\mathbf{h})|$;
- 4) $C(\mathbf{h}) = Cov[\mathbf{Z}(\mathbf{s}), \mathbf{Z}(\mathbf{s} + \mathbf{h})] = Cov[\mathbf{Z}(\mathbf{0}), \mathbf{Z}(\mathbf{h})]$;
- 5) Si $C_j(\mathbf{h})$ son funciones de covarianza válidas, $j = 1, \dots, k$, entonces $\sum_{j=1}^k b_j C_j(\mathbf{h})$ es una función de covarianza válida, si $b_j \geq 0$ para toda j ;
- 6) Si $C_j(\mathbf{h})$ son funciones de covarianza válidos, $j = 1, \dots, k$, entonces $\prod_{j=1}^k C_j(\mathbf{h})$ es una función de covarianza válida.
- 7) Si $C(\mathbf{h})$ es una función de covarianza válida en \mathbb{R}^d , entonces también es una función de covarianza válida en \mathbb{R}^p , $p < d$.

Algunas demostraciones y comentarios acerca de las propiedades de $C(\mathbf{h})$ se pueden encontrar en [Schabenberger y Gotway \(2005\)](#) y [Gaetan y Guyon \(2010\)](#).

Definición 2.3 Un campo aleatorio $\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ es intrínsecamente estacionario si,

$$\begin{aligned} E[\mathbf{Z}(\mathbf{s})] &= \boldsymbol{\mu} \\ \frac{1}{2}Var[\mathbf{Z}(\mathbf{s}) - \mathbf{Z}(\mathbf{s} + \mathbf{h})] &= \boldsymbol{\gamma}(\mathbf{h}) \end{aligned}$$

Donde la función $\boldsymbol{\gamma}(\mathbf{h})$ recibe el nombre de semivariograma del proceso espacial. Además, si $\boldsymbol{\gamma}(\mathbf{h}) = \boldsymbol{\gamma}^*(\|\mathbf{h}\|)$ es una función solamente de $\|\mathbf{h}\|$, entonces se dice que $\boldsymbol{\gamma}(\cdot)$ es isotrópico.

Se puede demostrar que la clase de procesos intrínsecamente estacionarios es más grande que la clase de procesos estacionarios de segundo orden ([Cressie, 1993, Ch. 2.5.2](#)). Para ver que un proceso estacionario de segundo orden es también intrínsecamente estacionario es suficiente con examinar la siguiente expresión,

$$\begin{aligned} Var[\mathbf{Z}(\mathbf{s}) - \mathbf{Z}(\mathbf{s} + \mathbf{h})] &= Var[\mathbf{Z}(\mathbf{s})] + Var[\mathbf{Z}(\mathbf{s} + \mathbf{h})] - 2Cov[\mathbf{Z}(\mathbf{s}), \mathbf{Z}(\mathbf{s} + \mathbf{h})] \\ &= 2\{Var[\mathbf{Z}(\mathbf{s})] - 2C(\mathbf{h})\} \\ &= 2\{C(\mathbf{0}) - C(\mathbf{h})\} = 2\boldsymbol{\gamma}(\mathbf{h}) \end{aligned}$$

Por 2) y 3), note que $\boldsymbol{\gamma}(\mathbf{h}) \geq 0$. Estacionariedad intrínseca no implica estacionariedad de segundo orden. Si el proceso es intrínsecamente estacionario pero no estacionario de segundo orden, se tiene que $C(\mathbf{h})$ es un parámetro que no existe, entonces no queda más que trabajar con el semivariograma $\boldsymbol{\gamma}(\mathbf{h})$. Entonces es preferible trabajar con $\boldsymbol{\gamma}(\mathbf{h})$.

2.2 Campo Aleatorio Gaussiano

Un campo aleatorio $\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ es un campo aleatorio gaussiano si la función de distribución acumulada,

$$Pr(Z(\mathbf{s}_1) < z_1, Z(\mathbf{s}_2) < z_2, \dots, Z(\mathbf{s}_k) < z_k)$$

Es la de una variable aleatoria gaussiana k variada para toda k .

Por las propiedades de la distribución gaussiana multivariada esto implica que cada $Z(\mathbf{s}_i)$ es una variable aleatoria gaussiana univariada. El inverso no es cierto. Inclusive si $Z(\mathbf{s}_i) \sim N(\mu(\mathbf{s}_i), \sigma^2(\mathbf{s}_i))$, esto no implica que la distribución conjunta de $Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)$ sea gaussiana multivariada.

2.3 Una representación en términos de un modelo

Un modelo estadístico es la representación matemática de un mecanismo de generación de datos. Es una abstracción de los procesos físicos, biológicos, químicos, etc., que generan los datos; haciendo hincapié en aquellos aspectos del proceso que son importantes para el análisis, y haciendo caso omiso de los aspectos intrascendentes (Schabenberger y Gotway, 2005).

Los modelos estadísticos más genéricos son una descomposición de una variable de respuesta en una estructura matemática que describe la media y una estructura estocástica aditiva que describe la variación y covariación entre las respuestas. Esta descomposición se expresa a menudo simbólicamente como:

$$Datos = Estructura + Error$$

Una versión en la que se supone una tendencia en $Z(\mathbf{s})$ está expresada como,

$$Z(\mathbf{s}) = \mathbf{f}(\mathbf{X}, \mathbf{s}, \boldsymbol{\beta}) + \mathbf{e}(\mathbf{s}) \quad (2.1)$$

Donde $Z(\mathbf{s}) = [Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)]'$, \mathbf{X} es una matriz $n \times p$ de covariables, $\boldsymbol{\beta}$ es un vector de parámetros y $\mathbf{e}(\mathbf{s})$ es un vector aleatorio con media $\mathbf{0}$ y varianza $Var[\mathbf{e}(\mathbf{s})] = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. La función f puede ser no lineal, por lo tanto, es necesario definir lo que representa el vector \mathbf{f} . Los elementos de \mathbf{f} se definen a continuación:

2.3. Una representación en términos de un modelo

$$\mathbf{f}(\mathbf{X}, \mathbf{s}, \boldsymbol{\beta}) = \begin{bmatrix} f(\mathbf{x}_1, \mathbf{s}_1, \boldsymbol{\beta}) \\ f(\mathbf{x}_2, \mathbf{s}_2, \boldsymbol{\beta}) \\ \vdots \\ f(\mathbf{x}_n, \mathbf{s}_n, \boldsymbol{\beta}) \end{bmatrix}$$

De (2.1) se puede ver que $E[\mathbf{Z}(\mathbf{s})] = \mathbf{f}(\mathbf{X}, \mathbf{s}, \boldsymbol{\beta})$ representa la tendencia a gran escala, la estructura de la media del proceso espacial. La variación y covariación de $\mathbf{Z}(\mathbf{s})$ se representa a través de las propiedades estocásticas de $\mathbf{e}(\mathbf{s})$. La suposición de estacionariedad es en términos del error $\mathbf{e}(\mathbf{s})$ del modelo, no en términos del atributo $\mathbf{Z}(\mathbf{s})$. La suposición de media cero de los errores en el modelo es un reflejo de la creencia de que el modelo es correcto en la media. Las propiedades de estacionariedad del campo aleatorio se reflejan en la estructura de $Var[\mathbf{e}(\mathbf{s})] = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. Las entradas de esta matriz de covarianza se pueden construir a partir de la función de covarianza $C(\mathbf{h})$ de un proceso estacionario de segundo orden, o equivalentemente en función de $\gamma(\mathbf{h})$.

Se puede hacer algunas simplificaciones y modificaciones a la estructura básica (2.1). La estructura a gran escala con frecuencia se puede expresar como una función lineal de las coordenadas espaciales, $E[\mathbf{Z}(\mathbf{s})] = \mathbf{X}(\mathbf{s})\boldsymbol{\beta}$. La matriz diseño (o regresora) \mathbf{X} es entonces un modelo de superficie de respuesta u otro polinomio en las coordenadas de \mathbf{s} , de ahí la dependencia de \mathbf{X} en \mathbf{s} . La matriz $\mathbf{X}(\mathbf{s})$ puede contener otras variables además de la información acerca de \mathbf{s} como es el caso de los modelos de regresión espacial. La función f es con frecuencia una función monótona e invertible. En este caso, se puede modelar,

$$\mathbf{Z}(\mathbf{s}) = f(\mathbf{x}'(\mathbf{s})\boldsymbol{\beta}) + \mathbf{e}(\mathbf{s}) \quad (2.2)$$

$$f^{-1}(E[\mathbf{Z}(\mathbf{s})]) = \mathbf{x}'(\mathbf{s})\boldsymbol{\beta}$$

La formulación como modelo de un proceso espacial es útil. Sin embargo, no manifiesta cual componente de parámetros es más importante para el modelador. Rara vez es igual la importancia dada al vector de parámetros de la media $\boldsymbol{\beta}$ y el vector de parámetros de covarianza $\boldsymbol{\theta}$. En regresión (o análisis de varianza espacial), se pone más énfasis en inferencias acerca de la función de la media y a $\boldsymbol{\theta}$ frecuentemente se le considera un parámetro de ruido. En aplicaciones de predicción espacial, la estructura de covarianza y los valores de los parámetros de $\boldsymbol{\theta}$ son cruciales, ya que el error cuadrado medio de predicción depende de ellos (Schabenberger y Gotway, 2005).

2.4 El Semivariograma

Sea $\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ un proceso espacial y sea,

$$\begin{aligned}\gamma^*(\mathbf{s}_i, \mathbf{s}_j) &= \frac{1}{2} \text{Var} [\mathbf{Z}(\mathbf{s}_i) - \mathbf{Z}(\mathbf{s}_j)] \\ &= \frac{1}{2} \{ \text{Var} [\mathbf{Z}(\mathbf{s}_i)] + \text{Var} [\mathbf{Z}(\mathbf{s}_j)] - 2 \text{Cov} [\mathbf{Z}(\mathbf{s}_i), \mathbf{Z}(\mathbf{s}_j)] \}\end{aligned}\quad (2.3)$$

Si $\gamma^*(\mathbf{s}_i, \mathbf{s}_j) \equiv \gamma(\mathbf{s}_i - \mathbf{s}_j)$ es función solamente de la diferencia de coordenadas $\mathbf{s}_i - \mathbf{s}_j$, entonces llamamos a $\gamma(\mathbf{s}_i - \mathbf{s}_j)$ el semivariograma del proceso espacial de covarianza estacionaria. Si $\mathbf{Z}(\mathbf{s})$ es intrínsecamente estacionario, entonces $\gamma(\mathbf{s}_i - \mathbf{s}_j)$ es un parámetro del proceso estocástico. En ausencia de estacionariedad, γ^* sigue siendo una función válida con la cual la matriz de varianzas y covarianzas $\text{Var}[\mathbf{e}(\mathbf{s})] = \mathbf{\Sigma}$ puede ser construida, pero no debe ser referida como semivariograma. La función $2\gamma(\mathbf{s}_i - \mathbf{s}_j)$ se conoce como el variograma, aunque la literatura no es consistente en este respecto. En este trabajo nos referiremos a γ como el semivariograma y a 2γ como el variograma.

Existen diversos modelos teóricos de semivariograma. En [Samper y Carrera \(1990\)](#) se presenta una discusión respecto a las características y condiciones que éstos deben de cumplir. La mayoría de los modelos dependen de tres parámetros comunes que son descritos a continuación:

Efecto pepita

Se denota por τ y representa una discontinuidad puntual del semivariograma en el origen. Puede ser debido a errores de medición en la variable o a la escala de la misma. En algunas ocasiones puede ser indicativo de que parte de la estructura espacial se concentra a distancias inferiores a las observadas.

Meseta

Es la cota superior del semivariograma. Puede definirse como el límite del semivariograma cuando la distancia h tiende a infinito. La meseta puede ser o no finita. Los semivariogramas que tienen meseta finita cumplen con la hipótesis de estacionariedad fuerte; mientras que cuando ocurre lo contrario, el semivariograma define un fenómeno natural que cumple sólo con la hipótesis intrínseca. La meseta se denota por $(\sigma^2 + \tau^2)$ cuando la pepita es diferente de cero. La pepita

2.5. Modelos de Covarianza estacionaria y el Semivariograma

no debe representar más del 50 por ciento de la meseta. Si el ruido espacial en las mediciones explica en mayor proporción la variabilidad que la correlación del fenómeno, las predicciones que se obtengan pueden ser muy imprecisas.

Rango

Se denota por ϕ . En términos prácticos corresponde a la distancia a partir de la cual dos observaciones son independientes. El rango se interpreta como la zona de influencia en torno a un punto más allá de la cual la autocorrelación es nula. Existen algunos modelos de semivariograma en los que no existe una distancia finita para la cual dos observaciones sean independientes; por ello se llama rango práctico a la distancia para la cual el semivariograma alcanza el 95% de la meseta. Entre más pequeño sea el rango, más cerca se está del modelo de independencia espacial. El rango no siempre aparece de manera explícita en la fórmula del semivariograma.

2.5 Modelos de Covarianza estacionaria y el Semivariograma

2.5.1 Validez del modelo

Se consideran modelos isotrópicos para la función de covarianza y el semivariograma de un proceso espacial. Se parte de modelos para las funciones de covarianza debido a que los semivariogramas para procesos estacionarios de segundo orden pueden ser construidos a partir de funciones de covarianza. Por ejemplo, si $C(\mathbf{h})$ es la función de covarianza de un proceso isotrópico con varianza σ^2 y sin efecto pepita, entonces,

$$\gamma(\mathbf{h}) = \begin{cases} 0 & h = 0 \\ \sigma^2(1 - C(\mathbf{h})) & h > 0 \end{cases}$$

No cualquier expresión matemática puede servir como un modelo para la dependencia espacial en un campo aleatorio. Sea $C(\mathbf{h})$ la función de covarianza isotrópica de un campo estacionario de segundo orden y $\gamma(\mathbf{h})$ el semivariograma isotrópico de un campo estacionario de segundo orden o intrínsecamente estacionario. Entonces, se cumplen las siguientes propiedades:

- Si $C(\mathbf{h})$ es válido en \mathbb{R}^d , entonces también es válido en \mathbb{R}^s , $s < d$ (Matérn, 1986). Si $\gamma(\mathbf{h})$ es válido en \mathbb{R}^d , también es válido en \mathbb{R}^s , $s < d$.

2.5. Modelos de Covarianza estacionaria y el Semivariograma

- Si $C_1(\mathbf{h})$ y $C_2(\mathbf{h})$ son funciones de covarianza válidas, entonces $aC_1(\mathbf{h}) + bC_2(\mathbf{h})$, $a, b \geq 0$ es una función de covarianza válida.
- 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.
- Una función de covarianza $C(\mathbf{h})$ válida es una función definida positiva, esto es, se tiene que,

$$\sum_{i=1}^k \sum_{j=1}^k a_i a_j C(\mathbf{s}_i - \mathbf{s}_j) \geq 0,$$

Para cualquier conjunto de números reales a_1, \dots, a_k y ubicaciones $\mathbf{s}_i, \mathbf{s}_j$.

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

$$2 \sum_{i=1}^m \sum_{j=1}^m a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0,$$

Para cualesquiera números reales a_1, \dots, a_m , de tal manera que $\sum_{i=1}^m a_i = 0$ y un número finito de sitios $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_m$.

- Una condición necesaria para que $\gamma(\mathbf{h})$ sea un semivariograma válido es que $2\gamma(\mathbf{h})$ crezca más lentamente que $\|\mathbf{h}\|^2$. Esto con frecuencia se refiere a la hipótesis intrínseca.

2.5.2 Algunas funciones de covarianza

Algunos modelos paramétricos comunmente usados en la práctica son los siguientes:

Covarianza Matérn

$$C(\mathbf{h}) = \sigma^2 \left\{ \frac{1}{2^{k-1} \Gamma(k)} \left(\frac{\mathbf{h}}{\phi} \right)^k K_\kappa \left(\frac{\mathbf{h}}{\phi} \right) \right\}, \quad \mathbf{h} > 0, \kappa > 0, \phi > 0 \quad (2.4)$$

Donde K_κ es la función Bessel de segundo tipo modificada de orden $k > 0$. El parámetro ϕ gobierna el grado de dependencia espacial. El suavizamiento del proceso aumenta con κ .

Es de particular importancia, en la modelación espacial, la función Bessel de segundo tipo, K_ν , de orden ν . Esta función está definida como:

2.5. Modelos de Covarianza estacionaria y el Semivariograma

$$K_\nu(t) = \frac{\pi I_{-\nu}(t) - I_\nu(t)}{2 \cdot 2\pi \sin \nu} \quad (2.5)$$

Donde $I_\nu(t)$ es la función Bessel modificada de primer tipo, definida por:

$$I_\nu(t) = \left(\frac{t}{2}\right)^\nu \sum_{i=0}^{\infty} \frac{\left(\frac{1}{4}t^2\right)^i}{i! \Gamma(\nu+i+1)} \left(\frac{t}{2}\right)^{2k}$$

Dado que el cálculo de estas funciones puede ser numéricamente complicado, se pueden utilizar aproximaciones para $t \rightarrow 0$:

$$\begin{aligned} K_0(t) &\approx -\ln(t) \\ K_\nu(t) &\approx \frac{\Gamma(\nu)}{2} \left(\frac{t}{2}\right)^{-\nu}, \nu > 0 \end{aligned}$$

En la sección 3.2 se presentan detalles acerca de la estimación de los parámetros de esta función de covarianza.

Covarianza Esférica

La expresión de este modelo está dada por:

$$C(\mathbf{h}) = \sigma^2 \left[1 - \frac{3}{2} \frac{\mathbf{h}}{\phi} + \frac{1}{2} \left(\frac{\mathbf{h}}{\phi} \right)^3 \right], \mathbf{h} \leq \phi$$

Este modelo de covarianza tiene una expresión polinómica simple y su forma coincide con lo que a menudo se observa: un crecimiento casi lineal hasta una cierta distancia y luego una estabilización.

El parámetro ϕ indica el rango de la covarianza esférica; es decir, la covarianza desaparece cuando el rango es alcanzado. Esto significa que el modelo esférico es de rango finito. De hecho, la familia de covarianzas esférica carece de flexibilidad en comparación con la clase Matérn de dos parámetros y la función de covarianza exponencial. También la función de correlación, $\rho(\mathbf{h}) = C(\mathbf{h})/\sigma^2$, es solo una vez diferenciable en $h = \phi$, lo cual causa dificultades técnicas en la estimación de máxima verosimilitud (Mardia y Watkins, 1989, Warnes y Ripley, 1987).

El semivariograma correspondiente a este modelo de covarianza, debe su popularidad en gran parte al hecho de que ofrece ilustraciones claras de la pepita, la cima o meseta y el rango, tres

2.5. Modelos de Covarianza estacionaria y el Semivariograma

características tradicionalmente asociadas con variogramas. Sin embargo, el semivariograma, es válido en las dimensiones $D = 1, 2, 3$, pero para $D \geq 4$ falla para corresponder a una matriz de covarianzas que sea definida positiva (Banerjee *et al.*, 2004).

Covarianza Exponencial

Un caso importante de la clase Matérn de funciones de covarianza se obtiene para $\kappa = \frac{1}{2}$. El modelo resultante se conoce como el modelo exponencial,

$$C(\mathbf{h}) = \sigma^2 \exp\left(-\frac{\mathbf{h}}{\phi}\right), \mathbf{h} > 0, \phi > 0$$

Aunque el modelo esférico es suave en el sentido de diferenciación continua, hace la suposición implícita de que las correlaciones son exactamente cero a distancias suficientemente grandes. Pero en algunos casos puede ser más apropiado asumir que aunque las correlaciones pueden llegar a ser arbitrariamente pequeñas a grandes distancias, estas nunca desaparecen.

