Universidad de Valparaíso

Facultad de Ciencias

Departamento de Estadística-CIMFAV


# Tamaño Muestral Efectivo en el Modelamiento de Variables Espaciales

por

Juan Carlos Herrera Órdenes

para la opción al grado de

Magíster en Estadística

Valparaíso. Chile

Viernes 12 de Octubre de 2012

Universidad de Valparaíso

Facultad de Ciencias

Departamento de Estadística-CIMFAV


# Tamaño Muestral Efectivo en el Modelamiento de Variables Espaciales

por

Juan Carlos Herrera Órdenes

para la opción al grado de

Magíster en Estadística

Valparaíso. Chile

Viernes 12 de Octubre de 2012

**Universidad de Valparaíso**
**Facultad de Ciencias**
**Departamento de Estadística-CIMFAV**

Los miembros del Comité de Tesis recomendamos que la Tesis «Tamaño Muestral Efectivo en el Modelamiento de Variables Espaciales», realizada por el estudiante Juan Carlos Herrera Órdenes, con número de matrícula xxxxxxx, sea aceptada para su defensa como opción al grado de Magíster en Estadística.

El Comité de Tesis

Dra. Mónica Catalán Reyes
Profesora Guía

Dr. Ronny Vallejos Arriagada
Profesor Co - Guía

Dr. Harvey Rosas
Profesor Informante

Vo. Bo.

Juan Carlos Herrera Órdenes
Departamento de Estadística-CIMFAV

Valparaíso. Chile, Viernes 12 de Octubre de 2012

*La estadística es el único tribunal de apelación*
*para juzgar el nuevo conocimiento.(P.C. Mahalanobis)*

*Conjeturar es barato;*
*conjeturar erróneamente es caro. (Proverbio Chino)*

*Tengo mis resultados hace tiempo,*
*pero no se como llegar a ellos. (C.F.Gauss)*

*a mi esposa Daisy e*
*hija Amanda*
*quiénes son la inspiración*
*en mi vida.*

# Índice general

**Agradecimientos** xi

**I Preliminares** 1

**1. Introducción** 3

1.1. Estimación del Tamaño Muestral . . . 3

1.1.1. Desigualdad de Tchebychev . . . 3

1.1.2. Error de estimación . . . 3

1.2. Comentarios . . . 5

**2. Estadística Espacial** 7

2.1. Introducción . . . 7

2.2. Tipos de Datos Espaciales . . . 7

2.2.1. Datos Geoestadísticos . . . 7

2.2.2. Lattice Data . . . 8

2.2.3. Patrones de puntos . . . 8

2.3. Procesos Espaciales . . . 8

2.4. Estructuras de Correlación Espacial . . . 9

2.4.1. Isotropía . . . 9

2.4.2. Ergodicidad . . . 10

2.4.3. Continuidad Espacial . . . 10

2.5. Modelos Paramétricos para el Semivariograma . . . 11

2.5.1. Parámetros del Semivariograma . . . 11

2.5.2. Modelo lineal . . . 12

2.5.3. Modelo Esférico . . . 12

2.5.4. Modelo Exponencial . . . 12

2.5.5. Modelo Gaussiano . . . 13
2.5.6. Modelo de Mathérn . . . 13
2.5.7. Modelo de Independencia (Efecto Pepita Puro) . . . 13
2.6. Estimación Muestral del Semivariograma . . . 14
2.6.1. Estimador de Matheron . . . 14
2.7. Estimación de los Parámetros del Semivariograma . . . 15
2.7.1. Estimación por Mínimos Cuadrados . . . 15
2.7.2. Estimación ML . . . 16
2.7.3. Estimación REML . . . 16
2.8. Criterios de Selección de Modelos . . . 17

**II Fundamentos** **19**

**3. Tamaño Muestral Efectivo: Algunas Propuestas** **21**
3.1. Introducción . . . 21
3.2. Efecto de la Correlación en el Tamaño Muestral Efectivo . . . 21
3.2.1. Demostración . . . 22
3.3. El Tamaño Muestral Efectivo en una Muestra Geográfica . . . 25
3.3.1. Factor de Inflación de la Varianza y el Tamaño Muestral Efectivo . . . 26
3.3.2. Tamaño Muestral Efectivo con estructura de Correlación Intra Clase y Autoregresiva . 27
3.4. Tamaño Muestral Efectivo para Variables Espaciales . . . 29
3.4.1. *ESS* como Cantidad de Información de Fisher . . . 29
3.4.2. Reducción Efectiva en el Tamaño Muestral . . . 30
3.4.3. Transformaciones para estabilizar la varianza (Método Delta) . . . 32
3.5. *ESS* con Modelos Paramétricos para el Semivariograma . . . 34
3.5.1. *ESS* con Modelo Exponencial . . . 34
3.5.2. *ESS* con Modelo Esférico . . . 35
3.5.3. *ESS* con Modelo Mathérn . . . 35
3.5.4. *ESS* con Modelo Gaussiano . . . 35
3.5.5. Ejemplos . . . 35
3.6. Comentarios . . . 36

**4. Muestreo Espacial** **39**

4.1. Introducción . . . 39
4.2. Propósitos y Realización del Muestreo . . . 40
4.3. Métodos de Muestreo e Inferencia Estadística . . . 40
4.3.1. Método de Muestreo e Inferencia Estadística . . . 40
4.4. Enfoques del Muestreo Espacial . . . 41
4.4.1. Basado en el Diseño . . . 41
4.4.2. Basado en el Modelo . . . 41
4.5. Decisiones importantes del Diseño . . . 42
4.5.1. Elección entre la inferencia basada en el diseño y basada en modelos . . . 42
4.6. Panorámica en los Tipos de Diseños de Muestreo . . . 43
4.6.1. Eligiendo una Estrategia Basada en el Diseño . . . 44
4.7. Esquemas de Muestreo Espacial . . . 48
4.7.1. Muestreo Basado en el Diseño . . . 48
4.7.2. Muestreo Basado en el Modelo . . . 50
4.8. Comentarios . . . 51

**5. Simulación del Tamaño Muestral Efectivo ESS 53**

5.1. Introducción . . . 53
5.2. Algoritmos para simular ESS . . . 53
5.2.1. Algoritmo N°1 . . . 53
5.2.2. Algoritmo N°2 . . . 54
5.3. Tamaño Muestral Efectivo con distintas Correlaciones Espaciales . . . 54
5.3.1. Correlación Intraclase $ESS_{intra}(n, \rho)$ . . . 54
5.3.2. Correlación Autoregresiva de Primer Orden (AR(1)) $ESS_{AR(1)}(n, \phi)$ . . . 55
5.3.3. Correlación Espacial $ESS(n, \theta)$ . . . 55
5.4. Gráficos . . . 56
5.5. Comentarios . . . 59

**III Estimación del Tamaño Muestral Efectivo en Jóvenes Pinos Radiata 61**

**6. Aplicación 63**

6.1. Introducción . . . 63
6.2. Método de Aplicación para ESS . . . 63

6.3. Variables de producción en plantaciones de jóvenes Pinos Radiata . . . . . 64
6.4. Análisis de datos Exploratorio Espacial . . . . . 64
6.5. Estimación Tamaño Muestral Efectivo . . . . . 69
6.5.1. Selección de Modelos, Estimación de Parámetros y Cálculo del Tamaño Muestral Efectivo 69
6.5.2. Muestreo y estimación de $ESS$ para Área Basal $m^2$ y Altura $m$ . . . . . 70
6.6. Comentarios . . . . . 71

**7. Conclusión** **73**

**8. Bibliografía** **75**

**A. Anexo N°1: Rutinas R - Simulación** **77**

**B. Anexo N°2: Rutinas R - Modelación Datos** **87**

# Índice de figuras


2.1. Comportamiento de un semivariograma acotado con una representación de los parámetros básicos. • SEMEXP corresponde al semivariograma experimental y ⇁ MODELO al ajuste de un modelo teórico. . . . . 11

2.2. Comparación de los modelos de Mathérn, Exponencial, Esférico, Gaussiano, Lineal y Pepita Puro, respecto a una escala simulada entre 0 y 1. . . . . 14

3.1. $ESS_{intra}(n, \rho)$ con $n = 20, 50, 100$ y $0 < \rho < 1$ . . . . . 31

3.2. $ESS_{AR(1)}(n, \phi)$ con $n = 20, 50, 100$ y $0 < \phi < 1$ . . . . . 32

3.3. Tamaño Muestral Efectivo con diferentes correlación espacial (a) Exponencial, (b) Esférica, (c) Mathérn y (d) Gaussiana. . . . . 36

4.1. Representación espacial 3D de los ingresos por viviendas en ciento de miles. . . . . 39

4.2. Similitudes y diferencias entre los tipos básicos de diseño . . . . . 43

4.3. Opciones principales para decidir sobre una estrategia de muestreo basado en el diseño. . . . 44

4.4. Árbol de decisiones para ayudar a la elección de un tipo de diseño para cantidades globales en el espacio. . . . . 45

4.5. Continuación Figura 4.4 . . . . . 46

4.6. Continuación Figura 4.5 . . . . . 47

4.7. Ejemplo teórico de una muestra aleatoria simple . . . . . 48

4.8. Ejemplo teórico de una muestra aleatoria simple estratificado . . . . . 48

4.9. Ejemplo teórico de una muestra en dos etapas. . . . . 49

4.10. Ejemplo teórico de una muestra en grupos. . . . . 49

4.11. Ejemplo teórico de una muestra sistemática. . . . . 50

4.12. Ejemplo teórico de una muestra en rejilla centrada. . . . . 50

4.13. Ejemplo teórico de una muestra geoestadística. . . . . 51

4.14. Código Fuente en R, extraído de Bivand et al. (2008). Applied Spatial Data Analysis Spatial with R. Pág 137 . . . . . 51

5.1. (a) Histograma para $\widehat{ESS}_{intra}(n, \hat{\rho}_{1_K})$, $\rho_1$=0.8 con $K = 2000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{intra}(n, \hat{\rho}_{1_K})$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{intra}(n, \hat{\rho}_{1_K})$ versus la distribución teórica Normal. . . . . 56

5.2. (a) Histograma para $\widehat{ESS}_{intra}(n, \hat{\rho}_{2_K})$, $\rho_2$=0.3 con $K = 2000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{intra}(n, \hat{\rho}_{2_K})$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{intra}(n, \hat{\rho}_{2_K})$ versus la distribución teórica Normal. . . . . 56

5.3. (a) Histograma para $\widehat{ESS}_{intra}(n, \hat{\rho}_{3_K})$, $\rho_1$=0.01 con $K = 2000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{intra}(n, \hat{\rho}_{3_K})$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{intra}(n, \hat{\rho}_{3_K})$ versus la distribución teórica Normal. . . . . 57

5.4. (a) Histograma para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{1_K})$, $\phi_1 = 0.7$ con $K = 2000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{1_K})$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{1_K})$ versus la distribución teórica Normal. . . . . 57

5.5. (a) Histograma para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{2_K})$, $\phi_2 = 0.3$ con $K = 2000$ y $n = 500$ , (c) QQ-Normal para para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{2_K})$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{2_K})$ versus la distribución teórica Normal. . . . . 57

5.6. (a) Histograma para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{3_K})$, $\phi_3 = 0.1$ con $K = 2000$ y $n = 500$ , (c) QQ-Normal para para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{3_K})$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{3_K})$ versus la distribución teórica Normal. . . . . 58

5.7. (a) Histograma para $\widehat{ESS}_{Exp}$, con $K = 1000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{Exp}$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{Exp}$ versus la distribución teórica Normal . . . . . 58

5.8. (a) Histograma para $\widehat{ESS}_{Esf}$, con $K = 1000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{Esf}$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{Esf}$ versus la distribución teórica Normal. . . . . 58

5.9. (a) Histograma para $\widehat{ESS}_{Math}$, con $K = 1000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{Math}$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{Math}$ versus la distribución teórica Normal. . . . . 59

6.1. Técnico Forestal realizando medición Área Basal $m^2$ y Altura $m$ de Pinos Jóvenes Radiata y su distribución en el espacio. . . . . 64

6.2. Histogramas y boxplot para Área Basal. . . . . 65

6.3. Histogramas y boxplot para Altura. . . . . 66

6.4. Puntos 3D y Mapa de la distribución de la Área Basal. . . . . 66

6.5. Puntos 3D y Mapa de la distribución de la Altura. . . . . 67

6.6. Dispersión Longitud-Latitud para Área Basal y Altura. . . . . 67

6.7. Dispersión Longitud-Latitud para Área Basal y Altura. . . . . 68

6.8. Distribución espacial de Área Basal y Altura desde distintos ángulos de visión. . . . . 68

6.9. Semivariogramas Muestral para Área Basal $m^2$ (a) y Altura $m$ (b) estimado mediante REML. 69

6.10. De las 688 unidades georeferenciadas, graficamos la distribución espacial de $\widehat{ESS}_{Esf} = 23$ para el Área Basal $m^2$ y $\widehat{ESS}_{Gauss} = 27$ para la Altura $m$, bajo un muestreo espacial aleatorio simple. . . . . 70

# Índice de tablas


4.1. Definición del método de diseño basado en modelos y basado en una combinación de un método de selección de unidades de muestreo y un método de inferencia estadística. . . . . 41

5.1. Estadísticas Descriptivas para $\widehat{ESS}_{intra}(n, \widehat{\rho}_K)$ con $n = 500$, $K = 2000$, $\rho_1$ =0.8, $\rho_2$ =0.3, $\rho_3$ =0.01, $\mu = 2$ y $\sigma^2 = 1$ . . . . . 54

5.2. Estadísticas Descriptivas para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_K)$, con $\phi_1$ =0.7, $\phi_2$ =0.3, $\phi_3$ =0.1, $\mu = 2$ y $\sigma = 1$, $K = 2000$ y $n = 500$ . . . . . 55

5.3. Estadísticas Descriptivas para cada $\widehat{ESS}_{esp}(n, \widehat{\theta})$ con distintos modelos, $n = 500$ y $K = 1000$ 55

6.1. Estadísticas Descriptivas para Área Basal $m^2$ y Altura $m$ . . . . . 65

6.2. Estimación Tamaño Muestral Efectivo vía $REML$ . . . . . 69

6.3. Estadísticas Descriptivas para $\widehat{ESS}_{Esf_1}$ y $\widehat{ESS}_{Esf_2}$ . . . . . 71

6.4. Intervalos de confianza del 95 % para $Y_1(s)$ e $Y_2(s)$. . . . . 71

# Agradecimientos

A mi profesor guía Dr. Ronny Vallejos por su apoyo y dedicación más allá de su labor académica.

Al apoyo incondicional de mis Padres, Familia y Esposa.

A todos quienes colaboraron con este trabajo, profesores que me brindaron parte de su tiempo y paciencia.

Muchas gracias.

# Resumen

En la actualidad para la selección de un conjunto de variables aleatorias (ó una muestra aleatoria) de tamaño $n$ provenientes desde una población de interés de tamaño $N$, hay que tener en cuenta dos requisitos; que las muestras seleccionadas sean independientes e idénticamente distribuidas. Esto es uno de los requisitos básicos para realizar inferencia estadística, donde $N \geq n$. Mas aún, bajo ciertas condiciones de regularidad y una probabilidad de significancia dada se puede establecer fórmulas analíticas que permiten calcular un número $n$ necesario de muestras, para estimar algún parámetro de interés (comúnmente, media y desviación estándar). En la literatura existen varias perspectivas o metodologías para realizar técnicas de muestreo (Ver, Hansen, 1953 y Chocran, 1977).

Otro escenario muy distinto, es cuando se está en presencia de un conjunto de variables cuyos atributos están georeferenciados, donde a menudo se puede encontrar que la información está duplicada, o bien, existe redundancia (se entiende por redundancia de la información, a los datos en la muestra que contienen la misma información que otro dato, o bien, están correlacionados). Debido a la característica de este conjunto de datos no se pueden aplicar las técnicas y fórmulas descritas en Chocran (1977) para obtener un tamaño de muestra. Sin embargo, existen autores como Haining (1990, 2004), Anselin (2010), Cressie (1993) y Griffith (2005, 2008) que discuten este tema en profundidad y han tratado de dar respuesta a esta situación.

Este impacto requiere tomar en cuenta la información entregada por la autocorrelación espacial, como también, poseer una estructura necesaria para determinar el *Tamaño Muestral Efectivo para el Modelamiento de Variables Espaciales*. Los impactos de la autocorrelación espacial pueden ser especificados y modelado entre otras, mediante el Factor de Inflación de Varianza estimada, denominada VIF. Esta conceptualización es similar a los impactos de multicolinealidad sobre los errores estándares de los coeficientes de regresión lineal múltiple estimados. (Griffith, 2005, 2008). *El Tamaño Muestral Efectivo para el modelamiento de variables espaciales*, está asociado a la configuración espacial de los datos, es decir, la distribución espacial de los datos en una zona geográfica. En la literatura se pueden encontrar perspectivas generales sobre el diseño muestral geográfico vistas en Muller (2001) y de Gruijter et al (2006).

Este trabajo se enfoca principalmente en la cuantificación de la correlación espacial y como esta afecta el *Tamaño Muestral Efectivo*. En el contexto de variables aleatorias espaciales, se define el *Tamaño Muestral Efectivo*, denotado como $ESS$ dado su acrónimo en inglés (Effective Sampling Size); a un subconjunto de elementos que pertenecientes una muestra general con atributos georefenciados de tamaño $n$, en donde existe una cantidad $ESS < n$ que mide el número de observaciones independientes necesarias para estimar la media (muestral) de un proceso. El valor o la cantidad que tome $ESS$ va a depender de cuan persistente sea la correlación espacial en la muestra georeferenciada.

Esta concepción motiva a estudiar los efectos de la correlación espacial en el $ESS$. Como también, estudiar expresiones analíticas que contengan la información de la estructura de correlación espacial (semivariograma) y el tamaño de la muestra, en este caso $n$ y como se reduce $ESS$, ya que bajo ciertas condiciones, esta cantidad fluctúa entre, $1 < ESS < n$.

Para validar la reducción de la información en muestras con atributos georeferenciados se realizarán estudios de simulación para distintos modelos y parámetros, de esta forma se cuantificará el $ESS$. Estos resultados se aplicarán a datos georeferenciados reales, tomados del ámbito forestal donde se miden carac-

terísticas de productividad de ciertos pinos radiata como son área basal ($m^2$) y altura ($m$).

La estructura de este trabajo se divide en tres partes. En la primera parte, **Preliminares** inicia con una introducción a los primeros cálculos en la estimación del tamaño muestral clásico. Luego, se definirán las características principales de la estadística espacial y los tipos de variables aleatorias espaciales, o bien, tipos de datos espaciales. También, se definirán estas variables bajo un proceso estocástico y sus propiedades. Algunas estructuras de correlación espacial, principalmente se hablará del semivariograma y función de correlación espacial asociado a modelos paramétricos para el semivariograma. Métodos de estimación de parámetros como; Mínimos Cuadrados (OLS), Mínimos cuadrados Generalizados o ponderados (GLS), Máxima Verosimilitud (MV) y Máxima Verosimilitud Restringida (REML).

La segunda parte, **Fundamentos** comienza planteando algunos aspectos de como la literatura ha tratado de dar una solución al cálculo del tamaño muestral efectivo en el modelamiento de variables espaciales, los autores que se pueden destacar son; Cressie (1993), Griffith (2005, 2008), Haining (1990, 2004), Vallejos (2010). El objetivo principal de este capítulo es destacar como la correlación espacial es un elemento importante al momento de calcular el tamaño muestral efectivo. Se plantean distintas expresiones analíticas para el tamaño muestral efectivo con varias estructuras de correlación espacial, como son, la correlación autoregresiva de primer orden e intraclase, en estos dos casos se pueden obtener expresiones analíticas explícitas para determinar el tamaño muestral efectivo. En este mismo sentido, la situación se hace más compleja cuando se quiere incorporar una estructura de correlación espacial, ya sea, lineal, exponencial, esférica, gaussiana y de Mathérn, debido a que no se pueden encontrar expresiones analíticas para el tamaño muestral efectivo. Para el caso que las expresiones del tamaño muestral efectivo tienen una forma explícita, se pueden utilizar métodos de transformaciones para estabilizar la varianza, esto nos permite encontrar la distribución asintótica de una función no lineal asociado a un estimador, ya sean estimados mediante OLS, MV u otro método. Como complemento a este análisis, se presentan algunos métodos de Muestreo Espacial, como son muestreo aleatorio, sistemático, estratificado, no alineado y de teselación hexagonal que tienen una cierta similitud con los métodos clásicos vistos en Cochran pero ahora las unidades de muestreo están georeferenciadas. Finalmente, se inspeccionan las expresiones para el cálculo del tamaño muestral efectivo mediante simulación. El objetivo de las simulaciones es ver el comportamiento distribucional del tamaño muestral efectivo al momento de estimar los parámetros.

Finalmente en la tercera parte, se realizará la aplicación de la **Estimación del Tamaño Muestral Efectivo en Jóvenes Pinos Radiata** a datos tomados del sector forestal, como son, área basal ($m^2$) y altura ($m$) en plantaciones de pinos, ambas medidas de productividad de bosques. El sitio de estudio, tiene una superficie de 1.244,43 hectáreas, está ubicado en el sector del ≪Escuadrón≫, al sur de Concepción en la porción sur de Chile ($36^o\ 54'S$,$73^o\ 54'O$) y pertenece a la empresa Forestal MININCO S.A. El objetivo es presentar como la correlación espacial interviene en el tamaño muestral efectivo. Y finalmente se presentará una **Conclusión** con los resultados más importantes y los aspectos más relevantes a la hora de realizar una selección de una muestra con atributos geográficos.

Palabras Claves: *Muestras Georeferenciadas, Correlación Espacial, Semivariograma, Tamaño Muestral Efectivo ESS.*

## Objetivos General y Específicos

### Objetivo general:

- Determinar el Tamaño Muestral Efectivo para el modelamiento de variables aleatorias espaciales correlacionadas.

### Objetivos específicos:

- Presentar problema de investigación y motivación del estudio de tesis.
- Presentar algunas propuestas realizadas por varios autores.
- Demostrar mediante simulación como los efectos de la autocorrelación espacial influyen en significativamente en el tamaño muestral efectivo.
- Investigar y estudiar la distribución del Tamaño Muestral Efectivo.
- Encontrar intervalos de confianza explícitos para el Tamaño Muestral Efectivo.
- Aplicar los resultados a datos georeferenciados.

# Parte I

# Preliminares

Capítulo 1

# Introducción

Para entender el cálculo del tamaño muestral efectivo, es necesario tener en cuenta algunos de los resultados que inician la formulación de esta medida. En este capítulo se comenzará con un resultado asociado al cálculo del tamaño muestral clásico, introduciendo intuitivamente los principales resultados usados actualmente.