El modelo exponencial tiene una ventaja sobre el modelo esférico en cuanto a que aunque su forma funcional es más simple, tiene un semivariograma válido en todas las dimensiones (y sin el requerimiento del rango finito del modelo esférico). La función de covarianza decae exponencialmente con el incremento de la distancia. El parámetro ϕ determina que tan rápido decae la covarianza. Para un valor de $\mathbf{h} = 3\phi$ la función de covarianza decrece aproximadamente un 95 por ciento su valor partiendo del origen, de modo que esta distancia se ha denominado el **rango práctico** del modelo exponencial.

Covarianza Gaussiana

Se puede determinar el rango práctico para valores particulares de κ en (2.4). Cuando $\kappa \rightarrow \infty$ el modelo de covarianza límite se conoce como el modelo gaussiano,

$$C(\mathbf{h}) = \sigma^2 \exp\left\{-\left(\frac{\mathbf{h}}{\phi}\right)^2\right\}, \mathbf{h} > 0, \phi > 0$$

Covarianza Circular

Sea $\theta = \min\left(\frac{\mathbf{h}}{\phi}, 1\right)$ y $g(\mathbf{h}) = \frac{2}{\pi} \left(\theta \sqrt{1 - \theta^2} + \sin^{-1} \sqrt{\theta}\right)$. El modelo circular está dado por,

$$C(\mathbf{h}) = \sigma^2 [1 - g(\mathbf{h})], \mathbf{h} < \phi$$

Covarianza Cúbica

$$C(\mathbf{h}) = \sigma^2 \left\{ 1 - \left[7 \left(\frac{\mathbf{h}}{\phi} \right)^2 - 8.75 \left(\frac{\mathbf{h}}{\phi} \right)^3 + 3.5 \left(\frac{\mathbf{h}}{\phi} \right)^5 - 0.75 \left(\frac{\mathbf{h}}{\phi} \right)^7 \right] \right\}, \text{ si } \mathbf{h} < \phi$$

2.6 Estimacion de Máxima Verosimilitud en CAG

Con el fin de proceder a hacer estimación por Máxima Verosimilitud (ML, por sus siglas en inglés), se tiene la necesidad de hacer una suposición acerca de la distribución de $\mathbf{Z}(\mathbf{s})$. No es suficiente con especificar los dos primeros momentos. La estimación de MV para modelos espaciales ha sido desarrollada solamente para el caso Gaussiano (Mardia y Marshall, 1984).

Consideramos el modelo gaussiano con una especificación lineal para la tendencia espacial. Esto permite la inclusión de una superficie de tendencia polinómica o, en general, covariables espacialmente referenciados. Por lo tanto, sea,

$$\mathbf{Z}(\mathbf{s}) \sim N^n (\mathbf{C}\boldsymbol{\beta}, \sigma^2 \mathbf{R}(\phi) + \tau^2 \mathbf{I}) \quad (2.6)$$

Donde \mathbf{C} es una matriz ($n \times p$) de covariables, $\boldsymbol{\beta}$ es el correspondiente vector de parámetros de regresión, $\tau^2 + \sigma^2$ es la varianza del campo, ϕ es un parámetro de escala, \mathbf{I} es la matriz identidad de dimensión $n \times n$ y $\mathbf{R}(\phi)$ es la matriz de correlaciones de dimensión $n \times n$ cuyos elementos están dados por,

$$(\mathbf{R}(\phi))_{ij} = \rho_Z(h_{ij}, \phi) = \rho(\mathbf{h}, \phi) = \frac{C(\mathbf{h})}{\sigma^2} \quad (2.7)$$

Es decir, $(\mathbf{R}(\phi))_{ij}$ es la correlación que existe entre $Z(s_i)$ y $Z(s_j)$, $s_i, s_j \in D$, $h_{ij} = \|s_i - s_j\|$.

La función de verosimilitud está dada por,

$$L(\boldsymbol{\beta}, \sigma^2, \phi, \tau^2) = -\frac{1}{2} \left\{ n \log(2\pi) + \log [|\sigma^2 \mathbf{R}(\phi) + \tau^2 \mathbf{I}|] + (\mathbf{Z}(\mathbf{s}) - \mathbf{C}\boldsymbol{\beta})' (\sigma^2 \mathbf{R}(\phi) + \tau^2 \mathbf{I})^{-1} (\mathbf{Z}(\mathbf{s}) - \mathbf{C}\boldsymbol{\beta}) \right\} \quad (2.8)$$

Un algoritmo para maximizar (2.8) es el siguiente.

2.7. Estimacion de Máxima Verosimilitud Restringida en CAG

Sean $v^2 = \frac{\tau^2}{\sigma^2}$ y $\mathbf{V} = \mathbf{R}(\phi) + v^2\mathbf{I}$. Dado \mathbf{V} , la función de log-verosimilitud se maximiza en,

$$\hat{\boldsymbol{\beta}}(\mathbf{V}) = (\mathbf{C}'\mathbf{V}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{V}^{-1}\mathbf{z} \quad (2.9)$$

y

$$\hat{\sigma}^2(\mathbf{V}) = n^{-1} \left\{ \mathbf{z} - \mathbf{C}\hat{\boldsymbol{\beta}}(\mathbf{V}) \right\}' \mathbf{V}^{-1} \left\{ \mathbf{z} - \mathbf{C}\hat{\boldsymbol{\beta}}(\mathbf{V}) \right\} \quad (2.10)$$

Sustituyendo a $\hat{\boldsymbol{\beta}}(\mathbf{V})$ y $\hat{\sigma}^2(\mathbf{V})$ en la log-verosimilitud se obtiene una log-verosimilitud concentrada,

$$L_0(v^2, \phi) = -\frac{1}{2} [n \log 2\pi + n \log \hat{\sigma}^2(\mathbf{V}) + \log |\mathbf{V}| + n] \quad (2.11)$$

Se debe de optimizar (2.11) con respecto a ϕ y v seguida de sustitución recursiva para obtener $\hat{\sigma}^2$ y $\hat{\boldsymbol{\beta}}$.

Los detalles prácticos de la optimización dependen de la familia de covarianza, en particular, bajo consideración. Por ejemplo, cuando se utiliza la función de covarianza Matérn, se sabe que el parámetro de forma κ presenta problemas de identificación. Por lo tanto, se prefiere elegir el valor de κ de un conjunto discreto, por ejemplo $\{0.5, 1.5, 2.5\}$, para cubrir diferentes grados de diferenciación media cuadrada del proceso en estudio, en vez de intentar optimizar sobre todos los valores positivos de κ .

Muchos autores han notado que los Estimadores de Máxima Verosimilitud (EMV) son seriamente sesgados. Un posible remedio sería incluir un paso de Jackknife en el proceso de estimacion (Miller, 1974), aunque esto incrementaría el tiempo de cálculo computacional hasta en un factor de n .

2.7 Estimacion de Máxima Verosimilitud Restringida en CAG

Una variante del método de Máxima Verosimilitud (ML, por sus siglas en inglés) es el método de Máxima Verosimilitud Restringida (REML, por sus siglas en inglés). Este método de estimación fue introducido por Patterson y Thompson (1971).

Bajo el modelo $E[\mathbf{Z}] = \mathbf{C}\boldsymbol{\beta}$ se pueden transformar los datos linealmente a $\mathbf{Z}^* = \mathbf{A}\mathbf{Z}$ de tal forma que la distribución de \mathbf{Z}^* no depende de $\boldsymbol{\beta}$. Entonces el principio que sigue el método REML es estimar los parámetros $\boldsymbol{\theta} = (v^2, \sigma^2, \phi)$ a través de máxima verosimilitud aplicada a los datos transformados \mathbf{Z}^* , los cuales determinan la estructura de covarianza de los datos. Siempre se

2.8. Predicción espacial y kriging

puede encontrar una matriz \mathbf{A} sin conocer los verdaderos valores de $\boldsymbol{\beta}$ o $\boldsymbol{\theta}$. Por ejemplo, la proyección para residuales de mínimos cuadrados ordinarios,

$$\mathbf{A} = \mathbf{I} - \mathbf{D}(\mathbf{D}\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T$$

tiene la propiedad que se necesita.

Como \mathbf{Z}^* es una transformación lineal de \mathbf{Z} , esta conserva la propiedad distributiva gaussiana multivariada. La restricción impuesta de que la distribución de \mathbf{Z}^* no depende de $\boldsymbol{\beta}$ reduce la dimensión de \mathbf{Z}^* de n a $n - p$, donde p es el número de elementos de $\boldsymbol{\beta}$.

Se calcula el estimador REML de $\boldsymbol{\theta}$ al maximizar el perfil de la verosimilitud de $\boldsymbol{\theta}$ basado en los datos transformados \mathbf{Z}^* . En efecto el perfil-verosimilitud puede ser escrito en términos de los datos originales \mathbf{Z} como,

$$L^*(\boldsymbol{\theta}) = -\frac{1}{2} \left\{ n \log(2\pi) + |\sigma^2 \mathbf{V}| + \log \left| \mathbf{D}^T (\sigma^2 \mathbf{V})^{-1} \mathbf{D} \right| + (y - \mathbf{D}\tilde{\boldsymbol{\beta}})^T (\sigma^2 \mathbf{V})^{-1} (y - \mathbf{D}\tilde{\boldsymbol{\beta}}) \right\} \quad (2.12)$$

Donde $\sigma^2 \mathbf{V}$ es la matriz de varianzas de \mathbf{Z} y $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{V})$ denota el estimador de máxima verosimilitud de $\boldsymbol{\beta}$ dado un valor de $\boldsymbol{\theta}$.

Algunas referencias acerca del método REML en el contexto geoestadístico son [Kitanidis \(1983\)](#) y [Zimmerman \(1989\)](#). En general el método REML produce estimadores menos sesgados que ML para los parámetros de los componentes de varianza en muestras pequeñas. Nótese que $L^*(\boldsymbol{\theta})$ depende de \mathbf{D} , y por lo tanto depende de una correcta especificación de $\mathbf{C}\boldsymbol{\beta}$ en el modelo. Aunque el método REML es ampliamente recomendado para modelos geoestadísticos, la experiencia ha mostrado que es más sensible que ML dependiendo del modelo escogido para $\mathbf{C}\boldsymbol{\beta}$ ([Diggle y Ribeiro, 2007](#)).

2.8 Predicción espacial y kriging

Considere el campo aleatorio $\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ observado en ubicaciones $\mathbf{s}_1, \dots, \mathbf{s}_n$, y el correspondiente vector de datos $\mathbf{Z}(\mathbf{s}) = [Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)]$. El dominio D es fijo y continuo, como cuando se está trabajando con datos geoestadísticos. La muestra es una observación incompleta de la superficie $Z(\mathbf{s}, w)$ que es el resultado de un experimento aleatorio con realización w . Uno de los problemas importantes en estadística espacial es la predicción de Z en alguna ubicación

2.8. Predicción espacial y kriging

especifica $\mathbf{s}_0 \in D$. Este puede ser una ubicación que forma parte del conjunto de ubicaciones en los que se ha observado $Z(\mathbf{s}, w)$, o una nueva ubicación (no observada).

Asuma que el campo aleatorio puede ser representado como,

$$\mathbf{Z}(\mathbf{s}) = \boldsymbol{\mu}(\mathbf{s}) + \mathbf{e}(\mathbf{s}), \quad \mathbf{e}(\mathbf{s}) \sim (\mathbf{0}, \boldsymbol{\Sigma}) \quad (2.13)$$

Podemos estar interesados en la estimación de $E[\mathbf{Z}(\mathbf{s})] = \boldsymbol{\mu}(\mathbf{s})$ o en la predicción de $\mathbf{Z}(\mathbf{s})$. En aplicaciones geoestadísticas la predicción es a menudo más importante que la estimación de la media.

Los métodos geoestadísticos de predicción son herramientas estadísticas para predecir $g(\mathbf{Z}(\mathbf{s}_0))$ a partir de un conjunto de observaciones $Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)$. Por lo general se conocen como métodos de Kriging, un término dado por G. Matheron en honor del ingeniero minero de Sudáfrica D.G. Krige, cuya investigación sobre la estimación de grado mineral en las minas de oro de Witwatersrand es considerado como un trabajo fundamental para el campo de la geoestadística (Krige, 1951, Matheron, 1963).

En este trabajo se presenta una revisión de algunos tipos de kriging existentes en la literatura.

2.8.1 Kriging Ordinario

Considere el modelo (2.13) y supóngase que $E[\mathbf{Z}(\mathbf{s})]$ es desconocida pero constante a través de las ubicaciones. Entonces el modelo resultante es llamado modelo de Kriging Ordinario; es decir,

$$\mathbf{Z}(\mathbf{s}) = \boldsymbol{\mu}\mathbf{1} + \mathbf{e}(\mathbf{s}), \quad \mathbf{e}(\mathbf{s}) \sim (\mathbf{0}, \boldsymbol{\Sigma}) \quad (2.14)$$

Donde $E[\mathbf{Z}(\mathbf{s})] = \boldsymbol{\mu}\mathbf{1}$ y $\text{Var}[\mathbf{Z}(\mathbf{s})] = \boldsymbol{\Sigma}$, con $\boldsymbol{\Sigma}$ desconocida.

2.8.2 Kriging Universal

Supongamos que tenemos datos observados $Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)$ en las ubicaciones espaciales $\mathbf{s}_1, \dots, \mathbf{s}_n$, y queremos predecir $Z(\mathbf{s}_0)$ en la ubicación \mathbf{s}_0 en el cual no tenemos una observación. Además supongamos que la forma del modelo lineal general es válido para los datos observados y los no observados:

2.8. Predicción espacial y kriging

$$\mathbf{Z}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{e}(\mathbf{s}), \quad \mathbf{e}(\mathbf{s}) \sim (\mathbf{0}, \boldsymbol{\Sigma}) \quad (2.15)$$

$$\mathbf{Z}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)'\boldsymbol{\beta} + \mathbf{e}(\mathbf{s}_0)$$

Donde $\mathbf{x}(\mathbf{s}_0)'\boldsymbol{\beta}$ es un vector $p \times 1$ de variables explicativas asociadas con la ubicación s_0 . Suponemos una matriz general de varianza-covarianza de los datos, $Var[\mathbf{Z}(\mathbf{s})] = \boldsymbol{\Sigma}$, y además que los datos y las características no observables están correlacionados espacialmente, de modo que $Cov[\mathbf{Z}(\mathbf{s}), \mathbf{Z}(\mathbf{s}_0)] = \boldsymbol{\sigma}$, un vector de dimensión $n \times 1$, y $Var[\mathbf{Z}(\mathbf{s}_0)] = \boldsymbol{\sigma}_0$. El modelo (2.15) es llamado Modelo de Kriging Universal.

Capítulo 3

Una prueba para la hipótesis de Campo Aleatorio Gaussiano

En este capítulo se propone una prueba para la hipótesis H_0 de un Campo Aleatorio Gaussiano. Es decir, se trata de determinar si un conjunto de datos espaciales puede ser considerado como una realización de un Campo Aleatorio Gaussiano.

A continuación se presenta una proposición que transforma a una variable aleatoria con distribución normal multivariada en una con distribución normal multivariada estandar, la cual será usada para la construcción de la prueba para probar H_0 .

3.1 Proposición

Supóngase que $\mathbf{X}_1, \dots, \mathbf{X}_m$ son vectores aleatorios independientes e idénticamente distribuidos en \mathbb{R}^p , $p \geq 1$. Sea $N^n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ que denota a la densidad normal n -variada con vector de medias $\boldsymbol{\mu}$ y matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$. Además sea $\mathbf{0}$ el vector nulo de orden n y sea \mathbf{I} la matriz identidad de orden $n \times n$. Entonces, por propiedades de la distribución normal multivariada, se tiene la siguiente proposición,

$$\mathbf{X} \sim N^n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ si y solo si } \mathbf{Z}^* = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu}) \sim N^n(\mathbf{0}, \mathbf{I}) \quad (3.1)$$

3.2 La familia Matérn de funciones de covarianza

En la subsección 2.5.2 fue presentada la función de covarianza Matérn.

Nótese que los parámetros ϕ y κ de la función no son ortogonales, en el sentido de que los parámetros de escala correspondientes a diferentes órdenes de correlación no son directamente comparables. Esto quiere decir que si la estructura de correlación es Matérn con parámetros ϕ y κ , entonces el mejor ajuste de orden $\kappa^* \neq \kappa$ también tendrá $\phi^* \neq \phi$. La relación entre el rango práctico y el parámetro de escala ϕ depende del valor de κ . El rango práctico es aproximadamente 3ϕ , 4.75ϕ y 5.92ϕ para la función Matérn con $\kappa = 0.5$, 1.5 y 2.5 , respectivamente, y $\sqrt{3}\phi$ para la función de correlación Gaussiana. Handkock y Wallis (1994) sugieren una reparametrización de κ y ϕ a un par casi ortogonal κ y $\alpha = 2\phi\sqrt{\kappa}$. La reparametrización no produce cambios en el modelo, pero es relevante para la estimación de los parámetros.

3.3 Estimación de los parámetros del semivariograma Matérn

En la sección 3.2 se mencionó el problema que existe en la estimación de los parámetros del semivariograma Matérn. A continuación se presenta una forma de estimar los parámetros. Cabe destacar que en este método se puede usar Máxima Verosimilitud, ó Máxima Verosimilitud Restringida.

1. Se va usar el paquete `geoR` (Diggle y Ribeiro, 2001) que se encuentra disponible en el software estadístico R (R Core Team, 2013).
2. Se ajusta un semivariograma empírico de tipo Cressie-Hawkins (Cressie, 1993) mediante el comando `variog` escogiendo la opción `bin` y una *tendencia de primer orden*, proporcionando un vector de valores `uvec` y *25 pares de bins* como mínimo. Los bins son grupos (o clases) de distancias que contienen los puntos del semivariograma estimado.
3. La ejecución de `variog` devolverá un objeto que contiene una lista de componentes. Aquí se presentan algunos que son necesarios para obtener valores iniciales para los parámetros del modelo Matérn:
 - Un vector de distancias \mathbf{u} ,
 - Un vector, \mathbf{v} , de valores estimados del semivariograma relacionados con las distancias dadas en \mathbf{u} ,

3.3. Estimación de los parámetros del semivariograma Matérn

- Número de pares en cada *bin*,
4. Se usa la media de \mathbf{v} como el estimador no paramétrico del rango, el cual se denota como ϕ_{rp} .
 5. Se ajusta una regresión lineal simple de \mathbf{v} con \mathbf{u} ; es decir, \mathbf{v} como variable dependiente y \mathbf{u} como variable independiente.
 6. Se toma el intercepto de la regresión como el estimador no paramétrico de τ^2 , el cual se denota como τ_{ini}^2 .
 7. Se toma la diferencia de la varianza de los datos menos τ_{ini}^2 como el estimador no paramétrico de σ^2 , el cual se denota por σ_{ini}^2 .
 8. El rango está relacionado con los parámetros del modelo Matérn mediante la expresión $rango = 2\phi\sqrt{\kappa}$, lo cual implica $\phi = \frac{rango}{2\sqrt{\kappa}}$. De aquí se deduce que el estimador no paramétrico de ϕ , denotado por ϕ_{ini} , es $\phi_{ini} = \frac{\phi_{rp}}{2\sqrt{\kappa}}$, para algún valor conocido de κ .
 9. Se considera la función de verosimilitud dada por la expresión $L(\boldsymbol{\beta}; \sigma^2, \phi, \tau^2, \kappa) = L(\boldsymbol{\beta}; \boldsymbol{\theta}; \mathbf{Z}(\mathbf{s})) = \ln \{|\Sigma(\boldsymbol{\theta})|\} + n \ln(2\pi) + (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta})' \Sigma(\boldsymbol{\theta})^{-1} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta}) = L$, donde $\boldsymbol{\beta}$ y $\mathbf{X}(\mathbf{s})$ son como en la sección 4.1.
 10. Se considera un conjunto de valores posibles para κ ; i.e., $\kappa = 0.3, 0.6, \dots, 4$.
 11. Se escoge un valor de κ denotado por κ_i , empezando por el valor más pequeño y aumentando sucesivamente hasta alcanzar el valor más grande.
 12. Se maximiza la función de verosimilitud, mediante el comando *likfit*, usando valores iniciales σ_{ini}^2 , τ_{ini}^2 y $\phi_{ini} = \frac{\phi_{rp}}{2\sqrt{\kappa_i}}$ con κ_i fijo. Se obtendrá el conjunto de estimaciones de máxima verosimilitud de los parámetros $\{\tilde{\boldsymbol{\beta}}_i; \tilde{\sigma}_i^2, \tilde{\phi}_i, \tilde{\tau}_i^2, \kappa_i\}$ de donde la verosimilitud proporciona un valor $L(\tilde{\boldsymbol{\beta}}_i; \tilde{\sigma}_i^2, \tilde{\phi}_i, \tilde{\tau}_i^2, \kappa_i) = L_i$ para cada κ_i .
 13. Se ordenan los valores L_i en orden ascendente; esto es, $L_1 < L_2 < \dots < L_q$.
 14. L_q es el valor, más grande, de la verosimilitud que se obtiene al variar los valores de κ en el conjunto considerado. Es claro que L_q proviene de algún conjunto $\{\tilde{\boldsymbol{\beta}}_q; \tilde{\sigma}_q^2, \tilde{\phi}_q, \tilde{\tau}_q^2, \kappa_q\}$. Por consiguiente $\tilde{\sigma}_q^2, \tilde{\phi}_q, \tilde{\tau}_q^2, \kappa_q$ son los estimadores de máxima verosimilitud que se usarán para estimar el semivariograma Matérn.