## 1.1 Estimación del Tamaño Muestral

En la planificación de un diseño muestral, hay que tener en cuenta varias situaciones, como por ejemplo, el escenario donde se aplicará el diseño, ver los alcances que se necesitan para la decisión del tamaño de la muestra. Esta decisión es de suma importancia, debido a que una muestra demasiado grande implica un gasto mayor de recursos, o bien, una muestra pequeña disminuye la utilidad de los resultados para la estimación del parámetro de interés.

Uno de los resultados importantes que facilita el cálculo del tamaño muestral, es la desigualdad de Tchebychev. Esta desigualdad es muy utilizada en probabilidad debido a que este resultado ofrece una cota inferior a la probabilidad de que el valor de una variable aleatoria con varianza finita esté a una cierta distancia de su esperanza matemática.

### 1.1.1 Desigualdad de Tchebychev

**Teorema 1.1** *Sea $X$ una variable aleatoria de media $\mu$ y varianza finita $\sigma^2$. Entonces para todo número real, $k > 0$.*

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \tag{1.1}$$

Note que haciendo la siguiente sustitución, $\epsilon = k\sigma$, se obtiene una desigualdad equivalente:

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}. \tag{1.2}$$

### 1.1.2 Error de estimación

Cuando se tiene una estimación $\widehat{\theta}$ respecto a un parámetro poblacional $\theta$ se necesita establecer un criterio referente a la precisión del parámetro de interés y la estimación, el cual llamamos error de estimación. El error de estimación o error de diseño denotado por $\epsilon$, es la diferencia en valor absoluto entre el parámetro poblacional $\theta$ y el estimador muestral $\widehat{\theta}$, es decir,

$$\epsilon = |\theta - \widehat{\theta}|. \tag{1.3}$$

Esta cantidad varía de manera aleatoria en muestreos repetitivos, y obviamente se desea que sea lo más pequeña posible.

Tomando en cuenta los resultados (1.1) y (1.3) consideremos el siguiente ejemplo: Supongamos que se tiene una muestra aleatoria de tamaño $n$, es decir, $\{y_1, \ldots, y_n\}$ provenientes de una población de tamaño $N$, donde $N \geq n$. Además, supongamos que cada una de las combinaciones $\binom{N}{n}$ posibles de la muestra aleatoria tiene la misma probabilidad de ser seleccionada, es decir, $\binom{N}{n}^{-1}$, donde $\binom{N}{n} = \frac{N!}{n!(N-n)!}$.

La inferencia está orientada hacia la media poblacional, digamos $\mu$. Sea $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ un estimador insesgado de $\mu$. Supongamos, que $\overline{y} \sim N(\mu, \sigma^2_{\overline{y}})$.

El objetivo es determinar un tamaño muestral necesario para estimar el parámetro $\mu$ de tal forma que el error de estimación no supere cierta presición, la cual se puede escribir como, $|\overline{y} - \mu| \leq \epsilon$. Ahora bien, utilizando la expresión (1.2) y (1.3), se puede obtener que:

$$\mathbb{P}(|\overline{y} - \mu| \leq \epsilon) \geq 1 - \frac{\sigma^2_{\overline{y}}}{\epsilon^2}. \tag{1.4}$$

Para determinar el tamaño muestral se puede resolver mediante dos vias alternativas. La primera es utilizar directamente la desigualdad de Tchebychev, fijando el error de estimación y una cierta probabilidad. La segunda alternativa es utilizar el teorema del límite central, utilizando la información de la distribución del estimador, es decir:

$$\mathbb{P}\left(|\overline{y} - \mu| \leq \epsilon\right) = \mathbb{P}\left(-\epsilon \leq \overline{y} - \mu \leq \epsilon\right) = \mathbb{P}\left(-\frac{\epsilon}{\sigma_{\overline{y}}} \leq \frac{\overline{y} - \mu}{\sigma_{\overline{y}}} \leq \frac{\epsilon}{\sigma_{\overline{y}}}\right) = 1 - \alpha \tag{1.5}$$

Se sabe que la siguiente transformación $Z = \frac{\overline{y} - \mu}{\sigma_{\overline{y}}} \sim N(0, 1)$, entonces, el término $\frac{\epsilon}{\sigma^2_{\overline{y}}}$ es el percentil $Z_{1-\alpha/2}$ de la distribución normal estándar, es decir,

$$\frac{\epsilon}{\sigma_{\overline{y}}} = Z_{1-\alpha/2}. \tag{1.6}$$

La varianza de la media muestral, denotada por $\sigma^2_{\overline{y}}$, para un muestro aleatorio simple es:

$$\sigma^2_{\overline{y}} = \mathbb{E}\left(\overline{y} - \overline{Y}\right)^2 = \frac{(N-n)S^2}{Nn}, \tag{1.7}$$

$$\sigma_{\overline{y}} = \sqrt{\frac{(N-n)S^2}{Nn}}. \text{ (Ver Cochran, 1977. pág. 23)} \tag{1.8}$$

donde $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})$ y $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. Por lo tanto, la presición puede ser escrita como:

$$\epsilon = Z_{1-\alpha/2}\sqrt{\frac{(N-n)S^2}{Nn}}, \tag{1.9}$$

realizando algunas operaciones algebraicas y despejando en función de $n$, se puede obtener la siguiente expresión:

$$n = \frac{NZ^2_{1-\alpha/2}S^2}{N\epsilon^2 + Z^2_{1-\alpha/2}S^2} = \frac{\frac{Z^2_{1-\alpha/2}S^2}{\epsilon^2}}{1 + \frac{Z^2_{1-\alpha/2}S^2}{N\epsilon^2}}. \tag{1.10}$$

En la práctica el tamaño muestral $n$ es un número natural, por lo que debemos tomar el valor entero más cercano al obtenido. Mientras que los valores de $\epsilon$ y $Z^2_{1-\alpha/2}$ dependen de la precisión deseada y el nivel de confianza que se desea imponer al diseño muestral.

En cuanto a $S^2$, debemos obtenerlo de investigaciones anteriores en la misma población, resultados de análisis de poblaciones parecidas a esta, o realizar un estudio piloto de la muestra.

Finalmente, el tamaño muestral óptimo se puede expresar como:

$$n_0 = \frac{Z^2_{\alpha/2}S^2}{\epsilon^2} \Rightarrow n = \frac{\frac{Z^2_{\alpha/2}S^2}{\epsilon^2}}{1+\frac{Z^2_{\alpha/2}S^2}{N\epsilon^2}} = \frac{n_0}{1+\frac{n_0}{N}}. \tag{1.11}$$

En este caso $n_0$ es el tamaño muestral inicial. El resultado anterior nos permite determinar un tamaño óptimo para un conjunto de variables aleatorias independientes cuando estas se encuentran dispuestas en lotes o a granel. La metodología asociada a este resultado comúnmente se denomina muestreo aleatorio simple (ver Cochran, 1977).

Consideremos $n_0$ para determinar el tamaño muestral inicial para estimar el parámetro de interés $\mu$ con un error máximo permisible $\epsilon$ prefijado y la varianza poblacional $\sigma^2$, descrito como:

$$n_0 = \left(\frac{SZ_{1-\alpha/2}}{\epsilon}\right)^2. \tag{1.12}$$

Considere los siguientes datos, $\epsilon = 0.12$, $S = 0.9$, $1-\alpha = 0.95 \Rightarrow 1-\alpha/2 = 0.975 \Rightarrow Z_{1-\alpha/2} = 1.96$, se puede ver lo siguiente: $n_0 = \left(\frac{(1.96)(0.9)}{0.12}\right)^2 = 261.09$.

Claramente en la práctica no se puede obtener un tamaño de muestra fraccionario por lo que se aproxima o redondea. En este se caso se puede observar que para un nivel de confianza del 95 %, un máximo de error prefijado de 0.12 y una dispersion de 0.9 unidades, se requiere un tamaño de 262 muestras para estimar el parámetro poblacional $\mu$.

## 1.2 Comentarios

Una de las ventajas de usar la desigualdad de Tchebychev es que no supone ninguna condición sobre la distribución del estimador. Sin embargo, el valor del tamaño muestral que se deduce suele estar por encima del verdadero tamaño muestral necesario para cumplir la presición prefijada.

Por otro lado, cuando se utiliza la aproximación normal, se debe suponer una distribución muy concreta, que en este caso, el estimador tiene una distribución normal. No obstante, este supuesto permite determinar fórmulas analíticas exactas para el tamaño muestral.

Estos ventajas y desventajas nos a ayuda establecer una característica en común que tiene la metodología de muestreo, cuando se quiere determinar un tamaño muestral. A grandes rasgos, se puede decir que las características principales que tiene la metodología de muestreo son las siguientes:

- Definir el diseño muestral a utilizar (Ya sea, Muestreo Aleatorio, Sistemático, Estratificado, Dos etapas, etc. bajo un Marco Muestral),
- Definir el estimador a utilizar y prefijar el error de estimación,
- Utilizar una ecuación o inecuación para determinar el tamaño muestral (fórmula asociada al diseño muestral),
- Ya calculado el tamaño muestral, aplicar la metodología de muestreo,
- Seleccionar la muestra y calcular el estimador muestral de interés
- Realizar inferencia.

Como ya se ha mencionado anteriormente, una de las características más importantes que tiene el muestreo clásico, es el hecho que las variables muestreadas tienen que ser mutuamente independientes e igualmente distribuidas, ya que, desde el punto distribucional conjunto de estas variables, facilita la simplificación del modelo probabilístico que subyace los datos, lo cual, facilita la estimación del parámetro. Otra característica que tienen los diseños muestrales en general, es que nos permite obtener información no redundante por cada variable aleatoria seleccionada, es decir, una variable aleatoria o un subconjunto de ellas, no puede ser representada como una combinación lineal de otras variables restantes, cada variable aleatoria incorpora información útil al modelo probabilístico, su función de autocorrelación es igual a cero para todas las combinaciones existentes entre las variables aleatorias.

Otro escenario muy distinto pero que presenta la misma problemática antes mencionada se puede encontrar cuando se está interesado en realizar un diseño de muestreo para variables aleatorias que se encuentran georeferenciadas. Este hecho presenta varias complicaciones debido a la presencia de correlación espacial en los datos. La correlación espacial en las variables aleatorias es un problema, ya que, no cumple las propiedades mencionadas anteriormente.

Sin embargo, en el contexto de variables espaciales, esta información va ser de suma importancia y tiene que ser considerada. En la literatura, existen medidas que entregan información sobre la existencia o no de correlación espacial, como es el caso del índice de Moran. Esta nueva perspectiva en datos georefenciados motiva proponer otras técnicas de estimación del tamaño muestral efectivo de las que se conocen comúnmente. Las cuales serán presentadas más adelante.

Capítulo 2

# Estadística Espacial

## 2.1 Introducción

Para determinar el tamaño muestral efectivo en el modelamiento de variables espaciales, es necesario establecer una estructura que relacione la información georeferenciada y la información que pueda existir en la autocorrelación espacial de las unidades experimentales seleccionadas, para esto es necesario comenzar definiendo algunos resultados asociado a la estadística espacial. La estadística espacial ó geoestadística es la aplicación de la teoría de las variables regionalizadas a la estimación de procesos o fenómenos en el espacio.

**Definición 2.1** ***Estadística Espacial:*** *La estadística espacial es una rama de la estadística que tiene como objetivo el análisis y modelamiento de datos georeferenciados en el espacio* $\mathbb{R}^2$, $\mathbb{R}^3$ *y* $\mathbb{R}^d$. *Usualmente se asume que* $\{s_1, s_2, \ldots, s_n\}$ *son las localidades donde se observan las variables* $\{Y(s_1), Y(s_2), \ldots, Y(s_n)\}$.