3.4 Construcción de una prueba para probar H_0

Suponga que se tiene una realización $\mathbf{z}(\mathbf{s}) = [z(s_1), \dots, z(s_n)]'$ de $\mathbf{Z}(\mathbf{s})$ que fue observada espacialmente, donde cada $z(s_i)$ es un único valor que ha sido observado en el sitio s_i con coordenadas (x_i, y_i) . Se desea probar si $\mathbf{z}(\mathbf{s})$ constituye una realización de un campo aleatorio Gaussiano.

Esto es, se tiene el contraste de hipótesis:

$$H_0 : \mathbf{Z}(\mathbf{s}) \text{ es un campo aleatorio gaussiano}$$

$$H_1 : \mathbf{Z}(\mathbf{s}) \text{ no es un campo aleatorio gaussiano}$$

Este contraste es equivalente a,

$$H_0 : \mathbf{Z}(\mathbf{s}) \text{ tiene distribución } N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$$

$$H_1 : \mathbf{Z}(\mathbf{s}) \text{ no tiene distribución } N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$$

Donde:

$$- \mathbf{X}(\mathbf{s}) = \begin{bmatrix} 1 & x_1 & y_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & y_n \end{bmatrix}$$

$$- \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)', \boldsymbol{\beta} \text{ desconocido,}$$

$$- \boldsymbol{\Sigma}(\boldsymbol{\theta}) \text{ desconocida se puede calcular a partir de una función de semivariograma (desconocida)}$$

$$\text{con parámetro } \boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \kappa), \boldsymbol{\theta} \text{ desconocido.}$$

En esta prueba se estima la función de semivariograma por medio de un semivariograma Matérn para así estimar $\boldsymbol{\Sigma}(\boldsymbol{\theta})$:

$$(\boldsymbol{\Sigma}(\boldsymbol{\theta}))_{ij} = \tau^2 \mathbf{I}(i = j) + \sigma^2 \frac{\Gamma(k)}{2^{k-1}} \left(\frac{d_{ij}}{\phi}\right)^k K_\kappa\left(\frac{d_{ij}}{\phi}\right), d_{ij} > 0, \kappa > 0, \phi > 0$$

Los valores σ^2 , ϕ , τ^2 y κ son los parámetros cima parcial, parámetro de escala, pepita y parámetro de forma respectivamente.

3.5. Estudio de simulación

Para probar H_0 se hace uso de la proposición que se encuentra en la sección 3.1.

Sean $\mathbf{X}(\mathbf{s})\hat{\boldsymbol{\beta}}$ y $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ la media y la matriz de varianzas y covarianzas estimadas. En $\mathbf{X}(\mathbf{s})\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}$ es el Estimador de Máxima Verosimilitud (EMV), ó el de Máxima Verosimilitud Restringida (EMVR), de $\boldsymbol{\beta}$ y la matriz $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ se estima usando una función de semivariograma Matérn con parámetro $\hat{\boldsymbol{\theta}}$, donde $\hat{\boldsymbol{\theta}}$ es el EMV, ó el EMVR, de $\boldsymbol{\theta}$. Ahora, sea $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1/2}$ la raíz cuadrada definida positiva simétrica de $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}$, la inversa de $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$.

Cuando $\mathbf{Z}(\mathbf{s})$ tiene distribución $N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, por la proposición (3.1), el vector aleatorio $\mathbf{Z}^* = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1/2}(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\hat{\boldsymbol{\beta}})$ tiene distribución aproximada $N^n(\mathbf{0}, \mathbf{I})$, lo que significa que las coordenadas del vector \mathbf{Z}^* , i.e. (Z_1, \dots, Z_n) , son aproximadamente independientes con distribución $N(0, 1)$. Es decir, \mathbf{Z}^* es una muestra aleatoria univariada de la distribución normal estándar.

Entonces, para probar la hipótesis H_0 se puede usar alguna prueba conocida para probar normalidad univariada. En este trabajo se usa la prueba de Anderson-Darling (AD) (Anderson y Darling, 1954) debido a que es la que proporciona mejores resultados en un estudio de simulación en términos de tamaño y potencia. Cabe aclarar que se hicieron estudios de simulación usando la prueba de Shapiro-Wilk y la de AD, pero la prueba de AD aproximó mejor el tamaño nominal de la prueba.

Finalmente, para probar H_0 se aplica la prueba de Anderson-Darling al vector \mathbf{Z}^* , con el objetivo de probar normalidad univariada. Por lo tanto, la regla de decisión es:

- Si AD rechaza normalidad univariada en \mathbf{Z}^* , entonces se rechaza H_0 .

De aquí en adelante, la prueba para probar H_0 recibe el nombre de prueba PCAG.

3.5 Estudio de simulación

3.5.1 Algoritmo para probar H_0

La prueba propuesta para H_0 que se presenta a continuación se programó en el software R:

3.5. Estudio de simulación

1. Usando los datos observados $\mathbf{z}(\mathbf{s})$, se estima por máxima verosimilitud los parámetros del modelo; i.e., $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ y $\hat{\boldsymbol{\theta}} = (\hat{\sigma}^2, \hat{\phi}, \hat{\tau}^2, \hat{\kappa})$. Los detalles computacionales de la estimación por máxima verosimilitud para el semivariograma Matérn se encuentran en la sección (3.3).
2. Usando $\hat{\boldsymbol{\theta}}$ y el semivariograma Matérn se calcula $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$.
3. Se obtiene el vector $\mathbf{Z}^* = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1/2} (\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\hat{\boldsymbol{\beta}})$
4. Usando el vector \mathbf{Z}^* , se aplica la prueba de Anderson-Darling para probar normalidad univariada.
5. En caso de que la hipótesis nula, de la prueba de Anderson Darling, sea rechazada se concluye que $\mathbf{Z}(\mathbf{s})$ no tiene distribución $N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$. Es decir, se rechaza H_0 .

3.5.2 Simulación

Se llevó a cabo un estudio de simulación de Monte Carlo para estimar el tamaño y la potencia de la prueba. Este estudio fue realizado bajo las siguientes condiciones:

- Se usa $N = 100$ replicaciones Monte Carlo.
- La simulación de las muestras se realiza en mallas irregulares de $(0, 50) \times (0, 50)$ con diferentes tamaños de muestra, n , $\{100, 200, 300, 400, 500\}$.
- Las muestras, bajo H_0 , son simuladas usando seis esquemas de dependencia (semivariogramas) espacial: *matern*, *exponencial*, *circular*, *gausiano*, *esférico* y *cúbico*. En la sección 2.5.2 se encuentra un resumen acerca de los distintos modelos de semivariograma.
- Para la simulación de las muestras bajo H_0 fueron considerados dos casos:
 1. Datos con media constante, con los siguientes parámetros: $\beta_0 = 3$, $\beta_1 = 0$, $\beta_2 = 0$, $\sigma^2 = 4$, $\tau^2 = 2$ y $\phi = 2, 0.5, 0.01$. Nótese que, en este caso, se fijaron los valores de β_0 , β_1 , β_2 , σ^2 y τ^2 y se hizo variar el valor de ϕ . Esto es con el objetivo de estudiar el tamaño estimado de la prueba bajo diferentes niveles de correlación.
 2. Datos con media dependiente de las coordenadas espaciales, con los siguientes parámetros: $\beta_0 = 3$, $\beta_1 = 3$, $\beta_2 = 3$, $\sigma^2 = 4$, $\tau^2 = 2$ y $\phi = 2, 0.5, 0.01$. En este caso, se fijaron los valores de β_0 , β_1 , β_2 , σ^2 y τ^2 y se hizo variar el valor de ϕ . Esto es con el objetivo de estudiar el tamaño de la prueba bajo diferentes niveles de correlación.

3.6. Discusión de resultados

- Las muestras, bajo H_1 , fueron simulados usando campos aleatorios no gaussianos: *T*, *Chi-cuadrada*, *Poisson*, *Logistico*, *Gumbel-Malik-Abraham* y *Binario*. Los detalles sobre la simulación de campos aleatorios no gaussianos se encuentran en el apéndice [A](#) de este trabajo.
- Los parámetros son estimados usando un semivariograma *Matérn*, por los métodos de máxima verosimilitud (ml) y máxima verosimilitud restringida (reml).
- El tamaño nominal de la prueba se fija en $\alpha = 0.05$.

3.6 Discusión de resultados

A continuación se presentan los resultados obtenidos del estudio de simulación; también se presenta la discusión de los mismos.

3.6. Discusión de resultados

Tabla 3.1: Tamaño estimado de la prueba PCAG.

Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$ con parámetros $\beta_0 = 3$, $\beta_1 = 0$, $\beta_2 = 0$, $\sigma^2 = 4$, $\phi = 2$, $\tau^2 = 1$ y $\kappa = 1.1$.

$H_0 : \mathbf{Z}(\mathbf{s}) \sim N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \kappa)$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula con el semivariograma Matérn.

Los modelos de semivariograma están definidos en la subsección 2.5.2.

	Tipo de Semivariograma					
	Matérn	Exponencial	Gausiano	Circular	Cúbico	Esférico
<i>n</i> = 100						
ml	0.08	0.08	0.08	0.09	0.09	0.10
reml	0.04	0.08	0.07	0.04	0.06	0.04
<i>n</i> = 200						
ml	0.05	0.07	0.09	0.10	0.10	0.08
reml	0.07	0.04	0.08	0.02	0.02	0.05
<i>n</i> = 300						
ml	0.08	0.07	0.06	0.09	0.09	0.05
reml	0.08	0.08	0.07	0.06	0.04	0.08
<i>n</i> = 400						
ml	0.07	0.06	0.07	0.09	0.10	0.03
reml	0.08	0.08	0.04	0.02	0.04	0.06
<i>n</i> = 500						
ml	0.07	0.07	0.06	0.09	0.09	0.04
reml	0.08	0.08	0.08	0.08	0.06	0.07

La Tabla 3.1 muestra los resultados de tamaño estimado de la prueba para datos con media constante ($\boldsymbol{\beta} = (3, 0, 0)$) con parámetros espaciales $\sigma^2 = 4$, $\phi = 2$, $\tau^2 = 1$ y $\kappa = 1.1$ para el semivariograma.

Se tienen estimaciones en donde el tamaño estimado es igual al tamaño nominal. Por ejemplo usando el semivariograma Matérn mediante ML y muestras de tamaño 200. Otro caso se tiene al usar el semivariograma esférico y REML con muestras de tamaño 200. Un caso más se localiza en la columna del semivariograma esférico en conjunción con ML usando muestras de tamaño 300.

Se tienen tamaños estimados que se aproximan al tamaño nominal. Por ejemplo, note que se tienen estimaciones de 0.04, 0.06 y 0.04 con muestras de tamaño 100 en el reglón del método

3.6. Discusión de resultados

REML en los semivariogramas circular, cúbico y esférico respectivamente. Se tienen estimaciones de 0.04 con muestras de tamaño 400 mediante REML usando los semivariogramas gaussiano y cúbico. Un caso más es del semivariograma exponencial mediante ML usando muestras de tamaño 400. Existen otros casos que se pueden encontrar al inspeccionar la Tabla 3.1.

Existen varios casos en donde el tamaño nominal es de 0.07. Como ejemplo, considerando el semivariograma Matérn y el método REML con muestras de tamaños 400 y 500. Otro caso es el del semivariograma gaussiano mediante REML y muestras de tamaño 300.

Por otro lado, se pueden encontrar tamaños estimados de 0.10 lo cual ya está muy lejos del tamaño nominal. Por ejemplo, el método ML produce tamaños estimados de 0.10 en los semivariogramas circular y cúbico en muestras de tamaños 200 y 400. También ML produce 0.10 con el semivariograma esférico con muestras de tamaño 100.

Las estimaciones más extremas, con respecto al tamaño nominal, se tienen con los semivariogramas circular, cúbico y esférico mediante ML en muestras de tamaños menores o iguales a 400. Estas estimaciones extremas son 0.09 y 0.10.

De manera global, el método ML produce estimaciones que van de 0.03 a 0.10 y REML produce estimaciones que van de 0.02 a 0.09. Es decir, las estimaciones de REML están mejor centradas en el tamaño nominal que ML. Es decir, REML produce mejores estimaciones que ML.

3.6. Discusión de resultados

Tabla 3.2: Tamaño estimado de la prueba PCAG.

Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$ con parámetros $\beta_0 = 3$, $\beta_1 = 3$, $\beta_2 = 3$, $\sigma^2 = 4$, $\phi = 2$, $\tau^2 = 1$ y $\kappa = 1.1$.

$H_0 : \mathbf{Z}(\mathbf{s}) \sim N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \kappa)$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula con el semivariograma Matérn.

Los modelos de semivariograma están definidos en la subsección 2.5.2.

	Tipo de Semivariograma					
	Matérn	Exponencial	Gausiano	Circular	Cúbico	Esférico
<i>n</i> = 100						
ml	0.06	0.08	0.10	0.10	0.08	0.10
reml	0.06	0.05	0.07	0.06	0.06	0.06
<i>n</i> = 200						
ml	0.05	0.07	0.09	0.09	0.09	0.09
reml	0.05	0.06	0.08	0.06	0.04	0.04
<i>n</i> = 300						
ml	0.07	0.06	0.09	0.07	0.06	0.09
reml	0.04	0.04	0.09	0.07	0.04	0.04
<i>n</i> = 400						
ml	0.06	0.07	0.07	0.08	0.08	0.08
reml	0.08	0.04	0.08	0.06	0.02	0.04
<i>n</i> = 500						
ml	0.07	0.07	0.08	0.08	0.07	0.08
reml	0.08	0.04	0.08	0.06	0.06	0.04

La Tabla 3.2 muestran los resultados de tamaño estimado de la prueba para datos con media no constante ($\boldsymbol{\beta} = (3, 3, 3)$) con parámetros espaciales $\sigma^2 = 4$, $\phi = 2$, $\tau^2 = 1$ y $\kappa = 1.1$ para el semivariograma.

Del mismo modo que para muestras con media constante, en este caso, se tienen estimaciones que están muy cerca del tamaño nominal. Por ejemplo, considerando el método REML y muestras de tamaño 100, se tienen estimaciones de 0.06 usando los semivariogramas circular, cúbico y esférico. Se tienen estimaciones de 0.04 usando el método REML con muestras de tamaño 300 con el semivariograma exponencial y Matérn. Otro caso es el de los semivariogramas circular y cubico mediante REML y muestras de tamaño 500. Se tienen tamaños estimados de 0.04.

También se tienen tamaños estimados que son iguales al tamaño nominal. Por ejemplo, los métodos de estimación ML y REML producen 0.05 con el semivariograma Matérn usando muestras de

3.6. Discusión de resultados

tamaño 200. Otro caso es del semivariograma exponencial en conjunción con REML y muestras de tamaño 100.

Se tienen tamaños estimados de 0.07 los cuales no están lejos del tamaño nominal. Por ejemplo, usando muestras de tamaño 400 el método ML produce estimaciones de 0.07 mediante los semivariogramas exponencial y gaussiano. Del mismo modo, ML produce 0.07 al usar los semivariogramas exponencial y Matérn con muestras de tamaño 500.

Nótese que al usar el semivariograma exponencial (considerando todos los tamaños de muestra) mediante REML se tienen tamaños estimados que van de 0.04 a 0.06. El desempeño de la prueba es bueno en este semivariograma. Lo mismo ocurre en el caso del semivariograma esférico mediante REML donde los tamaños estimados también van de 0.04 a 0.06.

Con respecto a muestras de tamaño 500, considerando todos los semivariogramas, las estimaciones de REML van de 0.04 a 0.08 y las estimaciones de ML van de 0.07 a 0.08. Es decir, las estimaciones que produce ML son mayores al tamaño nominal, lo cual es no deseable.

Por otro lado, los peores tamaños estimados están en la columna del semivariograma esférico mediante ML. Estos tamaños estimados van de 0.08 a 0.10 los cuales son mucho mayores que el tamaño nominal.

Con respecto al semivariograma Matérn las mejores estimaciones las produce ML. Las estimaciones van de 0.05 a 0.07; es decir iguales y no demasiado mayores que el tamaño nominal.

De forma global, el método REML produce tamaños estimados cuyas diferencias absolutas con el tamaño nominal no exceden de 0.04, para todos los semivariogramas y tamaños de muestra. El método ML presenta diferencias absolutas de hasta 0.05. Por lo tanto el método REML es mejor que ML.

3.6. Discusión de resultados

Tabla 3.3: Tamaño estimado de la prueba PCAG.

Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$ con parámetros $\beta_0 = 3$, $\beta_1 = 0$, $\beta_2 = 0$, $\sigma^2 = 4$, $\phi = 0.5$, $\tau^2 = 1$ y $\kappa = 1.1$.

$H_0 : \mathbf{Z}(\mathbf{s}) \sim N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \kappa)$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula con el semivariograma Matérn.

Los modelos de semivariograma están definidos en la subsección 2.5.2.

	Tipo de Semivariograma					
	Matérn	Exponencial	Gausiano	Circular	Cúbico	Esférico
<i>n</i> = 100						
ml	0.04	0.09	0.07	0.04	0.06	0.04
reml	0.02	0.08	0.05	0.03	0.05	0.07
<i>n</i> = 200						
ml	0.07	0.04	0.08	0.01	0.02	0.05
reml	0.06	0.05	0.04	0.05	0.03	0.04
<i>n</i> = 300						
ml	0.08	0.08	0.07	0.06	0.04	0.08
reml	0.07	0.03	0.07	0.04	0.05	0.08
<i>n</i> = 400						
ml	0.03	0.08	0.04	0.02	0.04	0.06
reml	0.02	0.08	0.08	0.08	0.04	0.03
<i>n</i> = 500						
ml	0.04	0.04	0.08	0.08	0.06	0.07
reml	0.04	0.02	0.08	0.03	0.04	0.03

La Tabla 3.3 muestra los resultados de tamaño estimado de la prueba para muestras con media constante ($\boldsymbol{\beta} = (3, 0, 0)$) con parámetros espaciales $\sigma^2 = 4$, $\phi = 0.5$, $\tau^2 = 1$ y $\kappa = 1.1$ para el semivariograma. Nótese que en esta tabla se usa el parámetro $\phi = 0.5$ y los valores de los parámetros restantes son los mismos que se usaron en las Tablas 3.1 y 3.2. Esto es con el objetivo de estudiar el desempeño de la prueba al usar diferentes niveles de correlación espacial manteniendo fijos los parámetros restantes.