**Definición 2.2** ***Primera Ley de la Geografía:*** *Todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes.*

## 2.2 Tipos de Datos Espaciales

Esta rama posee ciertas características propias, donde es importante identificar como están dispuesta las unidades experimentales espaciales. Consideremos el dominio $D \subset \mathbb{R}^d$. Entonces la variable $Y(s)$ es medida para cada $s \in D$. Se define así, un proceso estocástico espacial de dimensión $d$, como el conjunto,

$$\{Y(s), s \in D \subset \mathbb{R}^d\}. \tag{2.1}$$

Esto permite que los datos espaciales puedan ser pensados como las realizaciones de un campo aleatorio multivariado $\{Y(s), s \in D \subset \mathbb{R}^d\}$, donde $Y \in \mathbb{R}^m$, donde $m \geq 1$ y la ubicación $s$ puede ser un conjunto en el espacio euclidiano $d-$ dimensional con $d \geq 1$.

Cressie (1993) considero que la mayoría de los problemas de la estadística espacial están identificados en cuatro categorías según sea el domino espacial $D$ fijo ó aleatorio, o bien, $D$ continuo ó discreto.

### 2.2.1 Datos Geoestadísticos

Las características que tienen estos tipo de datos son: dado un dominio continuo $D$, las localidades $s \in D$ son no aleatorias, sino más bien fijas. Para cada ubicación en $s$ se tiene una variables aleatoria $Y(s)$ que fue medida para cualquier $s \in D$. Generalmente, estos datos son seleccionados según el juicio del investigador. Un ejemplo de estos datos se puede encontrar en el estudio de la contaminación del aire u otras áreas de las geociencia.

### 2.2.2 Lattice Data

En este tipo de datos el dominio $D$ es fijo y discreto (es decir, no aleatorio y contable) y existen dos casos; la grilla rectangular y la grilla no rectangular. Una grilla es definida como un conjunto de índices de localizaciones que tiene asociado un conjunto de vecinos. Comúnmente en datos definidos sobre grillas no rectangulares, el punto representativo de cada sitio es el centroide de la región.

### 2.2.3 Patrones de puntos

En este tipo de datos, tanto la observación $Y$ como la localidad $s$ son aleatorias, sin embargo, $s$ es importante y la variables $Y$ no es relevante. La idea es conocer la distribución de las localidades $s_i$ en el dominio $D$ y no cómo se distribuyen las observaciones $Y$.

## 2.3 Procesos Espaciales

**Definición 2.3** *Un proceso estocástico $\{Y(s), s \in D\}$ es un conjunto de variables aleatorias definidas sobre un mismo espacio de probabilidad $(\Omega, \mathfrak{F}, \mathbb{P})$ indexadas en un conjunto $D$, entonces $Y(s)$ una función,*

$$\begin{array}{rccc} Y: & \Omega \times D & \longrightarrow & \mathbb{R} \\ & (\omega, s) & \mapsto & Y(\omega, s) \end{array}$$

**Definición 2.4** *Sea $\{Y(s), s \in D \subset \mathbb{R}^d\}$ un proceso estocástico. Se dice que $\{Y(s)\}$ es de segundo orden si y sólo sí:*

$$\mathbb{E}[Y^2(s)] < \infty, \ \forall \ s \in D. \tag{2.2}$$

*Se define el espacio*

$$L^2 = \{Y(s) : \mathbb{E}[Y^2(s)] < \infty, \ \forall \ s \in D\}, \tag{2.3}$$

*entonces se cumple la siguiente propiedad*

$$Y(s) \in L^p \Rightarrow Y(s) \in L^q, \ \forall \ q \leq p, \ \forall \ s \in D.$$

**Definición 2.5** *Proceso Gaussiano: Sea $\{Y(s), s \in D\}$ un proceso estocástico, diremos que $\{Y(s)\}$ es un proceso gaussiano si y solo sí para cada $n \in \mathbb{N}$, $\{s_1, s_2, \ldots, s_n\} \in D$ y el vector $Y(s) = (Y(s_1), Y(s_2), \ldots, Y(s_n))^t$ es normal multivariante con media $\mu(s)$ y matriz de covarianza $\Sigma$ cuya función de densidad es:*

$$f_Y\left(\mu(s), \Sigma\right) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(Y(s) - \mu(s))^t \Sigma^{-1}(Y(s) - \mu(s))\right\}. \tag{2.4}$$

**Definición 2.6** *Sea $\{Y(s), s \in D \subset \mathbb{R}^d\}$ un proceso estocástico. Se dice que $\{Y(s)\}$ es estrictamente estacionario o fuertemente estacionario si y solo si $\forall \ n \in \mathbb{N}$ y para cada $h \in \mathbb{R}$ la distribución de los vectores $(Y(s_1), Y(s_2), \ldots, Y(s_n))$ y $(Y(s_1+h), Y(s_2+h), \ldots, Y(s_n+h))$ es la misma. Es decir, $\{Y(s)\}$ es invariante bajo traslaciones. O bien,*

$$\mathbb{P}\left(Y(s_1) \leq t_1, Y(s_2) \leq t_2, \ldots, Y(s_n) \leq t_n\right) = \mathbb{P}\left(Y(s_1+h) \leq t_1, Y(s_2+h) \leq t_2, \ldots, Y(s_n+h) \leq t_n\right), \\ \forall n \in \mathbb{N}, \forall s_1, \ldots, s_n, h \in \mathbb{R}^d. \tag{2.5}$$

**Definición 2.7** *Se dice que $\{Y(s)\}$ es un proceso débilmente estacionario si y sólo si:*

$$\mathbb{E}[Y(s)] = \mu, \forall s \in D \tag{2.6}$$

$$\mathbb{V}ar[Y(s)] = \mathcal{C}(0), \tag{2.7}$$

$$\mathcal{C}ov[Y(s_i), Y(s_j)] = \mathcal{C}(s_i - s_j) = \mathcal{C}(h). \tag{2.8}$$

Algunas propiedades de la función de autocovarianza para procesos débilmente estacionarios son:

1. $\mathcal{C}(0) \geq 0$.
2. $\mathcal{C}(h) = \mathcal{C}(-h)$, $\forall h \in D$.
3. Si $\mathcal{C}_j(h)$ son funciones de covarianza en $\mathbb{R}^d$, entonces:
   $\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \cdot \mathcal{C}(s_i - s_j) \geq 0 \ \forall\ s_i,\ s_j,\ a_i,\ a_j,\ \forall\ s_i,\ s_j \in D$ y $a_1, \ldots, a_n \in \mathbb{R}$.
4. Si $\mathcal{C}_j(h)$ son funciones de covarianza $j = 1, \ldots, k$, entonces:
   $\sum_{j=1}^{k} b_j \cdot \mathcal{C}(h)$ es una función de covarianza, para $b_1, \ldots, b_n$ constantes positivas.
5. Si $\mathcal{C}_j(h)$ son funciones de covarianza en $\mathbb{R}^d$, entonces:
   $\prod_{j=1}^{k} \mathcal{C}_j(h)$ es una función de covarianza en $\mathbb{R}^p$, $\forall\ p < d$.

**Definición 2.8** *Si $\{Y(s)\}$ es un proceso de segundo orden, entonces la función de covarianza se puede definir de la siguiente manera:*

$$\begin{array}{rccl} \mathcal{C}: & D \times D & \longrightarrow & \mathbb{R} \\ & (s_1, s_2) & \mapsto & \mathcal{C}(s_1, s_2) = \mathcal{C}ov[Y(s_1), Y(s_2)], \end{array}$$

*y además se define la función de autocorrelación como:*

$$\begin{array}{rccl} \rho: & D \times D & \longrightarrow & \mathbb{R} \\ & (s_1, s_2) & \mapsto & \rho(s_1, s_2) = \frac{\mathcal{C}ov[Y(s_1), Y(s_2)]}{\sqrt{\mathbb{V}ar[Y(s_1)]\mathbb{V}ar[Y(s_2)]}}. \end{array}$$

## 2.4 Estructuras de Correlación Espacial

### 2.4.1 Isotropía

**Definición 2.9** *Se define un campo aleatorio $\{Y(s) : s \in D\}$ como estrictamente isotrópico si y sólo sí sus distribuciones finitas conjuntas son invariantes bajo todos los movimientos rígidos. Esto es para cualquier matriz ortogonal $\boldsymbol{H}$ $d \times d$ y cualquier $s \in D$,*

$$\mathbb{P}\left(Y(\boldsymbol{H}s_1 + s) \leq t_1, \ldots, Y(\boldsymbol{H}s_n + s) \leq t_n\right) = \mathbb{P}\left(Y(s_1) \leq t_1, \ldots, Y(s_n) \leq t_n\right) \forall n \in \mathbb{N}, \forall s_1, \ldots, s_n \in \mathbb{R}^d. \quad (2.9)$$

La condición de isotropía equivale a suponer que no existe razón para distinguir una dirección de otra para el estudio del campo aleatorio en consideración.

Si un campo aleatorio no es isotrópico y mediante una transformación lineal de las coordenadas se convierte en isotrópico, entonces, se dice que el campo aleatorio es geométricamente anisotrópico.

A continuación, una condición adicional en un campo aleatorio $\{Y(s) : s \in D\}$ para que sea débilmente isotrópico o intrínsecamente estacionaria debe cumplir con la siguiente definición.

**Definición 2.10** *Suponga que $\{Y(s) : s \in D\}$ satisface (2.6) y (2.8),[así (2.8) puede re-escribir como $\mathbb{E}[Y(s_i) - Y(s_j)]^2 = 2\gamma(s_i - s_j)$]. Entonces $\{Y(s)\}$ se dice que es intrínsecamente estacionaria, o bien, satisface la hipótesis de variable intrínseca.*

**Definición 2.11** *Sea $\{Y(s), s\epsilon D\}$ un proceso espacial. Se dice que $\{Y(s)\}$ es intrínsecamente estacionario si y sólo sí: $\gamma(h)$ depende únicamente de $h$. En este caso se define el variograma como:*

$$2\gamma(h) = \mathbb{E}[Y(s+h) - Y(s)]^2,\ \forall\ s,\ s+h \in D \quad (2.10)$$

*Si* $\{Y(s)\}$ *es débilmente estacionario, entonces*

$$\begin{aligned}\mathbb{E}[Y(s+h)-Y(s)]^2 &= \mathbb{V}ar[Y(s+h)-Y(s)]+constante, &(2.11)\\ 2\gamma(h) &= \mathbb{V}ar[Y(s+h)-Y(s)], &(2.12)\\ 2\gamma(h) &= \mathbb{V}ar[Y(s+h)]+\mathbb{V}ar[Y(s)]-2\mathcal{C}ov[Y(s+h)-Y(s)]. &(2.13)\end{aligned}$$

*Ahora,*

$$\begin{aligned}\gamma(h) &= \frac{1}{2}\{2\mathcal{C}(0)-2\mathcal{C}(h)\}, &(2.14)\\ \gamma(h) &= \mathcal{C}(0)-\mathcal{C}(h). &(2.15)\end{aligned}$$

### 2.4.2 Ergodicidad

El concepto de ergodicidad está asociado a la función de covarianzas $\mathcal{C}(\cdot)$ como función del variograma $\gamma(\cdot)$. La podemos encontrar a partir de $\gamma(h)$ una vez que se conoce $\mathcal{C}(0)$. Entonces para un proceso estacionario ergódico $\{Y(s)\}$, tenemos que:

$$\lim_{\|h\|\mapsto\infty}\mathcal{C}(h)=0. \tag{2.16}$$

Lo anterior establece que la función de covarianzas decrece a medida que la separación entre las localidades aumenta en el espacio. Por lo tanto, podemos encontrar $\mathcal{C}(h)$ a partir del semivariograma

$$\begin{aligned}\lim_{\|h\|\mapsto\infty}\mathcal{C}(h) &= 0, &(2.17)\\ \mathcal{C}(h) &= \lim_{\|\mu\|\mapsto\infty}(\gamma(\mu)-\gamma(h)). &(2.18)\end{aligned}$$

### 2.4.3 Continuidad Espacial

Al momento de modelar datos espaciales necesitamos comprender las diferencias en la continuidad entre los modelos de covarianza cerca del origen y las consecuencias que esto tiene para la inferencia estadística. Un punto clave es la relación e implicancia que una función de covarianza $\mathcal{C}(\cdot)$ tiene en el suavizamiento de un proceso espacial. En este contexto, se define la continuidad en media cuadrática.

**Definición 2.12** *Sea* $\{X_n\}$ *una sucesión de variables aleatorias en* $L^2$ *y* $X$ *una variable aleatoria definida sobre el mismo espacio de probabilidad en* $(\Omega,\mathfrak{F},\mathbb{P})$*. Se sabe que* $\{X_n\}$ *converge en media cuadrática a* $X$ *si y sólo sí* $\mathbb{E}[X_n-X]^2\to 0$*,* $n\to\infty$*. Para un proceso espacial* $\{Y(s):s\in D\subset\mathbb{R}^d\}$ *con media constante y varianza constante, la continuidad en media cuadrática en* $s$ *implica que:*

$$\lim_{\|h\|\mapsto 0}\mathbb{E}\left[(Y(s+h)-Y(s))^2\right]=0, \tag{2.19}$$

*Para un proceso débilmente estacionario* $\mathbb{E}\left[(Y(s+h)-Y(s))^2\right]=2\gamma(h)$*. Así,*

$$\lim_{\|h\|\mapsto 0}\mathbb{E}\left[(Y(s+h)-Y(s))^2\right]=2(\mathcal{C}(0)-\mathcal{C}(h)), \tag{2.20}$$

*entonces el proceso es continuo en media cuadrática si* $\lim_{\|h\|\mapsto 0}\mathcal{C}(h)=\mathcal{C}(0)$*.*

En la práctica no siempre el semivariograma pasa por el origen. Cuando se está en este caso se denomina efecto pepita ver subsección 2.5.7.

## 2.5 Modelos Paramétricos para el Semivariograma

El modelado del semivariograma incluye dos etapas fundamentales, una vez construido el semivariograma experimental o empírico es necesario ajustar a este un modelo teórico, con el objetivo de determinar los parámetros descriptivos del semivariograma que posteriormente serán usados en la estimación.

### 2.5.1 Parámetros del Semivariograma

**Definición 2.13** *Dado un semivariograma $\gamma(t)$, llamaremos:*

$$\textit{pepita: nugget:} \lim_{t\to 0}\gamma(t) = \tau^2 \tag{2.21}$$

$$\textit{meseta: sill:} \lim_{t\to \infty}\gamma(t) = \tau^2+\sigma^2 \tag{2.22}$$

$$\textit{rango: range:} \min_t\{t:\gamma(t)=sill\} = \frac{1}{\phi} \tag{2.23}$$

$$\textit{meseta parcial: partial sill:} = sill - nugget = \sigma^2 \tag{2.24}$$

Los parámetros del semivariograma caracterizan tres elementos importantes en la variabilidad de un atributo georeferenciado que son: la discontinuidad en el origen (existencia de efecto de pepita), el valor máximo de variabilidad (meseta), y el área de influencia de la correlación espacial (rango) los cuales se describen a continuación.

- **Efecto Pepita (Nugget)**: Se denota por $\tau^2$ y representa una discontinuidad puntual del semivariograma en el origen ver Figura 2.1. 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.

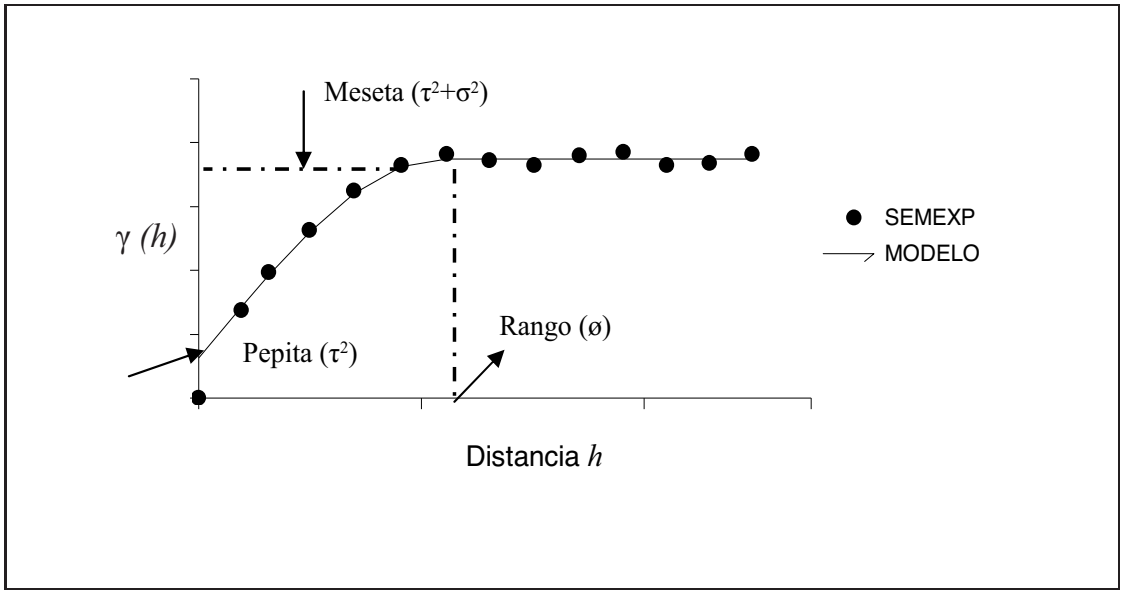


Figura 2.1: Comportamiento de un semivariograma acotado con una representación de los parámetros básicos. • SEMEXP corresponde al semivariograma experimental y $\rightharpoondown$ MODELO al ajuste de un modelo teórico.

- **Meseta (Sill)**: Es la cota superior del semivariograma. También puede definirse como el limite 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 $\sigma^2$ o por $\tau^2 + \sigma^2$ cuando el efecto pepita es diferente de cero.

- **Rango (Range)**: En términos prácticos, el rango $\phi$ corresponde a la distancia a partir de la cual dos observaciones son independientes. El rango se interpreta como la zona de influencia. 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 efectivo a la distancia para la cual el semivariograma alcanza el 95 % de la meseta. Entre más pequeño sea el rango, más cerca se esta del modelo de independencia espacial. El rango no siempre aparece de manera explícita en la fórmula del semivariograma. En el caso del modelo esférico (2.5.3), el rango coincide con el parámetro $\phi$, que se utilizará en las ecuaciones más adelante. Sin embargo, en el modelo exponencial (2.5.4), el rango efectivo es $\phi/3$ y en el modelo gaussiano (2.5.5) es $\phi/\sqrt{3}$.

### 2.5.2 Modelo lineal

El modelo lineal para modelar el semivariograma, es de la forma:

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 \cdot t & \text{si } t \geq 0,\ \tau^2, \sigma^2 > 0, \\ 0 & \text{, si } t = 0. \end{cases} \tag{2.25}$$

El modelo lineal define un modelo acotado en función de dos constantes positivas. Se debe notar que cuando $t \to \infty \Rightarrow \gamma(t) \to \infty$, por lo tanto, este proceso no es débilmente estacionario pero es intrínsecamente estacionario. Además, tanto el nugget como el sill son ambos infinitos.

$$\mathcal{C}(t) = \begin{cases} \tau^2 + \sigma^2 & \text{si } t \geq 0,\ \sigma^2 > 0, \\ 0 & \text{, si } t = 0. \end{cases} \tag{2.26}$$

### 2.5.3 Modelo Esférico

Este modelo es probablemente el más utilizado, es una expresión polinomial simple, en su forma representada en la figura 2.2, se puede observar un crecimiento casi lineal y después a cierta distancia finita del origen se alcanza una estabilización, la meseta. La tangente en el origen encuentra a la meseta en el punto de abscisa $(2/3)\phi$, donde a representa el valor del alcance. Un modelo con estructura de correlación esférica espacial, tiene la siguiente forma:

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 & \text{si } t \geq \frac{1}{\phi}, \\ \sigma^2 \left[ \frac{3}{2}\frac{t}{\phi} + \frac{1}{2}\left(\frac{t}{\phi}\right)^3 \right] & \text{si } 0 < t < \frac{1}{\phi}\ , \\ 0 & \text{, si } t = 0. \end{cases} \tag{2.27}$$

El semivariograma esférico es valido para cualquier dimensión en la que se esté trabajando.

### 2.5.4 Modelo Exponencial

Este modelo a diferencia del esférico crece inicialmente más rápido y después se estabiliza de forma asintótica. Como la meseta no se alcanza a una distancia finita, se usa con fines prácticos el *alcance efectivo* o *alcance práctico* $\tilde{\phi}$, valor que se obtiene en el punto de abscisa para el cual el modelo obtiene el 95 % de la meseta, con un valor $\tilde{\phi} = 3\phi$, donde $\phi$ es el parámetro de escala. La tangente en el origen encuentra a la meseta en el punto $\phi = 3\tilde{\phi}$. El modelo exponencial tiene la siguiente forma:

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi \cdot t)) & \text{, si } t > 0, \\ \tau^2 + \sigma^2 & \text{, si } t = 0. \end{cases} \tag{2.28}$$

La ventaja comparativa de utilizar un modelo exponencial por sobre un modelo esférico es que es mas simple en forma y también que es un variograma valido en todas las dimensiones del espacio. Cabe destacar además, que este modelo satisface la condición de ergodicidad el cual podemos pasar de una escala $\gamma$ y una $\mathcal{C}$ que será escrita en función de $\gamma$ de la siguiente manera:

$$\begin{aligned} \mathcal{C}(h) &= \lim_{u\to\infty}(\gamma(u)-\gamma(t)), && (2.29)\\ &= \tau^2+\sigma^2-[\tau^2+\sigma^2(1-\exp(1-\phi t))], && (2.30)\\ &= \sigma^2\exp(-\phi t), \forall t>0. && (2.31) \end{aligned}$$

Se tiene que

$$\mathcal{C}(t)=\begin{cases} \sigma^2\exp(-\phi t) & \text{, si } t>0,\\ \tau^2+\sigma^2 & \text{, si } t=0. \end{cases} \quad (2.32)$$

### 2.5.5 Modelo Gaussiano

Este modelo inicialmente presenta un comportamiento parabólico en el origen, después al igual que en el modelo Exponencial se alcanza la meseta de forma asintótica. El modelo gaussiano tienen la siguiente estructura:

$$\gamma(t)=\begin{cases} \tau^2+\sigma^2(1-\exp(-\phi^2t^2)) & \text{, si } t>0,\\ 0 & \text{, si } t=0. \end{cases} \quad (2.33)$$

El variograma gaussiano es una función analítica y forma una realización muy suavizada de un proceso espacial. Este modelo está definido sobre un rango efectivo $\phi$, una varianza a priori (sill) $\sigma^2$ y un efecto nugget $\tau^2$. Si se realiza el siguiente cálculo

$$\mathcal{C}(h) = \lim_{u\to\infty}(\gamma(u)-\gamma(t)). \quad (2.34)$$

Se obtiene, el siguiente resultado

$$\mathcal{C}(t)=\begin{cases} \sigma^2\exp(-\phi^2t^2) & \text{, si } t>0,\\ \tau^2+\sigma^2 & \text{, si } t=0. \end{cases} \quad (2.35)$$

### 2.5.6 Modelo de Mathérn

El modelo de Mathérn se define como sigue

$$\gamma(t)=\begin{cases} \tau^2+\sigma^2\left[1-\frac{(2\sqrt{\nu}t\phi)}{2^{\nu-1}\Gamma(\nu)}K_\nu(2\sqrt{\nu}t\phi)\right] & \text{, si } t>0,\\ \tau^2 & \text{, si } t=0. \end{cases} \quad (2.36)$$

Esta clase de modelos fue originalmente sugerida por Mathérn (1960). Sus elementos son; $\nu>0$, que denota la suavidad del proceso, $\phi$ es el parámetro de escala espacial, $\Gamma(\nu)$ es la función gamma usual, $K_\nu$ es la función de Bessel modificada por el escalar $\nu$.