Nótese que se tienen casos donde el tamaño estimado es igual al tamaño nominal. Por ejemplo, considere muestras de tamaño 200 en conjunción con el método REML. Los tamaños estimados son de 0.05 en los semivariogramas exponencial y circular. Otra vez, considerando muestras de tamaño 200 pero en conjunción con ML, se tiene un tamaño estimado de 0.05 con el semivariograma esférico. Otros casos son con el método REML usando muestras de tamaño 100 en donde se tiene 0.05 con los semivariogramas gaussiano y cúbico.

3.6. Discusión de resultados

También se tienen varios casos en donde el tamaño estimado es muy cercano al tamaño nominal; es decir, se tienen tamaños estimados de 0.04 y 0.06. Considerando muestras de tamaño 400 y el método ML, se tienen tamaños estimados de 0.04 con los semivariogramas gaussiano y cúbico. Los métodos ML y REML producen tamaños estimados de 0.04 al usar muestras de tamaño 500 con el semivariograma Matérn. Se tienen tamaños estimados de 0.04 y 0.06 en muestras de tamaño 500 mediante ML con los semivariogramas exponencial y cúbico, respectivamente. Existen otros casos que se pueden encontrar al inspeccionar la Tabla 3.3.

Se tienen otros casos en donde el tamaño estimado es cercano al tamaño nominal. Por ejemplo, REML y ML producen 0.07 al usar muestras de tamaño 300 con el semivariograma gaussiano. El método ML produce 0.07 al usar muestras de tamaño 200 con el semivariograma Matérn. Otra vez, ML produce 0.07 con muestras de tamaño 500 con el semivariograma esférico.

Por otro lado las estimaciones extremas son 0.01, 0.08 y 0.09. Por ejemplo, al usar muestras de tamaño 300, el método ML produce 0.08 con los semivariogramas exponencial y esférico. Existe un único caso de 0.01, el cual se localiza en la columna del semivariograma circular mediante ML en conjunción con muestras de tamaño 100. También existe un único caso de 0.09 el cual se encuentra en la columna del semivariograma exponencial mediante ML usando muestras de tamaño 100.

Para el caso de muestras de tamaño 500 (considerando todos los semivariogramas), las estimaciones de REML van de 0.02 a 0.08. Las estimaciones de ML van de 0.03 a 0.08. Es decir, ML es mejor en estos tamaños de muestra.

Se puede notar que de manera global REML produce tamaños estimados que van de 0.02 a 0.08 mientras que los que produce ML van de 0.01 a 0.09. Es decir, las estimaciones de REML están mejor centradas en el tamaño nominal que ML.

3.6. Discusión de resultados

Tabla 3.4: Tamaño estimado de la prueba PCAG.

Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$ con parámetros $\beta_0 = 3$, $\beta_1 = 3$, $\beta_2 = 3$, $\sigma^2 = 4$, $\phi = 0.5$, $\tau^2 = 1$ y $\kappa = 1.1$.

$H_0 : \mathbf{Z}(\mathbf{s}) \sim N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \kappa)$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula con el semivariograma Matérn.

Los modelos de semivariograma están definidos en la subsección 2.5.2.

	Tipo de Semivariograma					
	Matérn	Exponencial	Gausiano	Circular	Cúbico	Esférico
<i>n</i> = 100						
ml	0.04	0.05	0.04	0.07	0.04	0.02
reml	0.03	0.05	0.05	0.09	0.04	0.05
<i>n</i> = 200						
ml	0.07	0.04	0.04	0.05	0.06	0.06
reml	0.06	0.08	0.04	0.08	0.05	0.07
<i>n</i> = 300						
ml	0.07	0.03	0.04	0.04	0.04	0.02
reml	0.02	0.08	0.05	0.05	0.04	0.02
<i>n</i> = 400						
ml	0.06	0.04	0.04	0.08	0.02	0.04
reml	0.02	0.02	0.08	0.02	0.03	0.04
<i>n</i> = 500						
ml	0.06	0.04	0.08	0.07	0.01	0.06
reml	0.05	0.03	0.07	0.08	0.03	0.07

La Tabla 3.4 muestra los resultados de tamaño estimado de la prueba para datos con media no constante ($\boldsymbol{\beta} = (3, 3, 3)$) con parámetros espaciales $\sigma^2 = 4$, $\phi = 0.5$, $\tau^2 = 1$ y $\kappa = 1.1$ para el semivariograma. En esta tablas se tiene el parámetro $\phi = 0.5$ y los valores de los parámetros restantes son los mismos que se usaron en las tablas 3.1, 3.2 y 3.3.

Nótese que se tienen tamaños estimados iguales al tamaño nominal. El método REML produce 0.05 usando muestras de tamaño 500 en conjunción con el semivariograma Matérn. Otra vez REML produce tamaños estimados de 0.05 con muestras de tamaño 300 al usar los semivariogramas gaussiano y circular. El método ML produce 0.05 en el semivariograma exponencial con muestras de tamaño 100. Se pueden encontrar otros casos al inspeccionar la Tabla 3.4.

Se tienen tamaños estimados de 0.04 y 0.06. Para muestras de tamaño 300, se tiene 0.04 con el semivariograma cúbico por los dos métodos de estimación. Se tienen estimaciones de 0.06 y 0.04

3.6. Discusión de resultados

usando los semivariogramas Matérn y exponencial para muestras de tamaños 400 y 500 mediante ML. Se tienen tamaños estimados de 0.06 usando muestras de tamaño 200 mediante ML en los semivariogramas cúbico y esférico. Se pueden encontrar más casos al inspeccionar la Tabla 3.4.

También se tienen casos en donde el tamaño estimado es cercano al tamaño nominal. Por ejemplo, en la columna del semivariograma Matérn en conjunción con muestras de tamaño 100, el método REML produce un tamaño estimado de 0.03. En la columna del semivariograma esférico en conjunción con REML y muestras de tamaño 200 se tiene un tamaño estimado de 0.07. Existen algunos otros casos de 0.03 y 0.07 los cuales se pueden encontrar al inspeccionar la Tabla 3.4.

Por otro lado, las peores estimaciones, las más extremas, son 0.01 y 0.09. Se tiene únicamente un caso de 0.01, el cual se encuentra en la columna del semivariograma cúbico en conjunción con muestras de tamaño 500 mediante ML. Del mismo modo, se tiene únicamente un caso de 0.09, el cual se encuentra en la columna del semivariograma circular en conjunción con muestras de tamaño 100 mediante REML

Para el caso de muestras de tamaño 500, considerando todos los semivariogramas, REML produce tamaños estimados que van de 0.02 a 0.09 mientras que ML produce estimaciones que van de 0.01 a 0.08.

De forma global, los tamaños estimados por el método ML van de 0.01 a 0.08. El método REML presenta estimaciones que van de 0.02 a 0.09. Pero, considerando que las estimaciones extremas son solamente una por cada extremo, y que son presentadas, una por ML y la otra por REML, se puede concluir que ambos métodos son iguales; es decir producen estimaciones casi iguales.

3.6. Discusión de resultados

Tabla 3.5: Tamaño estimado de la prueba PCAG.

Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$ con parámetros $\beta_0 = 3$, $\beta_1 = 0$, $\beta_2 = 0$, $\sigma^2 = 4$, $\phi = 0.01$, $\tau^2 = 1$ y $\kappa = 1.1$.

$H_0 : \mathbf{Z}(\mathbf{s}) \sim N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \kappa)$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula con el semivariograma Matérn.

Los modelos de semivariograma están definidos en la subsección 2.5.2.

	Tipo de Semivariograma					
	Matérn	Exponencial	Gausiano	Circular	Cúbico	Esférico
<i>n</i> = 100						
ml	0.06	0.05	0.04	0.04	0.09	0.03
reml	0.07	0.03	0.04	0.04	0.02	0.03
<i>n</i> = 200						
ml	0.06	0.08	0.04	0.05	0.04	0.05
reml	0.05	0.03	0.04	0.04	0.06	0.03
<i>n</i> = 300						
ml	0.05	0.04	0.03	0.04	0.04	0.06
reml	0.06	0.04	0.06	0.08	0.04	0.02
<i>n</i> = 400						
ml	0.05	0.07	0.07	0.02	0.04	0.06
reml	0.07	0.08	0.03	0.08	0.04	0.08
<i>n</i> = 500						
ml	0.02	0.07	0.02	0.05	0.06	0.03
reml	0.08	0.08	0.03	0.08	0.05	0.06

La Tabla 3.5 muestra los resultados de tamaño estimado de la prueba para datos con media constante ($\boldsymbol{\beta} = (3, 0, 0)$) con parámetros espaciales $\sigma^2 = 4$, $\phi = 0.01$, $\tau^2 = 1$ y $\kappa = 1.1$ para el semivariograma. Nótese que en estas tablas se tiene el parámetro $\phi = 0.01$ y los valores de los parámetros restantes son los mismos que fueron usados en las tablas 3.1, 3.2, 3.3 y 3.4. Se está estudiando el comportamiento de la prueba a diferentes niveles de correlación espacial manteniendo fijos los valores de los parámetros restantes.

Se tienen estimaciones en donde el tamaño estimado es igual al tamaño nominal. Por ejemplo usando el semivariograma Matérn mediante REML y muestras de tamaño 200. Otro caso se tiene al usar el semivariograma esférico y REML con muestras de tamaño 200. Otros casos se encuentran en la columna del semivariograma Matérn en la intersección con ML donde se tienen tamaños estimados de 0.05 al usar muestras de tamaños 300 y 400. Existen muchos casos más que se pueden encontrar al inspeccionar la Tabla 3.5.

3.6. Discusión de resultados

También se tienen varios casos en donde el tamaño estimado es muy cercano al tamaño nominal; es decir, se tienen tamaños estimados de 0.04 y 0.06. Por ejemplo, considerando muestras de tamaño 300 y el método ML, se tienen tamaños estimados de 0.04 con los semivariogramas exponencial y circular. Los métodos ML y REML producen tamaños estimados de 0.04 al usar muestras de tamaño 400 con el semivariograma cúbico. Se tiene un tamaño estimado de 0.06 en muestras de tamaño 500 mediante REML usando el semivariograma esférico. Existen más casos que se pueden encontrar al inspeccionar la Tabla 3.5.

Existen varios casos en donde el tamaño nominal es de 0.07. Como ejemplo considerando el semivariograma Matérn en conjunción con REML usando muestras de tamaños 100 y 400. Otro caso es el del semivariograma gaussiano mediante ML y muestras de tamaño 400. Otro caso más es del semivariograma exponencial en conjunción con muestras de tamaño 500 en donde ML produce el tamaño estimado de 0.07.

Un caso notable está en la columna del semivariograma cúbico. Aunque este presente la estimación más extrema de 0.09, mediante ML, se puede ver que el tamaño estimado se aproxima al tamaño nominal cuando el tamaño de muestra crece. Es decir se tiene evidencia de consistencia con este semivariograma. Y lo mismo ocurre con REML al usar este semivariograma.

Por otro lado, los peores tamaños estimados son 0.02, 0.08 y 0.09. Por ejemplo, el método ML produce un tamaño estimado de 0.02 en el semivariograma cúbico usando tamaños de muestra 100. ML produce 0.08 con el semivariograma circular al usar muestras de tamaño 300. Se tiene únicamente un caso de 0.09, el cual se localiza en la columna del semivariograma cúbico mediante ML usando muestras de tamaño 100.

Para el caso de muestras de tamaño 500, considerando todos los semivariogramas, REML produce tamaños estimados que van de 0.03 a 0.08 mientras que ML produce estimaciones que van de 0.02 a 0.07.

De manera global, el método ML produce estimaciones que van de 0.02 a 0.09 y REML produce estimaciones que van de 0.02 a 0.08. Es decir, las estimaciones de REML están mejor centradas en el tamaño nominal que ML.

3.6. Discusión de resultados

Tabla 3.6: Tamaño estimado de la prueba PCAG.

Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$ con parámetros $\beta_0 = 3$, $\beta_1 = 3$, $\beta_2 = 3$, $\sigma^2 = 4$, $\phi = 0.01$, $\tau^2 = 1$ y $\kappa = 1.1$.

$H_0 : \mathbf{Z}(\mathbf{s}) \sim N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \kappa)$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula con el semivariograma Matérn.

Los modelos de semivariograma están definidos en la subsección 2.5.2.

	Tipo de Semivariograma					
	Matérn	Exponencial	Gausiano	Circular	Cúbico	Esférico
<i>n</i> = 100						
ml	0.05	0.06	0.02	0.06	0.02	0.05
reml	0.04	0.08	0.02	0.03	0.02	0.09
<i>n</i> = 200						
ml	0.03	0.05	0.06	0.04	0.07	0.05
reml	0.05	0.03	0.07	0.04	0.01	0.04
<i>n</i> = 300						
ml	0.06	0.06	0.05	0.02	0.03	0.05
reml	0.04	0.05	0.05	0.02	0.06	0.04
<i>n</i> = 400						
ml	0.08	0.06	0.07	0.05	0.06	0.02
reml	0.03	0.07	0.08	0.06	0.06	0.02
<i>n</i> = 500						
ml	0.04	0.06	0.06	0.03	0.03	0.06
reml	0.05	0.03	0.06	0.06	0.05	0.05

La Tabla 3.6 muestran los resultados de tamaño estimado de la prueba para datos con media no constante ($\boldsymbol{\beta} = (3, 3, 3)$) con parámetros espaciales $\sigma^2 = 4$, $\phi = 0.01$, $\tau^2 = 1$ y $\kappa = 1.1$ para el semivariograma. En esta tabla se usa el parámetro $\phi = 0.01$ y los valores de los parámetros restantes son los mismos que son usados en las tablas 3.1, 3.2, 3.3, 3.4 y 3.5.

Nótese que se tienen casos donde el tamaño estimado es igual al tamaño nominal. Por ejemplo, considere muestras de tamaño 500 en conjunción con el método REML. Los tamaños estimados son de 0.05 en los semivariogramas Matérn, cúbico y esférico. Otra vez, considerando muestras de tamaño 300 en conjunción con REML, se tienen tamaños estimados de 0.05 con los semivariogramas exponencial y gaussiano. Otros casos son, con el método ML usando muestras de tamaño 100 en donde se tiene 0.05 con los semivariogramas Matérn y esférico.

También se tienen varios casos en donde el tamaño estimado es muy cercano al tamaño nominal;

3.6. Discusión de resultados

es decir, se tienen estimaciones de 0.04 y 0.06. Considerando muestras de tamaño 300 y el método REML, se tienen tamaños estimados de 0.04 con los semivariogramas Matérn y esférico. Los métodos ML y REML producen tamaños estimados de 0.06 al usar muestras de tamaño 500 con el semivariograma gaussiano. Se tienen tamaños estimados de 0.04 y 0.06 en muestras de tamaño 200 mediante ML con los semivariogramas circular y gaussiano, respectivamente. Existen más casos que se pueden encontrar al inspeccionar la Tabla 3.6.

Se tienen otros casos en donde el tamaño estimado es cercano al tamaño nominal. Por ejemplo, REML produce 0.07 al usar muestras de tamaño 200 con el semivariograma gaussiano. El método ML produce 0.07 al usar muestras de tamaño 200 con el semivariograma cúbico. Otra vez, ML produce 0.07 con muestras de tamaño 400 usando el semivariograma gaussiano.

Un caso notable se puede ver en la columna del semivariograma exponencial (se está considerando todos los tamaños de muestra) en conjunción con el método ML. Los tamaños estimados van de 0.05 a 0.06, lo cual es muy deseable en una prueba de hipótesis, debido a que los tamaños estimados están muy cerca del tamaño nominal.

Por otro las estimaciones extremas son 0.01, 0.08 y 0.09. Por ejemplo, al usar muestras de tamaño 100, el método REML produce 0.08 con el semivariograma exponencial. Existe un único caso de 0.01, el cual se localiza en la columna del semivariograma cúbico mediante REML en conjunción con muestras de tamaño 200. También existe un único caso de 0.09 el cual se encuentra en la columna del semivariograma esférico mediante REML usando muestras de tamaño 100.

Para el caso de muestras de tamaño 500 (considerando todos los semivariogramas), las estimaciones de REML van de 0.03 a 0.06. Las estimaciones de ML van de 0.03 a 0.06. Es decir, se pueden considerar iguales ambos métodos para este tamaño de muestra.

Se puede notar que de manera global REML produce tamaños estimados que van de 0.01 a 0.09 mientras que los que produce ML van de 0.02 a 0.08. Es decir, las estimaciones de ML están mejor centradas en el tamaño nominal que REML.

Habiendo hecho el análisis de resultados del tamaño estimado de la prueba, se puede afirmar que los tamaños estimados son cercanos al tamaño nominal para muestras de tamaños iguales o mayores que 100, al usar ambos métodos de estimación, en ambos casos de datos con media constante y no constante. También se debe destacar que para valores de ϕ cercanos a 2 el método REML es mejor que el método ML. Sin embargo, las estimaciones que producen ambos métodos se van igualando conforme el valor de ϕ se va aproximando a 0.01.

3.6. Discusión de resultados

Tabla 3.7: Potencia estimada de la prueba PCAG.

Estudio de simulación de campos aleatorios no gaussianos, en una malla irregular de $(0,50) \times (0,50)$.

$H_1 : \mathbf{Z}(\mathbf{s}) \approx N^n(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ y $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ se calcula usando el semivariograma Matérn.

	Tipo de Campo Aleatorio					
	Binario	GMA	Logístico	T	Poisson	Chi-cuadrada
$n = 100$						
ml	1	0.92	0.91	0.22	0.43	1
reml	0.99	0.93	0.90	0.23	0.42	0.99
$n = 200$						
ml	0.99	0.97	0.96	0.20	0.49	1
reml	1	0.99	0.97	0.22	0.50	1
$n = 300$						
ml	0.98	1	1	0.21	0.57	1
reml	1	1	1	0.21	0.55	1
$n = 400$						
ml	1	1	1	0.22	0.62	1
reml	1	1	0.99	0.23	0.60	
$n = 500$						
ml	1	1	0.99	0.23	0.68	1
reml	1	1	1	0.24	0.69	1

La Tabla 3.7 muestra los resultados de la potencia estimada usando datos simulados de campos aleatorios no gaussianos.

Para tamaños de muestra iguales o mayores que 100, considerando los métodos ML y REML, se tienen potencias cercanas a uno al usar muestras provenientes de campos aleatorios de tipo Binario y Chi-Cuadrada. Las potencias son altas para muestras de tamaños iguales o mayores a 100 al usar estos semivariogramas.

Para los campos aleatorios de tipo Gumbel-Malik-Abraham y Logístico, considerando el método ML, se tienen potencias que van de 0.91 a 1. Se puede ver de manera clara, en estos dos semivariogramas, que la potencia crece rápidamente conforme el tamaño muestral aumenta. Lo mismo ocurre al considerar el método REML y estos dos semivariogramas. Es decir, se tiene evidencia de consistencia en la potencia estimada de la prueba.

3.7. Una aplicación

Para el campo aleatorio Poisson, considerando ambos métodos de estimación, se tienen potencias que van de 0.42 a 0.69 . Se puede ver que la potencia crece conforme el tamaño muestral aumenta. Otra vez se tiene evidencia de consistencia en la potencia estimada de la prueba.

Para el campo aleatorio T, considerando ambos métodos de estimación, se tienen potencias que van de 0.20 a 0.24. En este caso la potencia crece lentamente cuando el tamaño muestral aumenta. Esto podría deberse a la similitud que existe entre un Campo Aleatorio Gaussiano y un Campo Aleatorio T.

De manera general, las potencias estimadas son suficientemente significativas con respecto al tamaño nominal.

Finalmente ya habiendo hecho el analisis de resultados de tamaño y potencia de la prueba, se concluye que la metodología propuesta es adecuada para ser usada en aplicaciones.