En el modelo si $\nu=1/2$, entonces el variograma corresponde a un modelo exponencial. Si $\nu\to\infty$ en el modelo de Mathérn entonces el variograma corresponde a un modelo gaussiano.

### 2.5.7 Modelo de Independencia (Efecto Pepita Puro)

Es indicativo de carencia de correlación espacial entre las observaciones de una variable espacial. Es común sumar este modelo a otro modelo teórico del semivariograma, para obtener lo que se conoce como semivariograma anidado. Lo anterior se sustenta en una propiedad de los semivariogramas que dice

que cualquier combinación lineal de semivariogramas con coeficientes positivos es un semivariograma. Su expresión matemática es:

$$\gamma(t) = \begin{cases} 0 & , \text{si } t = 0, \\ \tau^2 & , \text{si } t \geq 0. \end{cases} \tag{2.37}$$

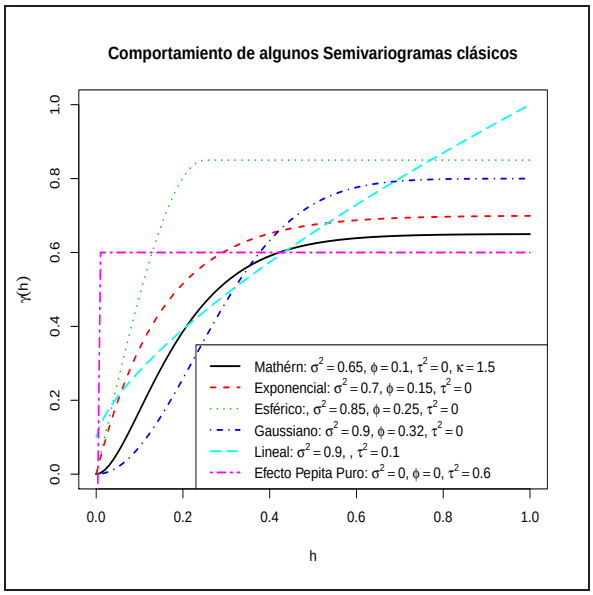


Figura 2.2: Comparación de los modelos de Mathérn, Exponencial, Esférico, Gaussiano, Lineal y Pepita Puro, respecto a una escala simulada entre 0 y 1.

En la figura 2.2, se presentan las formas que tiene cada semivariograma. Cada linea representa un semivariograma con sus respectivos parámetros.

## 2.6 Estimación Muestral del Semivariograma

### 2.6.1 Estimador de Matheron

Para analizar el semivariograma se parte desde un conjunto de datos espaciales observados, digamos $Y(s_1), \ldots, Y(s_n)$. Una manera de obtener una buena información sobre el conjunto de datos, es graficando las diferencias al cuadrado $\{Y(s_i) - Y(s_j)\}^2$ contra el retraso de la distancia $h$ (ó $\|h\|$). El estimador del semivariograma de los promedios de los cuadrados de las diferencias de puntos que se distancian $s_i - s_j = h$, se conoce comúnmente como el estimador clásico o Matheron desde que fue propuesto por Matheron (1962):

$$\gamma(h) = \frac{1}{2|N(h)|} \sum_{N(h)} [Y(s_i) - Y(s_j)]^2. \tag{2.38}$$

donde el conjunto $N(h)$ consiste en los pares de las ubicaciones $(s_i, s_j)$ de tal manera que $s_i - s_j = h$ y $|N(h)|$ denota el número de pares distintos en $N(h)$. Las recomendaciones típicas son que por lo menos 30 (más 50) pares de localidades deberían estar disponibles en cada intervalo. Una de las desventajas del semivariograma clásico es su sensibilidad a datos u observaciones atípicas en los datos.

Una de las bondades que este estimador tiene, es su fácil implementación computacional, su uniformidad y que alcanza el cero en la dirección cero, esto es: $\widehat{\gamma}(h) = \widehat{\gamma}(-h)$, $\widehat{\gamma}(h) = 0$.

Como el estimador de Matheron es un estimador basado en el cuadrado de la diferencia de las observaciones, muchos investigadores han tratado de establecer o aproximar de manera asintótica los momentos y

la distribución de este estimador. Es así, como Cressie (1985) estableció una aproximación para la varianza del estimador de modo que considerando $\{Y(s)\}$ un proceso gaussiano se tiene que:

Considere $\{Y(s)\}$ un proceso tal que $(2\gamma(h))^{-1}\{Y(s)-Y(s+h)\}^2 \sim \chi^2_1$ y $Var[\{Y(s)-Y(s+h)\}^2] = 2\times 4\gamma(h)^2$. Cressie (1985) muestra que la varianza con retardo $h_i$ puede ser aproximada como

$$\mathbb{V}ar(\hat{\gamma}(h_i)) \approx \frac{2\cdot\gamma(h_i)^2}{|N(h_i)|}. \tag{2.39}$$

La aproximación ignora las correlaciones entre $Y(s_i)-Y(s_j)$ y $Y(s_k)-Y(s_l)$. Si el estimador de Matheron es consistente, entonces $\hat{\gamma}(h)$ es un estimador insesgado de $\gamma(h)$.

## 2.7 Estimación de los Parámetros del Semivariograma

### 2.7.1 Estimación por Mínimos Cuadrados

Los principios de la geométrica de mínimos cuadrados nos permiten estimar los parámetros en un modelo que describe la media de un vector aleatorio, teniendo en cuenta la variación y covariación de los elementos del vector. Para aplicar la estimación de mínimos cuadrados y para modelar el semivariograma, la media de la *respuesta* tiene que ser modelada en función del semivariograma. Considere la posibilidad de un estimador semivariograma empírico en $k$ rezagos. Por ejemplo, un modelo de semivariograma $\hat{\gamma}(h)$ puede ser representado por los pseudo-datos:

$$\hat{\gamma}(\mathbf{h}) = [\hat{\gamma}(h_1), \hat{\gamma}(h_2), \ldots, \hat{\gamma}(h_k)]^t, \tag{2.40}$$

o

$$\tilde{\gamma}(\mathbf{h}) = [\tilde{\gamma}(h_1), \tilde{\gamma}(h_2), \ldots, \tilde{\gamma}(h_k)]^t. \tag{2.41}$$

u otro estimador empírico.

$$\widehat{\theta}_{OLS} = argmin_{\alpha\in\Theta}\sum_{i=1}^{k}\left(\widehat{\gamma}_n(h_i)-\gamma(h_i;\theta)\right)^2. \tag{2.42}$$

Los pasos necesarios en la derivación se pueden repetir para otros estimadores. El método de los mínimos cuadrados no exige supuestos distribucionales sobre $\hat{\gamma}(h)$ aparte de la existencia del primer y segundo momento. Considere un modelo estadístico de la forma:

$$\hat{\gamma}(h) = \gamma(h,\theta) + \epsilon(h), \tag{2.43}$$

donde $\epsilon(h) = [\epsilon(h_1), \epsilon(h_2), \ldots, \epsilon(h_k)]^t$, es asumido como un vector de errores $k\times 1$ con media 0 y matriz de covarianza de errores, $\mathbb{V}ar(\epsilon(h)) = R$. Vamos a escribir $R(\theta)$ si es necesario para enfatizar la dependencia explícita de $\theta$.

$$(\hat{\gamma}(h)-\gamma(h,\theta))^t R^{-1}(\theta)(\hat{\gamma}(h)-\gamma(h,\theta)). \tag{2.44}$$

Si $R$ no depende de $\theta$, entonces la suma de cuadrados generalizados se convierte en un problema de mínimos cuadrados ordinarios no lineal, y por lo tanto se minimizará

$$(\hat{\gamma}(\mathbf{h})-\gamma(h,\theta))^t(\hat{\gamma}(\mathbf{h})-\gamma(h,\theta)), \tag{2.45}$$

que se resuelve iterativamente. En el otro caso, realizamos un proceso iterativo que permita actualizar las estimaciones de $\hat{\theta}$, de tal manera que $R(\hat{\theta})$ continúe actualizándose. Ahora bien; la dificultad de minimizar la suma de cuadrados generalizados no está en la presencia de una matriz de ponderaciones, sino que yace

en la obtención de la matriz $R$. En Cressie (1993) se encuentran los principios básicos.
La aproximación de Mínimos Cuadrados Ponderados (WLS) ajustan un semivariograma reemplazando $R(\theta)$ por una matriz diagonal $W(\theta)$ cuyos valores están dados por $\mathbb{V}ar(\hat{\gamma}(h_m))$, de esta manera minimizamos la suma de cuadrados ponderados.

$$(\hat{\gamma}(\mathbf{h}) - \gamma(h,\theta))^t W^{-1}(\theta)(\hat{\gamma}(\mathbf{h}) - \gamma(h,\theta)), \tag{2.46}$$

$$= \sum_{m=1}^{k} \frac{|N(h_m)|}{2\gamma(h_m,\theta)^2}(\hat{\gamma}(h_m) - \gamma(h_m,\theta))^2. \tag{2.47}$$

Así, esta expresión la podemos escribir como la suma de cuadrados ponderados sobre $k$ localidades.

### 2.7.2 Estimación ML

Para estimar parámetros de un campo aleatorio espacial por el método de Máxima Verosimilitud necesitamos conocer la distribución espacial de $\{Y(s)\}$ y este se utiliza sólo para el caso en que la distribución es Gaussiana.

Denotemos el vector de observaciones $Y = (Y(s_1), Y(s_2), \ldots, Y(s_n))^t$ y asumimos que $Y(s) \sim N(\mu\mathbf{1}, \Sigma(\theta))$. La matriz de covarianza de $\{Y(s)\}$ ha sido parametrizada de tal forma que cualquier estimación de $\hat{\theta}$ la varianza y covarianza pueda ser estimada como $\hat{\mathbb{V}ar}(Y(S)) = \Sigma(\hat{\theta})$. La expresión para la log-verosimilitud es

$$\varphi(\mu;\theta;Y(s)) = \log_e|\Sigma(\theta)| + n\log_e 2\pi + (Y(s) - \mu\mathbf{1})^t\Sigma^{-1}(\theta)(Y(s) - \mu\mathbf{1}), \tag{2.48}$$

y es minimizada con respecto a $\mu$ y $\theta$ . Si $\theta$ es conocido, el mínimo de la expresión puede ser desarrollado como

$$\tilde{\mu} = (\mathbf{1}^t\Sigma^{-1}(\theta)\mathbf{1})^t\mathbf{1}^t\Sigma^{-1}(\theta)Y(s). \tag{2.49}$$

Que es el estimador mínimo cuadrático generalizado para $\mu$. El estimador máximo verosímil tiene importantes propiedades como por ejemplo, bajo condiciones de regularidad, o sea, para un proceso $Y(s_i)$ que se distribuye aproximadamente Normal, este tiene una distribución asintóticamente Gaussiana y es un estimador eficiente. Sin embargo, para muestras de tamaño $n$ finitas a menudo este estimador es sesgado.