3.7 Una aplicación

Se tiene un conjunto de datos que fueron recolectados en varias estaciones registradoras del Estado de Paraná, Brasil, pertenecientes a las siguientes empresas: COPEL, IAPAR, DNAEE, SUREHMA e INEMET. La base de datos fue organizada por Laura Regina Bernardes Kiihl (IAPAR, Instituto Agronómico del Paraná, Londrina, Brasil) y fue proporcionada por Jacinta Loudovico Zamboti (Universidad Estadual de Londrina, Brasil). Las coordenadas de los límites (bordes) del estado de Paraná fueron proporcionadas por Joao Henrique Caviglione (IAPAR). Este conjunto de datos fue utilizado por [Diggle y Ribeiro \(2001\)](#) para ilustrar los métodos discutidos en su artículo. Los datos se refieren a la precipitación media en diferentes años del período mayo-junio (estación seca). Fueron recogidos en 143 estaciones de recolección en todo el Estado de Paraná, Brasil.

3.7. Una aplicación

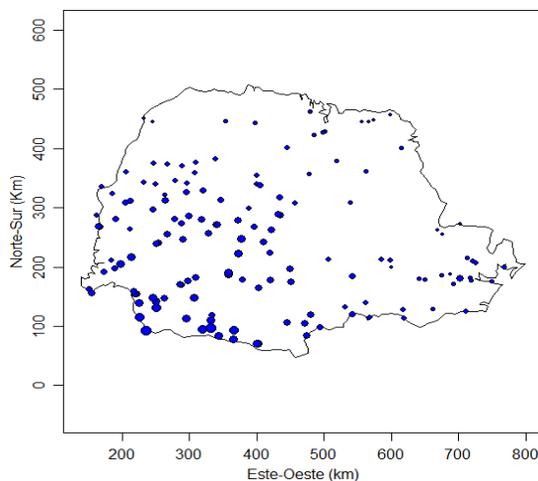


Figure 3.1: Mapa del estado de Paraná con 143 estaciones.

La figura 3.1 muestra las 143 estaciones distribuidas en el estado de Paraná, Brasil.

Se va a aplicar la metodología para verificar si los datos son una realización de un Campo Aleatorio Gaussiano. Se tiene el siguiente resultado:

El valor de la estadística de Anderson-Darling es $AD = 0.49667$, con $p - valor = 0.2095$.

Tomando $\alpha = 0.05$, se puede ver que $p - valor > \alpha$. Por lo tanto, no se rechaza la hipótesis nula y se concluye que los datos son una realización de un Campo Aleatorio Gaussiano.

Capítulo 4

Una prueba para la hipótesis de media constante de un Campo Aleatorio Gaussiano

En este capítulo se propone una prueba para la hipótesis de media constante de los datos, bajo el supuesto de que son una realización de un Campo Aleatorio Gaussiano.

Nótese que el hecho de probar Media Constante es equivalente a probar que se tiene un modelo de Kriging Ordinario lo cual es equivalente a probar Estacionariedad de Segundo Orden (o débil).

En la sección (2.2) se mostró una expresión general para modelar datos espaciales. Después en (2.13) se supone un modelo específico para $\mathbf{Z}(\mathbf{s})$. Esto es,

$$\mathbf{Z}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{e}(\mathbf{s}), \quad \mathbf{e}(\mathbf{s}) \sim (\mathbf{0}, \boldsymbol{\Sigma})$$

Usando este modelo se proponen términos particulares para sus componentes, resultando un modelo que se nombrará “modelo particular”.

4.1 Un modelo particular para $X(\mathbf{s})$

Sea el campo aleatorio $\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^2\}$, en donde $\mathbf{Z}(\mathbf{s}) = [Z(s_1), \dots, Z(s_n)]$, $\mathbf{s} = [s_1, \dots, s_n]'$, $s_i = (x_i, y_i)$, siendo x_i y y_i las coordenadas espaciales latitud y longitud respectivamente.

4.1. Un modelo particular para $X(\mathbf{s})$

Asuma que $\mathbf{Z}(\mathbf{s})$ tiene la forma del modelo (2.15), cuyas componentes son las siguientes expresiones particulares:

- $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]'$ es un vector de parámetros desconocidos,
- $\mathbf{X}(\mathbf{s}) = \begin{bmatrix} 1 & x_1 & y_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & y_n \end{bmatrix}$,
- $\mathbf{e}(\mathbf{s}) = [e(s_1), \dots, e(s_n)]' \sim N^n(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$.
- Los elementos de $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ van a ser calculados a partir de un modelo de semivariograma conocido, digamos exponencial dado por la expresión $[\boldsymbol{\Sigma}(\boldsymbol{\theta})]_{ij} = \sigma^2 \exp(-\phi d_{ij}) + \tau^2 I(i = j)$, $\sigma^2 > 0$, $\phi > 0$, $\tau^2 > 0$, donde σ^2 es la cima parcial, ϕ es el rango, τ^2 es la pepita y d_{ij} es la distancia euclidiana entre las ubicaciones s_i y s_j . Si el semivariograma no es conocido, entonces puede ser estimado usando un semivariograma Matérn como se propone en el capítulo 3.

Tomando en cuenta los puntos anteriores $\mathbf{Z}(\mathbf{s})$ tiene la forma,

$$\begin{bmatrix} Z(s_1) \\ Z(s_2) \\ \vdots \\ Z(s_n) \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \beta_2 y_1 \\ \beta_0 + \beta_1 x_2 + \beta_2 y_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \beta_2 y_n \end{bmatrix} + \begin{bmatrix} e(s_1) \\ e(s_2) \\ \vdots \\ e(s_n) \end{bmatrix} \quad (4.1)$$

Se puede ver que (4.1) es un modelo de Kriging Universal en donde $\mathbf{X}(\mathbf{s})\boldsymbol{\beta}$ es un modelo de superficie de respuesta de primer orden en las coordenadas de \mathbf{s} .

Nótese lo siguiente acerca de (4.1):

1. Cuando β_1 y β_2 son cero, $\mathbf{Z}(\mathbf{s})$ es un modelo de Kriging Ordinario.
2. $\mathbf{Z}(\mathbf{s})$ es un modelo de Kriging Universal si al menos β_1 o β_2 es diferente de cero.

4.2 Prueba de hipótesis lineal para efectos fijos

Los modelos con errores correlacionados tienen una solución de mínimos cuadrados generalizados para los efectos fijos. Similar al caso de cuando los errores son no correlacionados, se consideran hipótesis lineales que incluyen a $\boldsymbol{\beta}$, de la forma (Schabenberger y Gotway, 2005):

$$H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{I}_0 \quad \text{vs} \quad H_1 : \mathbf{L}\boldsymbol{\beta} \neq \mathbf{I}_0$$

Donde \mathbf{L} es una matriz de coeficientes de contraste de dimensión $l \times p$, $\boldsymbol{\beta}$ es un vector de dimensión $p \times 1$ e \mathbf{I}_0 es un vector $l \times 1$ conocido. La estadística de Wald (Wald, 1943) para probar H_0 está dada por la expresión:

$$\tilde{F} = \frac{(\mathbf{L}\tilde{\boldsymbol{\beta}} - \mathbf{I}_0)' \left[\mathbf{L} \left(\mathbf{X}(\mathbf{s})' \boldsymbol{\Sigma}(\tilde{\boldsymbol{\theta}})^{-1} \mathbf{X}(\mathbf{s}) \right)^{-1} \mathbf{L}' \right] (\mathbf{L}\tilde{\boldsymbol{\beta}} - \mathbf{I}_0)}{\text{rango}(\mathbf{L})} \quad (4.2)$$

En donde $\boldsymbol{\beta}$ es el estimador de mínimos cuadrados generalizados basado en $\boldsymbol{\theta}$, siendo este último el estimador de máxima verosimilitud o el de máxima verosimilitud restringida.

Bajo H_0 , \tilde{F} tiene distribución F con grados de libertad $\text{rango}\{\mathbf{L}\}$ y $(n - \text{rango}\{\mathbf{X}(\mathbf{s})\})$ (Harville y Jeske, 1992, Kackar y Harville, 1990, Prasad y Rao, 1990).

4.3 Una prueba para probar Kriging Ordinario

Se desea probar si la media del proceso es constante; es decir, si se tiene un modelo de Kriging Ordinario:

$$\begin{aligned} H_0 : \mathbf{Z}(\mathbf{s}) \text{ es un modelo de Kriging Ordinario} \\ H_1 : \mathbf{Z}(\mathbf{s}) \text{ es un modelo de Kriging Universal} \end{aligned}$$

Para hacer esto, se propone un contraste de hipótesis para el parámetro $\boldsymbol{\beta}$ del modelo (4.1),

$$H_0 : \beta_1 = \beta_2 = 0, \beta_0 > 0 \quad \text{vs} \quad H_1 : \beta_p \neq 0 \text{ para algún } p = 1, 2. \quad (4.3)$$

Para probar H_0 se hace uso de la estadística (4.2) asignando expresiones particulares a sus componentes de la siguiente forma:

4.4. Un estudio de simulación

- $\boldsymbol{\beta}$, $\mathbf{X}(s)$ y $\mathbf{e}(s)$ son como en la sección 4.1,
- \mathbf{L} es una matriz identidad de dimensión 3×3 ,
- \mathbf{I}_0 se estima por $\tilde{\mathbf{I}}_0 = \begin{bmatrix} n^{-1} \sum_{i=1}^n z(s_i) & 0 & 0 \end{bmatrix}'$.

La prueba rechaza H_0 cuando $\tilde{F} > F_\alpha$.

Nóte que \tilde{F} , en este caso, tiene una distribución que depende de \mathbf{I}_0 . Es decir que para probar H_0 se requiere calcular la distribución de \tilde{F} sujeto a que tenemos un valor desconocido \mathbf{I}_0 el cual se estima mediante la media de las observaciones $z(s_1), z(s_2), \dots, z(s_n)$.

Entonces, se hace uso del método de bootstrap paramétrico para encontrar la distribución de \tilde{F} bajo H_0 y así calcular F_α .

4.4 Un estudio de simulación

4.4.1 Bootstrap paramétrico para obtener F_α

Para aproximar la distribución de la estadística de prueba bajo la hipótesis nula, fue implementado un algoritmo de bootstrap paramétrico el cual se presenta a continuación:

1. Se usa un semivariograma exponencial para modelar la estructura de dependencias de $\boldsymbol{\Sigma}(\boldsymbol{\theta})$; es decir, $(\boldsymbol{\Sigma}(\boldsymbol{\theta}))_{ij} = \sigma^2 \exp(-\phi d_{ij}) + \tau^2 I(i = j)$, $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2)$.
2. Usando las observaciones disponibles, $\mathbf{z}(s) = [z(s_1), \dots, z(s_n)]'$, se calculan las estimaciones $\hat{\boldsymbol{\theta}}$, $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ y $\hat{\boldsymbol{\beta}}$.
3. Se calcula la media muestral de las observaciones disponibles, i.e., $\tilde{\mu}_0 = n^{-1} \sum_{i=1}^n z(s_i)$.
4. Usando la muestra $\mathbf{z}(s)$, la media $\tilde{\mu}_0$, el vector $\hat{\boldsymbol{\beta}}$ y la matriz de var-cov $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ se calcula la estadística \tilde{F} . Esto es,

$$\tilde{F} = \frac{(\hat{\boldsymbol{\beta}} - \tilde{\mathbf{I}}_0)' \left(\mathbf{X}(s)' \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1} \mathbf{X}(s) \right)^{-1} (\hat{\boldsymbol{\beta}} - \tilde{\mathbf{I}}_0)}{\text{rank}(\mathbf{L})} \quad \text{con } \tilde{\mathbf{I}}_0 = \begin{bmatrix} \tilde{\mu}_0 \\ 0 \\ 0 \end{bmatrix}.$$

4.4. Un estudio de simulación

5. Bajo la hipótesis nula, la distribución estimada de $\mathbf{Z}(\mathbf{s})$ es $N^n(\mathbf{X}(\mathbf{s})\tilde{\mathbf{I}}_0, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}))$.
6. Se extrae una muestra de tamaño 1 de $N^n(\mathbf{X}(\mathbf{s})\tilde{\mathbf{I}}_0, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}))$ y se obtiene la media muestral $\tilde{\mu}_0$ y las estimaciones de máxima verosimilitud $\hat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ y $\hat{\boldsymbol{\beta}}$.
7. Usando los valores estimados en el paso anterior, se calcula la estadística \tilde{F} .
8. Se repite m veces los pasos 6 y 7.
9. Usando los m valores calculados anteriormente, se obtiene el cuantil α , con $\alpha \in (0, 1)$, denotado por F_α . Este cuantil es el valor crítico para rechazar H_0 .
10. Si $\tilde{F} > F_\alpha$ entonces se rechaza H_0 y se concluye que $\beta_p \neq 0$, para algún $p = 1, 2$.

4.4.2 Simulación

Un estudio de simulación de Monte Carlo (MC) fue llevado a cabo con el fin de investigar las propiedades de la prueba en términos de su tamaño y potencia, bajo las siguientes consideraciones:

- Para la simulación MC se usa $N = 100$ replicaciones.
- La simulación de las muestras se realiza en mallas irregulares de $(0, 50) \times (0, 50)$ con diferentes tamaños de muestra, $n, \{100, 200, 300, 400, 500\}$.
- Se utiliza $m = 100$ muestras bootstrap para aproximar la distribución de la estadística de prueba.
- Las muestras son simuladas bajo cinco esquemas de dependencia (semivariogramas) espacial: *exponencial, circular, gaussiano, esférico y cúbico*. En la sección 2.5.2 se encuentra un resumen acerca de los distintos modelos de semivariograma.
- Para estimar el tamaño de la prueba, las muestras, bajo H_0 son simuladas con los siguientes parámetros: $\beta_0 = 3, \beta_1 = 0, \beta_2 = 0, \sigma^2 = 4, \tau^2 = 1$ y $\phi = 2, 0.5, 0.01$. En este caso se fijaron los valores de $\beta_0, \beta_1, \beta_2, \sigma^2$ y τ^2 y se hizo variar el valor de ϕ . Esto es con el objetivo de estudiar el tamaño estimado de la prueba bajo diferentes niveles de correlación.

4.5. Discusión de resultados

- Para estimar la potencia de la prueba, las muestras, bajo H_1 , son simuladas con los siguientes parámetros: $\sigma^2 = 4$, $\phi = 0.01$, $\tau^2 = 1$ y $\boldsymbol{\beta} = (3, 0.25, 0.25), (3, 3, 3)$, donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$. En este caso se fijaron los valores de σ^2 , ϕ y τ^2 y se consideraron dos casos para $\boldsymbol{\beta}$. Esto es con el objetivo de estudiar el comportamiento de la potencia estimada cuando H_1 se aproxima y cuando se aleja de $H_0 : \beta_1 = \beta_2 = 0, \beta_0 > 0$.
- Los parámetros son estimados suponiendo un semivariograma exponencial usando los métodos de máxima verosimilitud (ml) y máxima verosimilitud restringida (reml). Se usó el semivariograma exponencial debido a que su forma funcional es simple y a que los métodos de máxima verosimilitud no presentan problemas en el proceso de estimación de los parámetros. Se intentó usar el semivariograma Matérn pero este presentó problemas de estimación en el bootstrap paramétrico. También se iba a intentar usar el semivariograma esférico pero [Mardia y Watkins \(1989\)](#) y [Warnes y Ripley \(1987\)](#) comentan que este modelo presenta problemas en los métodos de estimación de máxima verosimilitud, por lo cual fue descartado.
- El tamaño nominal de la prueba se fija en $\alpha = 0.05$.
- Sólomente se usaron 100 replicaciones en este estudio de simulación debido a que un mayor número de replicaciones genera un trabajo computacional de varias horas que se puede extender a días y no se dispone del hardware computacional para hacer ese trabajo.

4.5 Discusión de resultados

A continuación se presentan los resultados obtenidos del estudio de simulación y una discusión de los mismos.

4.5. Discusión de resultados

Tabla 4.1: Tamaño estimado de la prueba para Kriging Ordinario.

Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$, con $\beta_0 = 3$, $\beta_1 = 0$, $\beta_2 = 0$, $\sigma^2 = 4$, $\phi = 2$ y $\tau^2 = 1$.

$H_0 : \beta_1 = \beta_2 = 0, \beta_0 > 0$.

Los distintos modelos de semivariograma están definidos en la subsección 2.5.2.

	Tipo de Semivariograma				
	Exponencial	Gausiano	Circular	Cúbico	Esférico
<i>n</i> = 100					
ml	0.07	0.03	0.06	0.07	0.07
reml	0.09	0.07	0.05	0.05	0.09
<i>n</i> = 200					
ml	0.07	0.07	0.08	0.04	0.07
reml	0.10	0.07	0.07	0.05	0.08
<i>n</i> = 300					
ml	0.04	0.02	0.04	0.07	0.06
reml	0.08	0.08	0.08	0.08	0.08
<i>n</i> = 400					
ml	0.04	0.02	0.06	0.04	0.03
reml	0.10	0.10	0.10	0.04	0.06
<i>n</i> = 500					
ml	0.04	0.03	0.06	0.06	0.06
reml	0.10	0.08	0.08	0.03	0.08

La Tabla 4.1 muestra los resultados de tamaño estimado de la prueba para datos simulados usando los parámetros $\beta = (3, 0, 0)$, $\sigma^2 = 4$, $\phi = 2$ y $\tau^2 = 1$.

Al analizar los resultados de la tabla se pueden encontrar resultados que dan evidencia del buen desempeño de la prueba. Se pueden ver casos en donde el tamaño estimado se aproxima bastante al tamaño nominal. Como por ejemplo en el caso del semivariograma exponencial y el metodo ML se tienen tamaños estimados de 0.06 para muestras de tamaño 300, 400 y 500 los cuales son muy cercanos al tamaño nominal. Estos resultados son los esperados para este semivariograma debido a que los parámetros son estimados usandolo. Se tiene otro ejemplo al considerar el método ML y tamaños de muestra 500. En este caso se tienen tamaños estimados de 0.6 en los semivariogramas circular, cúbico y esférico.

Luego, al observar la columna correspondiente al semivariograma exponencial y el método de

4.5. Discusión de resultados

estimación ML se puede notar que existe evidencia de consistencia en el tamaño estimado de la prueba, es decir, este se aproxima al tamaño nominal cuando el tamaño de muestra aumenta. Del mismo modo, la columna del semivariograma cúbico en conjunción con el método ML proporciona evidencia de consistencia en el tamaño estimado de la prueba.

Considerando muestras de tamaño 500, el semivariograma exponencial y el método ML se puede ver que el tamaño estimado se aproxima muy bien al tamaño nominal. Lo mismo ocurre con el semivariograma circular en conjunción con REML.

Se puede notar que el semivariograma circular y el método ML juntos producen tamaños estimados muy cercanos al tamaño nominal usando muestras de tamaño iguales o mayores que 300. De manera similar el semivariograma cúbico, el método ML y tamaños de muestra iguales o mayores que 400 producen tamaños estimados muy cercanos al tamaño nominal.

Con respecto al método REML; este produce tamaños estimados iguales al tamaño nominal en el caso del semivariograma cúbico con muestras de tamaños 100 y 200. Otro caso es el del semivariograma circular y muestras de tamaño 100 en donde REML produce un tamaño estimado igual al tamaño nominal. Y al considerar muestras de tamaño 400, se tienen tamaños estimados muy cercanos al tamaño nominal al usar los semivariogramas cúbico y esférico.

Se tienen únicamente dos casos en los cuales se tienen tamaños estimados de 0.02 por el método ML. Estos se encuentran en el semivariograma gaussiano con muestras de tamaño 200 y 300. Los casos restantes del método ML son tamaños estimados mayores a 0.02. También se tiene únicamente un caso en el cual se tiene un tamaño estimado de 0.08 por el método ML. Este se encuentra en el cruce entre el semivariograma circular y $n = 200$. Los casos restantes del método ML son tamaños estimados menores que 0.08.

Un caso notable es el de muestras de tamaño 300 en conjunción con el método REML en el cual se tienen tamaños estimados de 0.08 en todos los semivariogramas. De hecho el método REML produce tamaños estimados de la prueba que llegan a 0.10 lo cual no es deseable con respecto al tamaño nominal. Algunos casos se pueden encontrar en el renglón de muestras de tamaño 400 en conjunción con los semivariogramas exponencial, gaussiano y circular.