### 2.7.3 Estimación REML

La idea de la estimación Máximo Verosímil Restringida ó Máximo Verosímil Residual (REML) es estimar los parámetros de varianza y covarianza $\hat{\theta}$ para maximizar la verosimilitud de $KY(s)$ en vez de maximizar la verosimilitud de $Y(s)$, así se elige la matriz $K$, que es llamada matriz de contrastes residuales, de modo que $\mathbb{E}[KY(s)] = 0$.
La estimación REML se realiza sólo para el caso en que la media de $\{Y(s_i)\}$ es una función lineal debido a que no es muy evidente obtener la matriz $\phi$ en otros casos. Esta matriz no es única, no obstante, para la estimación e inferencia de los parámetros esto no es de mucha importancia. Si $\mathbb{E}[Y(s)] = \mu$ una elección de la matriz $K_{(n-1)\times n}$ podría ser

$$K = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \ldots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \ldots & -\frac{1}{n} \\ \ldots & \ldots & \ldots & \ldots \\ -\frac{1}{n} & -\frac{1}{n} & \ldots & 1-\frac{1}{n} \end{pmatrix}$$

entonces $KY(s)$ es un vector $(n-1)\times 1$ de diferencias entre las observaciones y la media muestral. Así, $KY(s) \sim N(\mathbf{0}, K\Sigma(\theta)K^t)$ y la estimación máximo verosímil restringida son los valores de $\theta$ que minimizan la función

$$\varphi_R(\theta;KY(s)) = \log_e|K\Sigma(\theta)K^t| + (n-1)\log_e 2\pi + Y(s)^t(K\Sigma(\theta)K^t)^{-1}KY(s), \tag{2.50}$$

De manera similar al perfilamiento de la máxima verosimilitud, esta expresión no provee información sobre la media $\mu$. En la estimación Máximo Verosímil $\mu$ fue perfilado fuera de la log-verosimilitud, mientras que en la estimación REML la verosimilitud fue maximizada desde un conjunto de datos diferentes formados por $KY(s)$ en vez de $Y(s)$ que son los datos originales. Por lo tanto, estrictamente este no es un estimador máximo verosímil de $\mu$. En cambio, el estimador obtenido al evaluar en el estimador REML, $\hat{\theta}_{reml}$ será un estimador mínimo cuadrático generalizado estimado

$$\hat{\mu} = (\mathbf{1}^t \Sigma^{-1}(\hat{\theta}_{reml})\mathbf{1})^t \mathbf{1}^t \Sigma^{-1}(\hat{\theta}_{reml}) Y(s). \tag{2.51}$$

## 2.8 Criterios de Selección de Modelos

Para seleccionar un modelo, existen varios métodos que son expresados respecto al error muestral entre la observación y la estimación, el objetivo es determinar una cantidad que nos diga que modelo se ajusta mejor respecto a un conjunto de modelos propuestos. Debe de existir un compromiso entre la bondad de ajuste y la complejidad del modelo. El Criterio de Información de Akaike (AIC) puede ayudar en nuestra selección, el cual se define como

$$AIC = -2ln(\text{máx.verosimilitud}) + 2(\text{núm.deparámetros}) \tag{2.52}$$

O bien,

$$AIC = n \cdot ln(R) + 2p \tag{2.53}$$

que es un estimador simplificado del Criterio de Información de Akaike. Donde $n$ es el numero de valores estimados $\{\hat{\gamma}(h_i), i = 1, \ldots, n\}$ del variograma muestral, $R$ es la suma residual de los cuadrados de las diferencias entre los valores experimentales $\hat{\gamma}(h_i)$ y los del modelo ajustado $\gamma(h_i)$, es decir $R = \sum_{i=1}^{n} (\gamma(h_i) - \hat{\gamma}(h_i))^2$, mientras que $p$ es el número (fijo) de parámetros del modelo de variograma ajustado $\gamma(h)$. Se considera que el modelo que presenta el menor AIC es el mejor.

# Parte II

# Fundamentos

Capítulo 3

# Tamaño Muestral Efectivo: Algunas Propuestas

## 3.1 Introducción

Para abordar el problema de la estimación del tamaño muestral efectivo en el modelamiento de variables espaciales hay que considerar previamente la estructura de correlación espacial presente en los datos, la cual puede ser representados mediante modelos paramétricos. En este capítulo se estudiará como la literatura ha tratado de dar respuesta a esta problemática, y en cuyo caso es determinar una expresión para al cálculo del tamaño muestral efectivo y como evaluar los impactos de la correlación espacial.

## 3.2 Efecto de la Correlación en el Tamaño Muestral Efectivo

Una primera visión se puede ver en Cressie (1993) donde considera el siguiente escenario. Supongamos que se observa una muestra de tamaño $n$, es decir, $Y_1, \ldots, Y_n$ que son independientes e idénticamente distribuidos ($i.i.d$) con distribución gaussiana de media $\mu$ desconocida y varianza $\sigma_0^2$ desconocida. El estimador insesgado de mínima varianza para $\mu$ es:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y(s_i). \tag{3.1}$$

en base a (3.1) se puede realizar una inferencia sencilla para $\mu$. Para esto supongamos que el estimador $\overline{Y}$ se distribuye $N\left(\mu, \frac{\sigma_0^2}{n}\right)$. Entonces, un intervalo de confianza al 95 % para $\mu$ es:

$$\left[\overline{Y} - (1.96)\sigma_0/\sqrt{n}, \overline{Y} + (1.96)\sigma_0/\sqrt{n}\right]. \tag{3.2}$$

Ahora bien, en lugar de una muestra independiente supongamos se tiene una muestra de tamaño $n$, es decir, $Y(s_1), \ldots, Y(s_n)$, pero ahora los datos están correlacionados positivamente con una correlación que disminuirá a medida que aumenta la separación entre los datos, es decir:

$$\mathbb{C}ov(Y(s_i), Y(s_j)) = \sigma_0^2 \cdot \rho^{|i-j|},\ i, j = 1, \ldots, n.\ 0 < \rho < 1. \tag{3.3}$$

Se sabe que tal correlación es el resultado de la función de correlación de un proceso autoregresiva de primercorden. Sin embargo, este mecanismo no es el que se tiene en mente. En el contexto de variables espaciales $Y(s_i)$ son datos espaciales en $\mathbb{R}^1$, por lo que la predicción de $Y(0)$ o $Y(3/2)$ son tan apropiadas como la de $Y(n+1)$.

Ahora se calcula la varianza de la media muestral con la estructura de correlación autoregresiva de primer orden, esto es;

$$\begin{aligned}\mathbb{V}ar\left(\overline{Y}\right) &= \mathbb{V}ar\left(\frac{1}{n}\sum_{i=1}^{n}Y(s_i)\right) = n^{-2}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{C}ov(Y(s_i),Y(s_j))\right\}, &(3.4)\\ &= \left\{\frac{\sigma_0^2}{n}\right\}\left[1+2\left(\frac{\rho}{1-\rho}\right)\left(1-\frac{1}{n}\right)-2\left(\frac{\rho}{1-\rho}\right)^2\left(\frac{1-\rho^{n-1}}{n}\right)\right]. &(3.5)\end{aligned}$$

La demostración de (3.5) se presenta a continuación

### 3.2.1 Demostración

Sean $Y(s_i) = Y_i$ e $Y(s_j) = Y_j$ dos procesos autoregresivos de orden 1, representados de la siguiente forma:

$$Y_i = \rho Y_{i-1} + \epsilon_i, \epsilon_i \sim N(0, \sigma_0^2(1-\rho^2)), \quad (3.6)$$
$$Y_j = \rho Y_{j-1} + \epsilon_j, \epsilon_j \sim N(0, \sigma_0^2(1-\rho^2)). \quad (3.7)$$

Si $Y_i$ e $Y_j$ son procesos invertibles, entonces la representación causal de (3.6) y (3.7) son respectivamente

$$Y_i = \sum_{k=0}^{\infty}\rho^k\epsilon_{i-k}, \quad (3.8)$$
$$Y_j = \sum_{l=0}^{\infty}\rho^k\epsilon_{j-l}. \quad (3.9)$$

Se quiere demostrar que

$$\mathbb{V}ar\left(\overline{Y}\right) = \left\{\frac{\sigma_0^2}{n}\right\}\left[1+2\left(\frac{\rho}{1-\rho}\right)\left(1-\frac{1}{n}\right)-2\left(\frac{\rho}{1-\rho}\right)^2\left(\frac{1-\rho^{n-1}}{n}\right)\right].$$

Primero se calcularan la covarianza entre $Y_i$ e $Y_j$ considerando las expresiones (3.8) y (3.9) respectivamente, es decir:

$$\begin{aligned}\mathbb{C}ov(Y_i,Y_j) &= \mathbb{C}ov\left(\sum_{k=0}^{\infty}\rho^k\epsilon_{i-k}, \sum_{l=0}^{\infty}\rho^k\epsilon_{j-l}\right) = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\rho^k\rho^l\mathbb{C}ov(\epsilon_{i-k},\epsilon_{j-l}),\\ &= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\rho^{k+l}\mathbb{C}ov(\epsilon_{i-k},\epsilon_{j-l}), \text{ si } i-k=j-l \Rightarrow k=i-j+l, \text{ entonces,}\\ &= \sum_{l=0}^{\infty}\rho^{i-j+2l}\mathbb{C}ov(\epsilon_{j-l},\epsilon_{j-l}),\\ &= \rho^{i-j}\sum_{l=0}^{\infty}\rho^{2l}\mathbb{C}ov(\epsilon_{j-l},\epsilon_{j-l}),\\ &= \rho^{|i-j|}\frac{\sigma_0^2(1-\rho^2)}{(1-\rho^2)}, \ \rho^{i-j}=\rho^{|i-j|}, \text{ matriz simétrica,}\\ &= \rho^{|i-j|}\sigma_0^2.\end{aligned}$$

Entonces, la expresión (3.4) la podemos escribir de la siguiente forma

$$\mathbb{V}ar\left(\overline{Y}\right) = n^{-2}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{C}ov(Y(s_i),Y(s_j))\right\},$$

$$= \quad n^{-2}\sigma_0^2 \sum_{i=1}^{n}\sum_{j=1}^{n} \rho^{|i-j|}. \tag{3.10}$$

En la expresión (3.10) se tiene que calcular $\sum_{i=1}^{n}\sum_{j=1}^{n}\rho^{|i-j|}$. Ahora bien, si se desarrolla esta sumatoria de la siguiente manera se puede obtener el siguiente patrón

| | $j=1$ | $j=2$ | $j=3$ | $j=4$ | $\ldots$ | $j=n$ |
|---|---|---|---|---|---|---|
| $i=1$ | $1+$ | $\rho+$ | $\rho^2+$ | $\rho^3+$ | $\ldots$ | $+\rho^{n-1}+$ |
| $i=2$ | $\rho+$ | $1+$ | $\rho+$ | $\rho^2+$ | $\ldots$ | $+\rho^{n-2}+$ |
| $i=3$ | $\rho^2+$ | $\rho+$ | $1+$ | $\rho+$ | $\ldots$ | $+\rho^{n-3}+$ |
| $i=4$ | $\rho^3+$ | $\rho^2+$ | $\rho+$ | $1+$ | $\ldots$ | $+\rho^{n-4}+$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $i=n$ | $\rho^{n-1}+$ | $\rho^{n-2}+$ | $+\rho^{n-3}+$ | $\rho^{n-4}+$ | $\ldots$ | $+1$ |