De manera global, el método ML produce estimaciones del tamaño de la prueba en el intervalo $(0.02, 0.08)$ mientras que REML produce estimaciones en el intervalo $(0.03, 0.10)$. ML produce diferencias absolutas, del tamaño estimado con el tamaño nominal, menores o iguales a 0.03. REML produce diferencias absolutas menores o iguales a 0.05, lo cual no es deseable. Es decir, el método ML produce mejores resultados que el método REML.

4.5. Discusión de resultados

Tabla 4.2: Tamaño estimado de la prueba de Kriging Ordinario.

Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$, con $\beta_0 = 3$, $\beta_1 = 0$, $\beta_2 = 0$, $\sigma^2 = 4$, $\phi = 0.5$ y $\tau^2 = 1$.

$H_0 : \beta_1 = \beta_2 = 0, \beta_0 > 0$.

Los distintos modelos de semivariograma están definidos en la subsección 2.5.2.

	Tipo de Semivariograma				
	Exponencial	Gausiano	Circular	Cúbico	Esférico
<i>n</i> = 100					
ml	0.07	0.03	0.02	0.04	0.03
reml	0.08	0.10	0.11	0.10	0.09
<i>n</i> = 200					
ml	0.10	0.04	0.08	0.06	0.02
reml	0.08	0.06	0.10	0.08	0.09
<i>n</i> = 300					
ml	0.08	0.03	0.06	0.03	0.02
reml	0.09	0.06	0.10	0.06	0.08
<i>n</i> = 400					
ml	0.06	0.04	0.02	0.06	0.07
reml	0.08	0.08	0.10	0.02	0.08
<i>n</i> = 500					
ml	0.03	0.03	0.07	0.06	0.03
reml	0.09	0.07	0.10	0.04	0.09

La Tabla 4.2 muestra los resultados de tamaño estimado de la prueba para datos simulados usando los parámetros $\beta = (3, 0, 0)$, $\sigma^2 = 4$, $\phi = 0.5$ y $\tau^2 = 1$. En este caso se usa $\phi = 0.5$ y los parámetros restantes toman los mismos valores de la tabla 4.1. Esto es con el objetivo de estudiar el tamaño estimado de la prueba sujeto a diferentes niveles de correlación manteniendo fijos los parámetros restantes.

Se tienen casos donde el tamaño estimado se aproxima muy bien al tamaño nominal. Por ejemplo, al usar el semivariograma exponencial y el método ML se tiene un tamaño estimado de 0.06 en muestras de tamaño 400. Al usar el semivariograma cúbico y el método REML se tiene un tamaño estimado de 0.06 con muestras de tamaño 500. El método ML produce estimaciones de 0.06 y 0.04 en muestras de tamaño 400 con los semivariogramas exponencial y gaussiano, respectivamente. Considerando la columna del semivariograma gaussiano y todos los tamaños de muestra, el método ML produce estimaciones de 0.3 y 0.4 los cuales están muy cerca del tamaño

4.5. Discusión de resultados

nominal.

Se puede notar que los tamaños estimados que mejor se aproximan al tamaño nominal están dados por el método ML en la columna del semivariograma cúbico, considerando todos los tamaños de muestra. De hecho, las mejores estimaciones se obtienen mediante el semivariograma cúbico usando el método ML.

Se puede ver que el método REML es el que produce las peores estimaciones. Por ejemplo, considerando la columna correspondiente al semivariograma exponencial y todos los tamaños de muestra, los tamaños estimados por el método REML son iguales o mayores a 0.08. Otro caso en donde se tienen estimaciones iguales o mayores 0.08 se obtiene al considerar la columna correspondiente al semivariograma esférico y todos los tamaños de muestra. Luego, considerando la columna correspondiente al semivariograma circular y todos los tamaños de muestra, los tamaños estimados por REML son iguales o mayores 0.10 los cuales están lejos del tamaño nominal. De hecho, las peores estimaciones están dadas en la columna del semivariograma circular usando el método REML.

Finalmente se tiene que de manera global, el método ML presenta estimaciones en el intervalo $(0.02, 0.10)$ mientras que REML presenta estimaciones en el intervalo $(0.02, 0.11)$. ML presenta diferencias absolutas, del valor estimado con el valor nominal, menores o iguales a 0.05. REML presenta diferencias absolutas menores o iguales a 0.06. Es decir, el método ML produce mejores tamaños estimados de la prueba que el método REML.

4.5. Discusión de resultados

Tabla 4.3: Tamaño estimado de la prueba de Kriging Ordinario.

Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$, con $\beta_0 = 3$, $\beta_1 = 0$, $\beta_2 = 0$, $\sigma^2 = 4$, $\phi = 0.01$ y $\tau^2 = 1$.

$H_0 : \beta_1 = \beta_2 = 0, \beta_0 > 0$.

Los distintos modelos de semivariograma están definidos en la subsección 2.5.2.

	Tipo de Semivariograma				
	Exponencial	Gausiano	Circular	Cúbico	Esférico
<i>n</i> = 100					
ml	0.08	0.02	0.06	0.04	0.03
reml	0.08	0.10	0.10	0.10	0.02
<i>n</i> = 200					
ml	0.07	0.08	0.06	0.04	0.03
reml	0.10	0.10	0.10	0.10	0.10
<i>n</i> = 300					
ml	0.03	0.09	0.02	0.08	0.02
reml	0.09	0.08	0.10	0.02	0.08
<i>n</i> = 400					
ml	0.03	0.07	0.02	0.07	0.04
reml	0.06	0.10	0.06	0.02	0.08
<i>n</i> = 500					
ml	0.03	0.07	0.02	0.07	0.03
reml	0.08	0.10	0.04	0.02	0.10

La Tabla 4.3 muestra los resultados de tamaño estimado de la prueba para datos simulados usando los parámetros $\beta = (3, 0, 0)$, $\sigma^2 = 4$, $\phi = 0.01$ y $\tau^2 = 1$. En este caso se usó $\phi = 0.01$ y los parámetros restantes tomaron los mismos valores de las tablas 4.1 y 4.2. Esto es con el objetivo de estudiar el comportamiento de la prueba a diferentes niveles de correlación.

De manera similar como en las tablas anteriores, en este caso se tienen estimaciones que están muy cercanas al tamaño nominal de la prueba. Por ejemplo en la intersección entre la columna del semivariograma exponencial y $n = 400$, el método REML produce un tamaño estimado de 0.06. En la intersección entre la columna del semivariograma circular y tamaños de muestra 400 y 500, el método REML produce tamaños estimados de 0.06 y 0.04 respectivamente. En la intersección de la columna del semivariograma circular y tamaños de muestra 100 y 200, el método ML produce tamaños estimados de 0.06 en ambos casos. En la intersección entre la columna del semivariograma cúbico y tamaños de muestra 100 y 200, el método ML produce

4.5. Discusión de resultados

tamaños estimados de 0.04 en ambos casos.

Considerando la columna del semivariograma cúbico y los tamaños de muestra 100,...,500, se puede ver que las estimaciones que produce el método ML son muy cercanos al tamaño nominal. Se tienen las mejores estimaciones mediante este semivariograma usando ML.

Para el caso de datos simulados con el semivariograma exponencial y considerando los tamaños de muestra 100,...,500, se tiene que el método ML presenta valores estimados cercanos al tamaño nominal, los cuales son los resultados esperados para este semivariograma debido a que las muestras aleatorias son simuladas y ajustadas con este semivariograma.

Para muestras de tamaño 500 y considerando todos los semivariogramas, el método ML produce estimaciones que se encuentran en el intervalo (0.02,0.07). El método REML presenta estimaciones en el intervalo (0.02,0.10). Es decir, las estimaciones que produce ML no son mayores que 0.08, y las que produce REML llegan a 0.10. El método REML produce tamaños estimados que alcanzan 0.10 lo cual es extremo con respecto al tamaño nominal. Un caso se puede encontrar en la intersección del semivariograma gaussiano con $n = 100$. Otro caso está en la intersección del semivariograma circular y $n = 300$. Cabe destacar que esos son los peores casos que se tienen en la tabla 4.3.

Un caso notable se tiene al considerar el tamaño de muestra 200 y el método REML. Se tienen tamaños estimados de 0.10 en todos los semivariogramas

De manera global, el método ML presenta estimaciones en el intervalo (0.02,0.09) mientras que REML presenta estimaciones en el intervalo (0.02,0.10). ML presenta diferencias absolutas, del valor estimado con el valor nominal, menores o iguales a 0.04. REML presenta diferencias absolutas menores o iguales a 0.05. Es decir, el método ML presenta mejores resultados que el método REML.

4.5. Discusión de resultados

Tabla 4.4: Potencia estimada de la prueba de Kriging Ordinario.

Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$, con $\beta_0 = 3$, $\beta_1 = 3$, $\beta_2 = 3$, $\sigma^2 = 4$, $\phi = 0.01$ y $\tau^2 = 1$.

$H_1 : \beta_p \neq 0$ para algún $p = 1, 2$.

Los distintos modelos de semivariograma están definidos en la subsección 2.5.2.

	Tipo de Semivariograma				
	Exponencial	Gausiano	Circular	Cúbico	Esférico
<i>n</i> = 100					
ml	0.95	0.97	0.94	0.97	0.94
reml	0.89	0.88	0.88	0.92	0.94
<i>n</i> = 200					
ml	0.95	0.97	0.96	0.97	0.95
reml	0.96	0.94	0.90	0.92	0.95
<i>n</i> = 300					
ml	0.96	0.97	0.95	0.97	0.97
reml	0.98	0.98	0.91	0.96	0.98
<i>n</i> = 400					
ml	0.98	0.97	0.98	0.99	0.98
reml	0.98	0.98	0.96	0.96	0.99
<i>n</i> = 500					
ml	0.98	0.98	0.99	0.99	0.99
reml	0.97	0.95	0.97	0.98	0.97

La Tabla 4.4 muestra los resultados de potencia estimada de la prueba para datos simulados usando los parámetros $\beta = (3, 3, 3)$, $\sigma^2 = 4$, $\phi = 0.01$ y $\tau^2 = 1$.

Existe evidencia de consistencia en la potencia de la prueba; es decir, las potencias estimadas crecen cuando el tamaño de la muestra crece. Por ejemplo considere la columna del semivariograma exponencial y el método ML. Se puede ver que la potencia aumenta conforme aumenta el tamaño de la muestra. Otro ejemplo se tiene al considerar el semivariograma circular y el método REML. Se puede ver el aumento en la potencia estimada al aumentar el tamaño de la muestra. Este comportamiento de la potencia estimada se repite en la mayoría de los casos al observar los semivariogramas y los métodos de estimación.

Las potencias más altas producidas por REML se obtienen al considerar la columna del semivariograma esférico. En este caso se tiene que para tamaños de muestra 100, ..., 500 las potencias son

4.5. Discusión de resultados

mayores a 0.90.

Con respecto al semivariograma exponencial las potencias estimadas que produce ML con muestras de tamaños 100, ..., 500 son altas. Este resultado es el esperado debido a que las muestras son generadas y ajustadas con este semivariograma.

Nótese que para muestras de tamaño 100 el método ML produce potencias estimadas mayores a 0.90 en todos los casos de semivariogramas. Un ejemplo es el semivariograma gaussiano con el cual ML produce un tamaño de 0.97.

Las potencias estimadas más altas se tienen al considerar el semivariograma cúbico en conjunción con el método ML al usar tamaños de muestra 400 y 500. En este caso las potencias estimadas son de 0.99. De hecho, se puede notar que las potencias globalmente más altas las produce ML con muestras de tamaño 500. También cabe destacar que las potencias estimadas más bajas las produce el método REML con muestras de tamaño 100.

Considerando todos los semivariogramas y métodos de estimación, las potencias estimadas son cercanas a uno para muestras de tamaños iguales o mayores a 300.

Se puede concluir que el método ML produce potencias más altas que el método REML.

4.5. Discusión de resultados

Tabla 4.5: Potencia estimada de la prueba de Kriging Ordinario.

Estudio de simulación en una malla irregular de $(0, 50) \times (0, 50)$, con $\beta_0 = 3$, $\beta_1 = 0.25$, $\beta_2 = 0.25$, $\sigma^2 = 4$, $\phi = 0.01$ y $\tau^2 = 1$.

$H_1 : \beta_p \neq 0$ para algún $p = 1, 2$.

Los distintos modelos de semivariograma están definidos en la subsección 2.5.2.

	Tipo de Semivariograma				
	Exponencial	Gausiano	Circular	Cúbico	Esférico
<i>n</i> = 100					
ml	0.34	0.36	0.32	0.43	0.30
reml	0.32	0.25	0.34	0.30	0.29
<i>n</i> = 200					
ml	0.36	0.39	0.44	0.43	0.53
reml	0.38	0.34	0.36	0.38	0.34
<i>n</i> = 300					
ml	0.40	0.38	0.42	0.35	0.42
reml	0.44	0.35	0.42	0.42	0.36
<i>n</i> = 400					
ml	0.44	0.46	0.40	0.55	0.40
reml	0.59	0.51	0.50	0.48	0.27
<i>n</i> = 500					
ml	0.48	0.55	0.50	0.51	0.50
reml	0.46	0.46	0.46	0.53	0.50

La Tabla 4.5 muestra los resultados de potencia estimada de la prueba para datos simulados usando los parámetros $\beta = (3, 0.25, 0.25)$, $\sigma^2 = 4$, $\phi = 0.01$ y $\tau^2 = 1$. Nótese que se usaron los valores $\beta_1 = 0.25$, $\beta_2 = 0.25$ y los parámetros restantes tomaron los mismos valores de la tabla 4.4. En este caso, se dice que la hipótesis alternativa esta muy cerca de la hipótesis nula.

Las potencias crecen cuando el tamaño de muestra crece en todos los semivariogramas y métodos de estimación. Es decir, se tiene evidencia de consistencia en la potencia estimada de la prueba. Como ejemplo, considere la columna del semivariograma gaussiano y el método ML. Se puede ver que la potencia estimada aumenta cuando el tamaño de la muestra aumenta. Otro ejemplo se tiene al considerar la columna del semivariograma circular y el método REML. La potencia estimada crece con el aumento en el tamaño de la muestra.

Con respecto al semivariograma exponencial las potencias estimadas que produce ML con mues-

4.6. Una aplicación

tras de tamaños 100,...,500 son aceptables con respecto al tamaño nominal. Este resultado es el esperado debido a que H_1 está muy cerca de H_0 aunque las muestras hayan sido generadas y ajustadas con este semivariograma.

Las potencias estimadas más altas las produce el método ML con muestras de tamaño 500. En este caso la potencia estimada más alta es de 0.55, la cual es también la más alta de manera global. Las potencias más bajas se encuentran al considerar el renglón del método REML y tamaños de muestra 100. En este caso la potencia estimada más baja es de 0.25, la cual es la más baja de manera global.

Se puede ver que las potencias no son altas al usar muestras de tamaño 500 lo cual es debido a que los parámetros β_1 y β_2 son cercanos a cero; es decir, H_1 está muy cerca de H_0 .

Se debe de aclarar que las potencias estimadas en todos los casos de semivariogramas, métodos de estimación y tamaños de muestra, son suficientemente significativas debido a que, a pesar de no ser altas, son mucho mayores que el tamaño nominal.

Finalmente, después de haber hecho el análisis de resultados del tamaño estimado y potencia estimada de la prueba, se puede concluir que el desempeño de la prueba es adecuado ya que aproxima muy bien al tamaño nominal sujeta a diferentes niveles de correlación y las potencias obtenidas son suficientemente significativas (con respecto al tamaño nominal). Entonces, la metodología propuesta es adecuada para ser usada en aplicaciones.

4.6 Una aplicación

En el ejemplo de la sección 3.7 se probó la hipótesis de que los datos son una realización de un campo aleatorio gaussiano. Ahora se va a probar si la media (o tendencia) del campo aleatorio es constante, o dependiente de las coordenadas espaciales.

Primeramente se va a hacer un análisis exploratorio, ajustando semivariogramas empíricos, con tendencias constante y de primer orden, a los datos. Después se van a ajustar semivariogramas exponenciales, con tendencias constante y de primer orden, a los semivariogramas empíricos. Esto es con el propósito de analizar gráficamente si el ajuste con tendencia constante es mejor que el ajuste con tendencia de primer orden.

4.6. Una aplicación

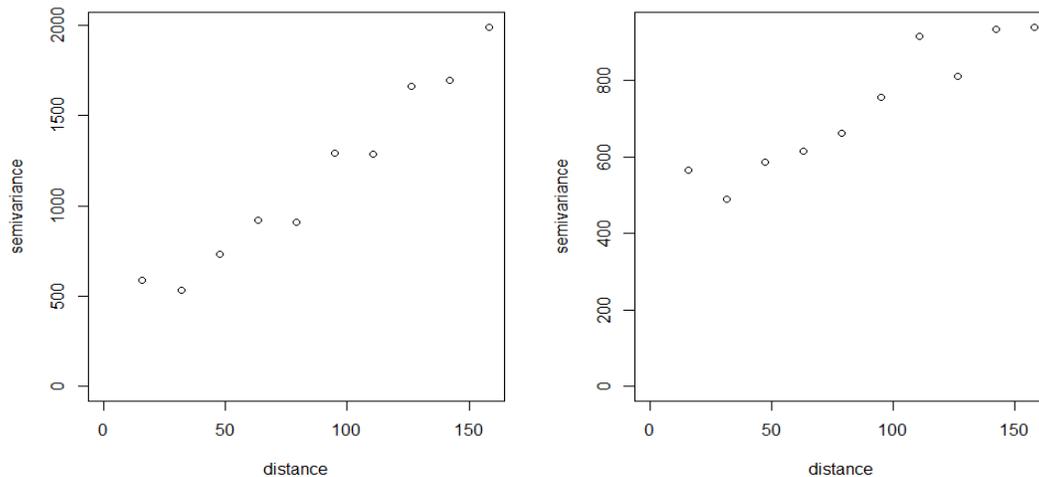


Figure 4.1: Semivariograma empírico, de tipo Cressie-Hawkins, ajustado a los datos de precipitación del estado de Paraná, Brasil. El semivariograma de la izquierda fue ajustado usando una tendencia constante y el de la derecha, usando una tendencia de primer orden.

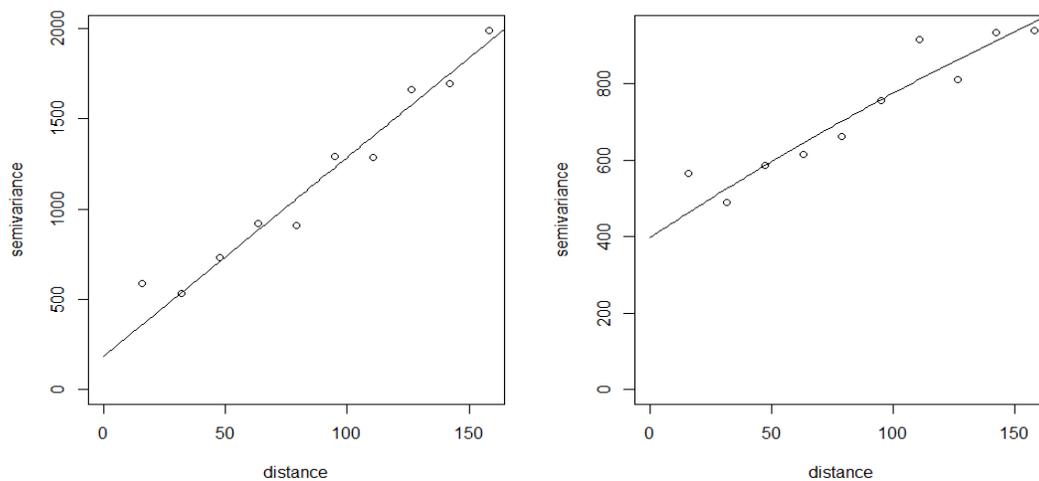


Figure 4.2: Semivariograma teórico (exponencial), ajustado al semivariograma empírico. El semivariograma de la izquierda fue ajustado usando una tendencia constante y el de la derecha, usando una tendencia de primer orden.

En la figura 4.2 se puede ver que el semivariograma teórico con tendencia constante es el que mejor se ajusta a su correspondiente semivariograma empírico. Por lo tanto, se puede esperar el

4.6. Una aplicación

no rechazo de la hipótesis nula.

Al aplicar la metodología para probar la media constante del Campo Aleatorio se tienen los siguientes resultados:

$$\begin{aligned}\tilde{F}_0 &= 7.084844 \\ F_{0.05} &= 13.88685\end{aligned}$$

Como $\tilde{F}_0 < F_{0.05}$, no se rechaza la hipótesis nula; es decir, los datos tienen media constante, por lo tanto no se rechaza el modelo de Kriging Ordinario para estos datos.

Capítulo 5

CONCLUSIONES

En este trabajo de investigación se proponen dos pruebas estadísticas para los cuales se realizaron estudios de simulación para estudiar sus propiedades estadísticas de tamaño y potencia. La primera es una prueba para la hipótesis de un campo aleatorio gaussiano. La segunda es una prueba para la media constante de un campo aleatorio gaussiano.

La estimación de los parámetros, en ambas pruebas, fue hecha por los métodos de máxima verosimilitud y máxima verosimilitud restringida. Algunos autores, entre ellos [Cressie \(1993\)](#) mencionan que los estimadores de máxima verosimilitud son sesgados. Sin embargo, en los resultados de simulación se vió que la sesgadez no afecta las estimaciones de tamaño y potencia de la prueba, debido a que el tamaño estimado se aproxima al tamaño nominal cuando el tamaño de muestra aumenta, y lo mismo ocurre con la potencia de ambas pruebas.

En la prueba para el campo aleatorio gaussiano fue usado el semivariograma Matérn para modelar la dependencia espacial. Esto fue conveniente debido a que este semivariograma consta de cuatro parámetros, los cuales abarcan un espacio paramétrico amplio. Es decir, usando este semivariograma se puede modelar, de manera general, la dependencia espacial de casi cualquier campo aleatorio. También se propuso un método para estimar los cuatro parámetros de este modelo. Los resultados del estudio de simulación muestran la consistencia de este método.

En la prueba para la media del campo aleatorio, fueron simulados datos con media constante y no constante, y se consideraron semivariogramas exponenciales, gaussianos, circulares, cúbicos y esféricos. Los resultados de simulación muestran que estos semivariogramas no influyen, en la prueba, para decidir si la media del campo aleatorio es constante. Para estimar los parámetros del modelo, se optó por usar el semivariograma Exponencial, el cual mostró un buen desempeño en

5. CONCLUSIONES

la estimación del tamaño y la potencia de la prueba.

En las dos pruebas ya mencionadas fueron considerados diferentes niveles de correlación con el objetivo de estudiar el desempeño de ambas pruebas sujetas a estos. Los resultados mostraron que el nivel de correlación no influye de manera significativa en la estimación del tamaño y la potencia de ambas pruebas. En el caso de la prueba para la media del campo aleatorio gaussiano fueron considerados diferentes valores de los elementos de β , con el fin de estudiar su influencia en la estimación de la potencia. Los resultados muestran que para valores de β_1 y β_2 cercanos a cero la potencia no es alta, pero es suficientemente mayor al tamaño nominal de la prueba.

En la prueba para el campo aleatorio gaussiano el método de máxima verosimilitud restringida, para estimar los parámetros, mostró un mejor desempeño, que el método de máxima verosimilitud, en la aproximación del tamaño nominal de la prueba. Y en la prueba para la media del campo aleatorio gaussiano el método de máxima verosimilitud mostró un mejor desempeño que el método de máxima verosimilitud restringida.

REFERENCIAS

- Anderson, T. W. y Darling, D. A. (1954). A Test of Goodness-of-Fit. *Journal of the American Statistical Association*, 49, 765-769.
- Balakrishnan, N., Ma, C. y Wang, R. (2015). Logistic vector random fields with logistic direct and cross covariances. *Journal of Statistical Planning and Inference*, 161, 109-118.
- Banerjee, S., Bradley, P. C. y Alan, E. G. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.
- Barry, R. P. (1996). A diagnostic to assess the fit of a variogram model to spatial data. *Journal of Statistical Software*, 1(1), 1-11.
- Casellas, J., Tarrés, J., Piedrafita, J. y Varona, L. (2006). Análisis del ajuste de los modelos de riesgos proporcionales mediante bootstrap paramétrico. *Información técnica económica agraria*.
- Clark, R. y Allingham, S. (2011). Robust resampling confidence intervals for empirical variograms. *Mathematical Geosciences*, 43(2), 243-259.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons. Revised edition.
- Diggle, P. J. y Ribeiro, J. P. J. (2001). geoR: a package for geostatistical analysis. *R-NEWS*, 1, 2, 15–18. ISSN 1609-3631.
- Diggle, P. J. y Ribeiro, P. J. J. (2007). *Model-based Geostatistics*. Springer Science+Business Media, LLC.
- Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics.*, 7, 1-26.
- Furrer, H. (2003). *The Term Structure of Interest Rates as a Random Field. Applications to Credit Risk*. Master thesis, ETH ZURICH and UNIVERSITY OF ZURICH.
- Gaetan, C. y Guyon, X. (2010). *Spatial Statistics and Modeling*. Springer Science+Business Media, LLC.
- Guzmán, M. M., Villaseñor, A. J. A. y Gonzalez, E. E. (2015). Two statistical tests for the semivariogram function of spatial Gaussian processes. *Journal of Applied Statistics*.

REFERENCIAS

- Handcock, M. S. y Wallis, J. R. (1994). An approach to statistical spatial temporal modeling of meteorological fields (with discussion). *Journal of the American Statistical Association*, 89, 368-390.
- Harville, D. A. y Jeske, D. R. (1992). Mean squared error of estimation or Prediction under a general linear model. *Journal of the American Statistical Association*, 87, 724-731.
- Heagerty, P. y Lele, S. R. (1998). A Composite Likelihood Approach to Binary Spatial Data. *Journal of the American Statistical Association*, 93, 1099-1111.
- Henze, N. (2002). Invariant tests for multivariate normality: a critical review. *Statist.Pap.*, 43, 767-506.
- Hwu, T. J., Han, C. P. y Rogers, K. J. (2002). The combination test for multivariate normality. *Statist. Computat. Simul.*, 72(5), 379-390.
- Kackar, R. N. y Harville, D. A. (1990). Approximation for Standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79, 853-862.
- Kitanidis, P. K. (1983). Statistical estimation of polynomial generalized covariance functions and hydrological applications. 22, 499-507.
- Kotz, S., Balakrishnan, N. y Johnson, N. L. (2000). *Continuous Multivariate Distributions*, tomo 1. John Wiley & Sons, Inc., segunda edición.
- Krige, D. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of Chemical, Metallurgical and Mining Society of South Africa, Association*, 52, 119-139.
- Malik, H. J. y Abraham, B. (1973). Multivariate Logistic Distributions. *Ann. Statist.*, 1, 588-590.
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57, 519-530.
- Mardia, K. V. y Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71, 135-46.
- Mardia, K. V. y Watkins, A. J. (1989). On multimodality of the likelihood in the spatial linear model. *Biometrika*, 76, 289-296.
- Matérn, B. (1986). *Spatial Variation*. Springer-Verlag, New York, segunda edición. Lecture Notes in Statistics.
- Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58, 1246-1266.
- Mecklin, C. J. y Mundfrom, D. J. (2005). A Monte Carlo comparison of the Type I and II error rates of tests of multivariate normality. *Statist. Computat. Simul.*, 75(2), 93-107.
- Miller, R. G. (1974). The jackknife - a review. *Biometrika*, 61, 1-15.

REFERENCIAS

- Oliver, M. A. y Webster, R. (2014). A tutorial guide to geostatistics: Computing and modeling variograms and kriging. *Catena*, 113, 56-69.
- Opitz, T. (2013). Extremal t processes: Elliptical domain of attraction and a spectral representation. *arXiv:1207.2296v6*.
- Padoan, S. A. y Moreno, B. (2015). Analysis of Random Fields Using CompRandFld. *Journal of Statistical Software*, 63.
- Patterson, H. D. y Thompson, R. (1971). *Recovery of inter-block information when block sizes are unequal*, tomo 58.
- Prasad, N. G. N. y Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 161-171.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Royston, J. P. (1982). An extension of Shapiro and Wilk W test for normality to large samples. *Appl. Statist.*, 31(2), 115-124.
- Royston, P. (1983). Approximating the Shapiro-Wilk W test for non-normality. *Statist. Comput.*, 2, 117-119.
- Samper, C. F. J. y Carrera, R. J. (1990). *Geoestadística, aplicaciones a la hidrología subterránea*. Centro Internacional de métodos numéricos en Ingeniería. Universidad Politécnica de Catalunya. Barcelona.
- Schabenberger, O. y Gotway, C. (2005). *Statistical Methods for spatial data analysis*. Chapman & Hall /CRC.
- Solanas, A. y Sierra, V. (1992). Bootstrap: fundamentos e introducción a sus aplicaciones. *Anuario de Psicología*, 55, 143-154.
- Srivastava, D. K. y Mudholkar, G. S. (2003). Goodness of fit tests for univariate and multivariate normal models. In: Khattree, R., Rao, C. R., eds. *Handbook of statistics 22: Statistics in Industry*, North Holland: Elsevier, 869-906.
- Stephens, M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association*, 69, 730-737.
- Stephens, M. A. (1976). Asymptotic Results for Goodness-of-Fit Statistics with Unknown Parameters. *Annals of Statistics*, 4, 357-369.
- Stephens, M. A. (1977). Goodness of Fit for the Extreme Value Distribution. *Biometrika*, 64, 583-588.

REFERENCIAS

- Stephens, M. A. (1979). Tests of Fit for the Logistic Distribution Based on the Empirical Distribution Function. *Biometrika*, 66, 591-595.
- Stoyan, D., Kendall, W. y Mecke, J. (1995). *Stochastic Geometry and its Applications*. John Wiley & Sons, New York. Second edition.
- Székely, G. J. y Rizzo, M. L. (2005). A new test for multivariate normality. *Multivariate Anal.*, 93(1), 58-80.
- Thode, H. C. (2002). *Testing for normality*. New York: Marcel & Dekker Inc.
- Villaseñor, A. J. A. y Gonzalez, E. E. (2009). A generalization of Shapiro-Wilk's test for multivariate normality. *Communications in Statistics - Theory and Methods.*, 38, 1870-1883.
- Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54, 426-482.
- Warnes, J. J. y Ripley, B. D. (1987). Problems with likelihood estimation of covariance functions of spatial gaussian processes. *Biometrika*, 74, 640-642.
- Zimmerman, D. L. (1989). Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. 21, 655-672.

APÉNDICES

A. Conceptos Teóricos

A1. Bootstrap Paramétrico

El método de Bootstrap fue descrito de forma sistemática por [Efron \(1979\)](#).

En ocasiones, el investigador conoce la función de distribución correspondiente a la variable aleatoria objeto de estudio, aunque se desconozcan los parámetros de la misma, punto que conduce al Bootstrap paramétrico. Suponga que $\hat{\theta} = \hat{\theta}(x)$ es un estimador de θ , siendo éste un parámetro o vector de parámetros de F , por lo cual se puede expresar $F = F_{\theta}$. Usando la función de distribución F , obtenga $\hat{\theta}$ a partir de datos muestrales. Estimado el parámetro o vector de parámetros, se puede recurrir a la estimación $\hat{F} = F_{\hat{\theta}}$ ([Solanas y Sierra, 1992](#)).

El bootstrap paramétrico se usa de manera rutinaria para aproximar la distribución de cualquier estadístico de interés (π) ([Casellas *et al.*, 2006](#)). Este procedimiento consiste en tres pasos característicos: *a*) definición del modelo asumido para los datos, *b*) re-muestreo mediante simulación de Monte Carlo y cálculo del estadístico π , y *c*) construcción de la distribución de bootstrap de π .

A2. La prueba de Anderson Darling

La prueba de Anderson-Darling ([Anderson y Darling, 1954](#)) se utiliza para probar si una muestra de datos proviene de una población con una distribución específica ([Stephens, 1974](#)). Se trata de una modificación de la prueba de Kolmogorov-Smirnov (K-S) y le da más peso a las colas que la prueba de K-S. La prueba de Anderson-Darling (AD) hace uso de una distribución específica en el cálculo de los valores críticos. Esto tiene la ventaja de permitir una prueba sensible y la desventaja de que los valores críticos deben calcularse para cada distribución. Actualmente, existen tablas de valores críticos para las distribuciones normal,

APÉNDICES

uniforme, lognormal, exponencial, Weibull, de valores extremos tipo I, Pareto generalizada y logística. No se proporcionan las tablas de valores críticos en este documento. Para mayor información consulte [Stephens \(1974\)](#), [Stephens \(1976\)](#), [Stephens \(1977\)](#) y [Stephens \(1979\)](#) ya que esta prueba se aplica por lo general con un programa de software estadístico que va a imprimir los valores críticos pertinentes.

La prueba de Anderson-Darling se define como:

$$\begin{aligned}H_0 &: \text{Los datos siguen una distribución especificada} \\H_1 &: \text{Los datos no siguen una distribución especificada}\end{aligned}$$

La estadística de prueba se define como,

$$A^2 = -N - S$$

Donde:

$$-S = \sum_{i=1}^N \frac{(2i-1)}{N} \{ \ln F(y_i) + \ln(1 - F(Y_{n+1-i})) \}.$$

- F es la función de distribución acumulada de la distribución especificada. Nótese que los Y_i son los datos ordenados.

Los valores críticos para la prueba de Anderson-Darling dependen de la distribución específica que se está probando. Existen valores tabulados para algunas distribuciones específicas (normal, lognormal, exponencial, Weibull, logística, valores extremos tipo 1) ([Stephens, 1974, 1976, 1977, 1979](#)). Esta prueba es de una cola y la hipótesis de que la distribución es de una forma específica se rechaza si la estadística de prueba, A , es mayor que el valor crítico. Tenga en cuenta que para una distribución dada, la estadística de Anderson-Darling puede multiplicarse por una constante (que por lo general depende del tamaño de la muestra, n). Estas constantes están dadas en los distintos trabajos de Stephens.

Muchas pruebas y procedimientos estadísticos se basan en supuestos de distribución específicos. Por ejemplo, el supuesto de normalidad es particularmente común en las pruebas estadísticas clásicas, en confiabilidad se tiene el supuesto de que los datos siguen una distribución Weibull.

Existen varias técnicas no paramétricas y otras robustas que no hacen supuestos distribucionales fuertes. Sin embargo, las técnicas basadas en supuestos de distribución específicas son en general más potentes que las técnicas no paramétricas y las robustas. Por lo tanto, si los supuestos de distribución pueden ser validados, entonces estas técnicas reciben preferencia en su uso.

A3. Una breve revisión acerca de campos aleatorios no gaussianos

Para estimar la potencia de la prueba en el capítulo 3 se simularon distintos modelos de campos aleatorios no gaussianos con tamaños de muestra 100, 200, 300, 400 y 500. A las muestras simuladas se les aplicó el algoritmo que estima el tamaño de la prueba, usando un tamaño nominal de $\alpha = 0.05$. Es decir, se contó el número de rechazos usando muestras provenientes de H_1 .

Fueron simulados datos provenientes de los siguientes campos aleatorios no gaussianos:

1. Logístico de tipo Gumbel-Malik-Abraham
2. Logístico
3. Poisson
4. Binario
5. T
6. Gumbel
7. Chi-Cuadrada

Se utilizaron los paquetes `CompRandFLd` y `RandomFields`, disponibles en el software estadístico R ([R Core Team, 2013](#)), para simular los distintos campos aleatorios. El uso de estos paquetes es de la siguiente forma:

1. Se especifica un modelo para la matriz de varianza-covarianza del campo aleatorio mediante el comando *model*. Además de los modelos de covarianza ya conocidos, se tienen modelos como el bernoulli, el Chi-cuadrada, el poisson, el brownresnick, el opitz, el schlater, el smith, etc.
2. Usando el modelo de covarianza en 1, se ejecuta el comando *RFsim* (paquete `CompRandFLd`) o *RFsimulate* (paquete `RandomFields`) para obtener realizaciones del campo aleatorio.

A continuación se presenta una revisión de campos aleatorios no gaussianos que se usaron en este trabajo.

Campo aleatorio Logístico de tipo Gumbel-Malik-Abraham

Para la construcción de esta subsección se recurrió a los resultados del libro de [Kotz et al. \(2000\)](#).

Sean U , V y W variables aleatorias (v.a.'s) independientes e idénticamente distribuidas (iid) con distribución de valores extremos con densidad:

$$p_U(u) = e^{-u-e^{-u}}, \quad -\infty < u < \infty \quad (.1)$$

y función generadora de momentos,

$$M(t) = \Gamma(1-t), |t| < 1$$

Considere la transformación,

$$X = V - U \text{ y } Y = W - U$$

Por cambio de variables, se puede obtener la densidad conjunta de X y Y , que resulta ser logística bivariada estandar.

No es difícil generalizar el resultado anterior al caso k -dimensional.

Sean U_0, U_1, \dots, U_k variables aleatorias independientes con distribución de valores extremos con densidad común como en (.1). Las v.a.'s,

$$X_i = U_i - U_0$$

tienen distribución conjunta logística k -variada estandar, con función de distribución,

$$F_{\mathbf{X}}(\mathbf{x}) = \frac{1}{1 + \sum_{i=1}^k e^{-x_i}}, -\infty < x_i < \infty$$

A $F_{\mathbf{X}}(\mathbf{x})$ se le conoce como la Distribución Logística k -variada de tipo Gumbel-Malik-Abraham cuyas propiedades son,

$$\text{Var}(X_i) = \frac{\pi^2}{3} \text{ para todo } i,$$

$$\text{Cov}(X_i, X_j) = \frac{\pi^2}{6},$$

$$\text{Corr}(X_i, X_j) = \frac{1}{2}$$

Se simularon 100 muestras de un campo aleatorio logístico de tipo Gumbel-Malik-Abraham y se aplicó el procedimiento para determinar la potencia de la prueba.

Campo aleatorio Logístico

Por el teorema de existencia de Kolmogorov se puede obtener un campo aleatorio escalar $\{Z(x), x \in D\}$ con distribuciones finito-dimensionales de tipo Gumbel-Malik-Abraham ([Malik y Abraham, 1973](#)) de la forma,

$$P(Z(x_1) \leq z_1, \dots, Z(x_n) \leq z_n) = \left\{ 1 + \sum_{k=1}^n \exp\left(-\frac{z_k - \mu(x_k)}{\sigma(x_k)}\right) \right\}^{-1}, \mathbf{z} \in \mathbb{R}^n,$$

donde $\mu(x)$ y $\sigma(x)$, $x \in D$, son funciones reales con $\sigma(x) > 0$, n es un numero natural arbitrario, y $x_k \in D$ ($k = 1, \dots, n$).

Alternativamente, este campo aleatorio puede ser formulado como,

$$Z(x) = \{Y(x) - Y_0\} \sigma(x) + \mu(x), \quad x \in D,$$

donde $\{Y(x), x \in D\}$ es un conjunto de v.a.'s iid con distribucion de valores extremos con densidad $\exp\{-y - \exp(-y)\}$, $y \in \mathbb{R}$, Y_0 es una v.a. con la misma densidad, Y_0 y $\{Y(x), x \in D\}$ son independientes.

Se simularon 100 realizaciones de un campo aleatorio logistico con $\sigma(x) = 3$, $\mu(x) = 2$ y se aplicó el método para determinar la potencia de la prueba.

Campo aleatorio T

Representacion espectral de un proceso extremo tipo T

Suponga que estan dadas las siguientes condiciones,

- Un indice de la cola $\alpha > 0$ y una funcion de correlacion Cov^* ,
- replicasiones W_i , que son iid, de un campo aleatorio gaussiano W en $X \subset \mathbb{R}^d$ con funcion de dispersion $Cov = Cov^*$ y
- un proceso poisson $\{V_i\} \sim PRM(\alpha v^{-(\alpha+1)} dv)$ en $(0, \infty)$.

El proceso definido por,

$$\mathbf{Z} = \{Z(x)\} = \left\{ m_\alpha^{-\alpha-1} \max_{i=1,2,\dots} V_i W_i(x) \right\} \quad (x \in X),$$

con $m_\alpha = \sqrt{\pi}^{-1} 2^{0.5(\alpha-2)} \Gamma(0.5(\alpha+1))$ y $\Gamma(\cdot)$ la funcion gamma, es un proceso extremo t con distribuciones marginales $\alpha - Fréchet$. Su estructura de dependencia esta caracterizada por α grados de libertad general y la funcion de correlacion Cov^* .

Se simularon 100 realizaciones de un campo aleatorio T usando una funcion de covarianza exponencial ($C(r) = e^{-r}$) y se aplicó el método para determinar la potencia de la prueba.

Campo aleatorio Chi-cuadrada

Sean Z_1, \dots, Z_n campos aleatorios gaussianos estacionarios independientes con funcion de media $m(\mathbf{x}) = E[Z_i(\mathbf{x})] = 0$, $i = 1, \dots, n$, funcion de varianza comun $R(\mathbf{y}) = E[Z(\mathbf{y})Z(\mathbf{x} + \mathbf{y})]$ y varianza $\sigma^2 = R(\mathbf{0})$. Para $\mathbf{x} \in R^d$, el proceso

$$Y(\mathbf{x}) = Z_1^2(\mathbf{x}) + \dots + Z_n^2(\mathbf{x})$$

es un campo Chi-cuadrada con parametro n .

Para obtener la funcion de covarianza, R^* , del campo, asuma que $\sigma^2 = R(\mathbf{x}, \mathbf{x}) = 1$. Por independencia de los campos Z_i, Z_k , $i \neq k$, obtenemos

$$R^*(\mathbf{x}, \mathbf{y}) = 2nR^2(\mathbf{x}, \mathbf{y})$$

donde R denota la función de covarianza común de los campos Z_i .

Se simularon 100 realizaciones de un campo aleatorio Chi-cuadrada usando una funcion de covarianza exponencial ($C(r) = e^{-r}$) y se aplicó el procedimiento para determinar la potencia de la prueba.

Campo aleatorio Poisson

Un proceso, se dice que es de tipo Poisson si cumple con las siguientes dos propiedades:

(i) Si $N(A)$ denota el numero de eventos en la subregión $A \subset D$, entonces $N(A) \sim Poisson(\lambda \nu(A))$, donde $0 < \lambda < \infty$ denota la funcion de intensidad constante del proceso, (ii) Si A_1 y A_2 son dos subregiones disjuntas de D , entonces $N(A_1)$ y $N(A_2)$ son independientes.

[Stoyan et al. \(1995\)](#) llaman a (ii) la propiedad de “Aleatoriedad Completa”. Nótese que la propiedad (ii) proviene de (i) pero el inverso no es cierto. El número de eventos en A se puede distribuir como una variable Poisson con intensidad espacialmente variable pero los eventos pueden permanecer independientes en conjuntos disjuntos. Se considera a (i) y (ii) como la definición de aleatoriedad espacial completa.

Un proceso puntual que satisface (i) y (ii) es llamado un Proceso Poisson Homogeneo (PPH). Si la funcion de intensidad $\lambda(\mathbf{s})$ varia espacialmente, la propiedad (i) no es conocida, pero la (ii) si. Un proceso de este tipo es llamado Proceso Poisson No Homogeneo (PPNH) y está caracterizado por las siguientes propiedades.

(i) Si $N(A)$ denota el número de ventos en la subregión $A \subset D$, entonces $N(A) \sim \text{Poisson}(\lambda(A))$, donde $0 < \lambda(\mathbf{s}) < \infty$ es la intensidad en la ubicación \mathbf{s} y $\lambda(A) = \int_A \lambda(\mathbf{s}) d\mathbf{s}$; (ii) Si A_1 y A_2 son dos subregiones disjuntas de D , entonces $N(A_1)$ y $N(A_2)$ son independientes.

El PPH (intensidad constante) es, obviamente, un caso particular del PPNH. [Stoyan et al. \(1995\)](#) se refieren al PPH como el proceso Poisson estacionario y etiquetan al PPNH como el proceso Poisson general.

Se simularon 100 realizaciones de un campo aleatorio Poisson, basado en la funcion de densidad normal (truncada) y con funcion de intensidad 30. Se aplicó el algoritmo para determinar la potencia de la prueba.

Campo aleatorio Binario

Un enfoque para la definición de un campo aleatorio binario espacio-temporal es el siguiente ([Heagerty y Lele, 1998](#)).

Sea $\{Y(\mathbf{s}, t), (\mathbf{s}, t) \in I\}$ un campo aleatorio gaussiano espacio-temporal de segundo orden estacionario con función de correlación $\rho(\mathbf{h}, u)$ que asumimos conocidos hasta un vector de parámetros ϑ . Un campo aleatorio espacio-temporal binario está definido como,

$$Z(\mathbf{s}, t) = 1I_{b, +\infty}(Y(\mathbf{s}, t) + \varepsilon(\mathbf{s}, t))$$

Donde para todos $(\mathbf{s}, t) \in I$, $\varepsilon(\mathbf{s}, t) \sim N(0, \tau^2)$ es ruido blanco y $b \in \mathbb{R}$ es un umbral. Supongamos que ε es independiente de Y . Con esta definición, para todo $(\mathbf{s}, t) \in I$, la probabilidad marginal univariada de Z es,

$$p_1 \equiv pr(Z(\mathbf{s}, t) = 1) = \Phi\left(\frac{\mu - b}{\omega}\right)$$

Donde μ es la media de Z y Φ es la distribución gaussiana estándar univariada. Además, para cualquier $(\mathbf{s}, t), (\mathbf{s}', t') \in I$ con $(\mathbf{s}, t) \neq (\mathbf{s}', t')$ las probabilidades bivariadas conjuntas son iguales a,

$$\begin{aligned} p_{11}(\mathbf{h}, u) &\equiv \Phi_2\left(\frac{\mu - b}{\omega}, \frac{\mu - b}{\omega}, \frac{\sigma^2 \rho(\mathbf{h}, u; \vartheta)}{\omega^2}\right) \\ p_{00}(\mathbf{h}, u) &= 1 - 2p_1 + p_{11}(\mathbf{h}, u), \\ p_{00}(\mathbf{h}, u) &\equiv p_{01}(\mathbf{h}, u) = p_1 - p_{11}(\mathbf{h}, u) \end{aligned}$$

donde Φ_2 es la distribución Gaussiana estándar bivariada. Note que la marginal y la probabilidad conjunta de éxito p_1 y $p_{11}(\mathbf{h}, u)$ caracterizan completamente la distribución bivariada. Probabilidades conjuntas de orden superior se derivan de manera similar, utilizando el principio de inclusión-exclusión.

Se simularon 100 realizaciones de un campo aleatorio binario basado en un campo aleatorio gaussiano latente con funcion de correlación exponencial. Se aplicó el algoritmo para determinar la potencia de la prueba.

Apéndice A: Programas de R Usados en el Trabajo

En todos los casos se usa el programa R. Se debe de instalar los paquetes necesarios, los cuales pueden obtenerse gratuitamente en la página web del proyecto R, <http://www.r-project.org>. Las rutinas que se presentan funcionan en R-3.0.0 ó superior.

Programa Monte Carlo para estimar el tamaño de la prueba para la hipótesis de Campo Aleatorio Gaussiano (3).

```
#### Monte Carlo #####

rm(list=ls())
library(spdep)
library(geoR)
library(MASS)
library(snowfall)
library(nortest)
cov.mod<-"exponential"
CMH1<- "matern"
betaKr<- rbind(3,0,0) # Vector que determina el tipo de media, ó tendencia #
n=100 # Tamano de muestra, numero de sitios #
m=100 # Numero de muestras Monte Carlo #
alfa=0.05
rech= as.numeric(0)
sigmac<- 4
fhi<- 3
nugg<-2
kapa=0
if(cov.mod=="matern"){kapa=1}
gridt<- matrix(NA,nrow=n,ncol=2)
Zs<- matrix(NA,nrow=n,ncol=1)
unos<- matrix(1,nrow=n, ncol=1)
k=0
sfInit(parallel=TRUE, cpus=4)
sfLibrary(snowfall)
sfLibrary(MASS)
sfLibrary(geoR)
sfLibrary(nortest)
```

APÉNDICES

```
sfExportAll( )
vswt<- function(b){
  k2=0
  while (k2==0) {
    can<- grf(n, grid = "irreg", xlims = c(0, 50), ylims = c(0, 50), nsim = 1,
    cov.model= cov.mod,cov.pars = c(sigmac,fhi), kappa= kapa, nugget =nugg,
    lambda = 0, mean = matrix(0,nrow=n,ncol=1), method="cholesky", RF=TRUE,
    messages=FALSE)
    gridt<- as.matrix(can$coords)
    gridf<- cbind(unos,gridt)           # Se crea la matriz Xs (de tres columnas) #
    mediaKr<-gridf%*%betaKr           # Media, ó Tendencia del proceso #
    Zs<- mediaKr+can$data              # Vector de datos Z de s #
    cmd<- cbind(gridt,Zs)              # Matriz de sitios y datos #
    geod<- as.geodata(cmd, coords.col=1:2, data.col=3)
    vemp<-variog(geod, option="bin",trend="1st", estimator.type="modulus",
    uvec=seq(0,sqrt(2*10*10)/2,l=11), pairs.min=25, nugget.tolerance=0)
    reg<-lm(vemp$v~vemp$u)
    ini.rmat<-mean(vemp$v)
    ini.pep=reg$coefficients[1]
    ini.sig<-var(Zs)-ini.pep
    mk=10
    conj=5
    met1<-matrix(NA,nrow=mk,ncol=8)
    met1p<-matrix(NA,nrow=mk,ncol=8)
    delta=conj/mk
    ki=0
    for (i in 1:mk){
      ki=delta+ki
      phii=ini.rmat/(2*sqrt(ki))
      sigml<-likfit(geod,trend="1st", ini.cov.pars = c(ini.sig,phii), fix.nugget=F,
      nugget=ini.pep, cov.model=CMH1, fix.kappa= T,kappa=ki, lik.method="ML",
      components=F,messages=F)
      met1p[i,1]=sigml$beta[1]; met1p[i,2]=sigml$beta[2]; met1p[i,3]=sigml$beta[3];
      met1p[i,4]=sigml$sigmasq; met1p[i,5]=sigml$phi; met1p[i,6]=sigml$tau;
      met1p[i,7]=ki; met1p[i,8]=sigml$loglik
    }
    met1po<-met1p[order(met1p[,8]),]
    beta0=met1po[1,1]
    beta1=met1po[1,2]
    beta2=met1po[1,3]
```

```

sigmasq0=met1po[1,4]
phi0=met1po[1,5]
nugg0=met1po[1,6]
kappa0=met1po[1,7]
emv<-cbind(beta0,beta1,beta2,sigmasq0,phi0,nugg0,kappa0)
mediaEV<- gridf%*(rbind(beta0,beta1,beta2))
sigma.mat<- varcov.spatial(coords = gridt, dists.lowertri = NULL,
                           cov.model = CMH1, nugget = nugg0,
                           cov.pars = c(sigmasq0,phi0),kappa=kappa0,
                           inv = FALSE, det = FALSE,
                           func.inv = c("cholesky"),
                           scaled = FALSE, only.decomposition = FALSE,
                           sqrt.inv = FALSE, try.another.decomposition=T,
                           only.inv.lower.diag = F)

sig=sigma.mat[[1]]
decomp<-eigen(sig, only.values = FALSE, EISPACK = FALSE)
eival<- decomp$values
eivec<- decomp$vectors
if(any(eival <= 0)) {cat("Eigenval contiene ceros y/o valores negativos\n")
                    cat("  ", "\n")
                    k2=0 } else {diag12<- diag(sqrt(eival),n,n)
                    sigm12<- t(eivec)%*%solve(diag12)%*(eivec)
                    S<- sigm12%*(Zs-mediaEV)
                    adt=ad.test(S)
                    adpv<- adt$p.value
                    rech<- ifelse(adpv<=alfa, 1, 0)
                    k2=1}
}
return (cbind(rech,adpv,emv,mean(Zs),var(Zs)))
}
results<-sfLapply(1:m, vswt)
sfStop()
vcuan<- unlist(rbind(results))
ecm<-matrix(vcuan,nrow=m,ncol=11,byrow=T)
vres<-ecm[,1]
norech<-length(which(vres==0))
rechs<- length(which(vres==1))
size<-rechs/m
cov.mod
CMH1

```

```
betaKrn
m
rechs
size
#####
```

Programa Monte Carlo para estimar el tamaño de la prueba para Kriging Ordinario (4).

```
rm(list=ls())
library(spdep)
library(geoR)
library(MASS)
library(stats)
niterr<-100      # Número de muestras Monte Carlo #
m=100           # Número de cuantiles a obtener en cada ciclo bootstrap #
n=300           # Tamaño de muestra, número de sitios #
alfa=0.05       # Cuantil nivel de significancia para rechazar H0 #
rechazos=norech= 0
resb<-matrix(NA,nrow=niterr,ncol=3)
cov.mod="exponential"
betaKr<- rbind(3,0,0)
sigmac<- 4
fhi<- 3
nugg<- 2
gridt<- matrix(NA,nrow=n,ncol=2)
Zs<- matrix(NA,nrow=n,ncol=1)
unos<- matrix(1,nrow=n, ncol=1)
rm(.Random.seed)
runif(1)
x<- .Random.seed[1:(niterr+50)]
i=0
r=1
while (r<=niterr){
i<-i+1
set.seed(x[i+2])
cat(" ", "\n")
cat("Iteracion r=",r," de niterr=",niterr," \n")
```

APÉNDICES

```
cat("seed",x[i+2],"\n")
can<- grf(n, grid = "irreg", xlims = c(0, 50), ylims = c(0, 50), nsim=1,
cov.model =cov.mod,   cov.pars = c(sigmac,fhi), kappa = 0, nugget =nugg,
lambda = 0,   mean = matrix(0,nrow=n,ncol=1), method="cholesky",RF=TRUE,
messages=FALSE)
gridt<- as.matrix(can$coords)
gridf<- cbind(unos,gridt)
mediaKr<- gridf%*%betaKr           # Vector de medias (media o tendencia) #
Zs<- mediaKr+can$data              # Vector de datos Z de s #
#Zs<- mediaKr+Ys
cmd<- cbind(gridt,Zs)              # Matriz de sitios y datos #
Fc<- matrix(NA,nrow=m,ncol=1)
Xs<- cbind(unos,gridt)            #Matriz X(s)
geod<- as.geodata(cmd, coords.col=1:2, data.col=3)
vemp<-variog(geod, option="bin", trend="1st",estimator.type="modulus",
uvec=seq(0,sqrt(2*10*10)/2,l=11),pairs.min=25,nugget.tolerance=0,messages=F)
reg<-lm(vemp$v~vemp$u)
ini.phi<-mean(vemp$v)
ini.pep=reg$coefficients[1]
ini.sig<-var(Zs)-ini.pep
sigml<- likfit(geod, trend="1st", ini.cov.pars = c(ini.sig,ini.phi),
fix.nugget=F,nugget=ini.pep,lik.method="ML",components=F,messages=F)
beta=sigml$beta
beta0=sigml$beta[1]
beta1=sigml$beta[2]
beta2=sigml$beta[3]
sigmasq0=sigml$sigmasq
phi0=sigml$phi
nugg0=sigml$nugget
kapa=sigml$kappa
beta=rbind(beta0,beta1,beta2)
pesp=rbind(sigmasq0,phi0,nugg0)
params<- c(sigmasq0,phi0)
sigma.mat<- varcov.spatial(coords = gridt, dists.lowertri = NULL,
cov.model = "exponential", nugget = nugg0,
cov.pars = c(sigmasq0,phi0),
inv = FALSE, det = FALSE,
func.inv = c("cholesky"),
scaled = FALSE, only.decomposition = FALSE,
sqrt.inv = FALSE, try.another.decomposition = T,
```

APÉNDICES

```
        only.inv.lower.diag = F)
sig=sigma.mat[[1]]
IOzs=rbind(mean(Zs),0,0)
L=diag(1,3,3)
rL<- nrow(L)
LBI0=(L%*(beta)-IOzs)
LXSig<- solve(L%*solve(t(Xs)%*solve(sig)%*Xs)%*t(L))
FO<- (t(LBI0)%*(LXSig)%*LBI0)/rL

decomp<-eigen(sig, only.values = FALSE, EISPACK = FALSE)
eival<- decomp$values
eivec<- decomp$vectors
if(any(eival<= 0)) {cat("Eigenval contiene ceros y/o valores negativos\n")}
cat("  ", "\n")} else {
Fcu<- matrix(NA,nrow=m,ncol=3)

for (k in 1:m){          # Inicia Bootstrap Paramétrico #
Fcu[k,2]<-IOzs[1]
znorm<- mvrnorm(1,(Xs)%*(IOzs),sig)
zsim<- cbind(gridt,znorm)
geod<- as.geodata(zsim, coords.col=1:2, data.col=3)
vemp<-variog(geod, option="bin",trend="1st",estimator.type="modulus",
uvec=seq(0,sqrt(2*10*10)/2,l=11),pairs.min=25,nugget.tolerance=0,messages=F)
reg<-lm(vemp$v~vemp$u)
ini.phi<-mean(vemp$v)
ini.pep=reg$coefficients[1]
ini.sig<-var(Zs)-ini.pep
sigml<- likfit(geod, trend="1st", ini.cov.pars = c(ini.sig,ini.phi),
fix.nugget=F,nugget=ini.pep, lik.method="ML",components=F,messages=F)
beta=sigml$beta
beta0=sigml$beta[1]
beta1=sigml$beta[2]
beta2=sigml$beta[3]
sigmasqj=sigml$sigmasq
phij=sigml$phi
nuggj=sigml$nugget
kapa=sigml$kappa
betavecj=rbind(beta0,beta1,beta2)
params<- c(sigmasqj,phij)
sigma.mat<- varcov.spatial(coords = gridt, dists.lowertri=NULL,
```

APÉNDICES

```

cov.model = "exponential", nugget = nuggj,
cov.pars = c(sigmasqj,phij),
inv = FALSE, det = FALSE,
func.inv = c("cholesky"),
scaled = FALSE, only.decomposition = FALSE,
sqrt.inv=FALSE, try.another.decomposition=FALSE,
only.inv.lower.diag = F)

sigj=sigma.mat[[1]]
I0=rbind(mean(znorm),0,0)
LBIO=(L%*%betavecj-I0)
LXSigj<- solve(L%*%solve(t(Xs)%*%solve(sigj)%*%Xs)%*%t(L))
Fcuan[k,1]<- (t(LBIO)%*%(LXSigj)%*%LBIO)/rL
Fcuan[k,3]<-I0[1]
I0zs=I0
}          # Finaliza Bootstrap Paramétrico #
vcuan<-Fcuan[,1]
Fco<- sort(vcuan, decreasing = FALSE)
Falfa<- quantile(Fco,1-alfa)
cat("F.cero=",FO,"\n")
cat("F.alfa=",Falfa,"\n")
if (FO>Falfa) {
  cat("FO>F.alfa. Se rechaza H0 con alfa=",alfa,"\n")
  cat("Se tiene modelo de Kriging Universal","\n")
  rechazos= rechazos+1} else
  {cat("FO<=F.alfa. No se rechaza H0 con alfa=",alfa,"\n")
  cat("Se tiene modelo de Kriging Ordinario","\n")
  norech=norech+1}
resb[r,1]=x[r+2]
resb[r,2]=FO
resb[r,3]=Falfa
r=r+1
}
}
size=rechazos/niterr
size
```