Entonces, lo anterior se puede generalizar de la siguiente forma

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n}\rho^{|i-j|} &= n\cdot 1 + 2(n-1)\rho + 2(n-2)\rho^2 + 2(n-3)\rho^3 + \ldots + 2(n-(n-1))\rho^{n-1},\\
&= n + 2[(n-1)\rho + (n-2)\rho^2 + (n-3)\rho^3 + \ldots + (n-(n-1))\rho^{n-1}],\\
&= n + 2\left[\sum_{i=1}^{n-1}(n-i)\rho^i\right],\\
&= n + 2\left[\sum_{i=1}^{n-1}n\rho^i - \sum_{i=1}^{n-1}i\rho^i\right].
\end{aligned}$$

Desarrollando la expresión (3.11) se obtiene

$$\begin{aligned}
\sum_{i=1}^{n-1}(n-i)\rho^i &= \sum_{i=1}^{n-1}n\rho^i - \sum_{i=1}^{n-1}i\rho^i,\\
&= n\sum_{i=1}^{n-1}\rho^i - \rho\sum_{i=1}^{n-1}i\rho^{i-1},\\
&= n\sum_{i=1}^{n-1}\rho^i - \rho\sum_{i=1}^{n-1}\frac{\partial}{\partial\rho}\rho^i,\\
&= n\sum_{i=1}^{n-1}\rho^i - \rho\frac{\partial}{\partial\rho}\sum_{i=1}^{n-1}\rho^i,\\
&= \frac{n\rho(1-\rho^{n-1})}{1-\rho} - \rho\frac{\partial}{\partial\rho}\left(\frac{\rho(1-\rho^{n-1})}{1-\rho}\right),\\
&= \frac{n\rho(1-\rho^{n-1})}{1-\rho} - \frac{\rho[(n-1)\rho^n - n\rho^{n-1}+1]}{(1-\rho)^2},\\
&= \frac{n\rho(1-\rho^{n-1})(1-\rho) - [(n-1)\rho^{n+1} - n\rho^n + \rho]}{(1-\rho)^2},\\
&= \frac{n\rho - n\rho^2 + \rho^{n+1} - \rho}{(1-\rho)^2}, \text{ sumamos un cero, es decir, } (\rho^2-\rho^2),\\
&= \frac{n\rho - n\rho^2 + \rho^{n+1} - \rho + (\rho^2-\rho^2)}{(1-\rho)^2},
\end{aligned}$$

$$
\begin{aligned}
&= \frac{n\rho - n\rho^2 - \rho + \rho^2 + \rho^{n+1}}{(1-\rho)^2}, \\
&= \frac{n\rho(1-\rho) - \rho(1-\rho) - \rho^2(1-\rho^{n-2})}{(1-\rho)^2}, \\
&= \frac{(n\rho - \rho)(1-\rho) - \rho^2(1-\rho^{n-1})}{(1-\rho)^2}, \\
&= \frac{\rho(n-1)(1-\rho) - \rho^2(1-\rho^{n-1})}{(1-\rho)^2}, \\
&= \left(\frac{\rho}{1-\rho}\right)(n-1) - \left(\frac{\rho}{1-\rho}\right)^2 (1-\rho^{n-1}).
\end{aligned}
$$

Finalmente, reuniendo los términos anteriores se tiene que

$$
\begin{aligned}
\mathbb{V}ar(\overline{Y}) &= \frac{\sigma_0^2}{n^2}\left\{n + 2\left[\left(\frac{\rho}{1-\rho}\right)(n-1) - \left(\frac{\rho}{1-\rho}\right)^2(1-\rho^{n-1})\right]\right\}, \\
&= \left\{\frac{\sigma_0^2}{n}\right\}\left[1 + 2\left(\frac{\rho}{1-\rho}\right)\left(1 - \frac{1}{n}\right) - 2\left(\frac{\rho}{1-\rho}\right)^2\left(\frac{1-\rho^{n-1}}{n}\right)\right].
\end{aligned}
$$

Ahora continuando con el ejemplo, consideremos los siguientes valores, $n = 10$ y $\rho = 0.26$, entonces $\mathbb{V}ar(\overline{Y}) = \left[\frac{\sigma_0^2}{10}\right](1.608)$ y un intervalo de confianza del 95 % para $\mu$ es:

$$
\left[\overline{Y} - (2.485)\sigma_0/\sqrt{10}, \overline{Y} + (2.485)\sigma_0/\sqrt{10}\right]. \tag{3.11}
$$

Por lo tanto, la presencia de autocorrelación positiva en los datos conduce a intervalos de confianza que son demasiados estrechos, para $n = 10$ y $\rho = 0.26$, la probabilidad de cobertura es del 87.8 % y no del 95 %. [Generalmente, es más común encontrar dependencia espacial positiva, es decir, $\rho > 0$]. La comprensión intuitiva del efecto de la correlación espacial puede ser obtenida de la siguiente relación:

$$
\mathbb{V}ar\left(\overline{Y}\right) = \frac{\sigma_0^2}{n'}. \tag{3.12}
$$

la expresión (3.5) del denominador fue escalada y la denotamos por $n'$ donde

$$
n' = \frac{n}{\left[1 + 2\left(\frac{\rho}{1-\rho}\right)\left(1 - \frac{1}{n}\right) - 2\left(\frac{\rho}{1-\rho}\right)^2\left(\frac{1-\rho^{n-1}}{n}\right)\right]}, 0 < \rho < 1. \tag{3.13}
$$

La expresión (3.5) puede ser interpretado como el número de observaciones independientes. Por ejemplo, Si $n = 10$ y $\rho = 0.26$, entonces $n' = 6.2$, el cual se puede interpretar como que se necesitan 6 observaciones independientes para alcanzar la misma precisión que 10 observaciones correlacionadas. Para muestras grandes, $n' \approx \frac{n(1-\rho)}{1+\rho}$, está expresión muestra que la correlación tiene un efecto incluso en muestras grandes. Los modelos espaciales, más complicados presentan el mismo comportamiento en general. Haining (1988) considera modelos Gaussianos con media constante en $\mathbb{R}^2$ que son especificados como modelos autoregresivos condicionales (CAR), simultáneos autoregresivos (SAR) y medias móviles cada una con un parámetro de varianza $\sigma^2$ desconocida. El compara la varianza de $\overline{Y}$ asumiendo independencia, con la varianza de $\overline{Y}$ asumiendo dependencia positiva; también se compara en este último con el estimador máximo verosímil $\hat{\mu}$ de la media poblacional $\mu$. En general, la inferencia clásica esta basada sobre $\overline{Y}$ y $\frac{\sigma^2}{n}$ es engañosa; para dependencia espacial positiva $\widehat{\mathbb{V}ar}(\overline{Y})$ y $\widehat{\mathbb{V}ar}(\bar{\mu})$ son típicamente más grande que $\frac{\hat{\sigma}^2}{n}$.

Para una clase de modelos de Series de tiempo que incluyen correlación intraclase el estimador máximo verosímil de $\mu$ es:

$$
\hat{\mu} \equiv \frac{Y_1 + (1-\rho)\sum_{i=2}^{n-1} Y_i + Y_n}{n - (n-2)\rho}. \tag{3.14}
$$

el cual es un estimador insesgado con varianza mínima. Por otro lado, la expresión (3.14) muestra claramente el impacto de la correlación espacial en el tamaño de la muestra para estimar $\mu$.

## 3.3 El Tamaño Muestral Efectivo en una Muestra Geográfica

Paralela y alternativamente Griffith (2005) propuso una medida basada en el siguiente modelo:

$$\mathbb{Y}_{n\times 1} = \mu_{1\times 1}\mathbf{1}_{n\times 1} + \epsilon_{n\times 1}. \tag{3.15}$$

donde $\mathbb{Y}_{n\times 1}$, es la variable que contiene atributos georeferenciados (o bien, una muestra con datos espaciales), $\mu_{1\times 1}$ (constante) es la media general de los datos georeferenciados y $\epsilon_{n\times 1}$ es la perturbación aleatoria. Una base para establecer el tamaño muestral efectivo está en la distribución muestral de los datos georeferenciados que se presentan en términos de la media muestral, que son múltiples extensiones para la exploración de la muestra o el coeficiente de correlación. Este enfoque se basa en el supuesto que los datos espaciales provienen de una distribución normal.

Para la medición natural de la variabilidad en algunos fenómenos georeferenciados, se puede usar la estimación de la inflación de varianza cuando la autocorrelación se pasa por alto (ver Haining 2003, § 8.1). Ahora bien, supongamos que la matriz $V$, $n\times n$ es la matriz de covarianza que contiene todas las covariaciones entre la estructura de las $n$ observaciones georeferenciadas. ( Más precisamente, se puede decir que $\sigma^2 V$ es la matriz de covarianza) de tal manera que $\mathbb{Y} = \mu\mathbf{1} + \epsilon$ se puede re-escribir como un modelo que tiene perturbaciones aleatorias independientes, es decir, $\mathbb{Y} = \mu\mathbf{1} + V^{-1/2}\epsilon = \mu\mathbf{1} + \epsilon^*$, donde $\epsilon^* = V^{-1/2}\epsilon$ vector de perturbaciones sin correlación. Ahora bien, supongamos $\epsilon^*$ es una muestra independiente e idénticamente distribuida $N(0, \sigma_\epsilon^2)$, donde $\sigma_\epsilon^2$ denota la varianza poblacional de las covariaciones de $\epsilon^*$. Cuando $\Sigma = \mathbb{I}$, con $\mathbb{I}$ la matrix identidad $n \times n$, representa un conjunto de $n$ observaciones independientes. Usando la notación matricial, la estimación de la varianza poblacional basada en la muestra $\mathbb{Y}$ e ignorando la autocorrelación espacial está dada por:

$$\mathbb{E}\left(\hat{\sigma}_Y^2\right) = \mathbb{E}\left[\frac{(\mathbb{Y} - \hat{\mu}1)^t(\mathbb{Y} - \hat{\mu}1)}{n}\right] = \frac{Tr(V^{-1})}{n}\sigma_{\epsilon^*}^2. \tag{3.16}$$

donde $\hat{\sigma}_Y^2$ denota la estimación de la varianza de la muestra $\mathbb{Y}$, denotado por $\sigma_Y^2$ y $Tr(\cdot)$ denota el operador de la traza. La cantidad $\frac{Tr(V^{-1})}{n}$ es el factor de inflación de la varianza (VIF) similar al generado por la multicolinealidad en los modelos de regresión lineal múltiple; esta cantidad expresa el grado de colinealidad entre las observaciones georefenciados y la precisión de $\mathbb{Y}$ con respecto a la dispersión espacial.

Una vez más usando la notación de matrices, se puede determinar que la media muestral es $\hat{\mu} = (\mathbf{1}^t\mathbf{1})^{-1}\mathbf{1}^t\mathbb{Y} = \frac{1}{n}\mathbf{1}^t\mathbb{Y}$ y al calcular la varianza poblacional de la variable $\mathbb{Y}$, haciendo omisión de la autocorrelación espacial se puede obtener que (Ver Rencher, (2002)):

$$\mathbb{V}ar\left(\hat{\mu}\right) = \frac{\mathbf{1}^t V^{-1}\mathbf{1}}{n^2}\sigma_{\epsilon^*}^2. \tag{3.17}$$

Reorganizando las condiciones del lado derecho de esta ecuación y tomando en cuenta los arreglos algebraicos necesarias nos da:

$$\mathbb{V}ar\left(\hat{\mu}\right) = \frac{\frac{Tr(V^{-1})}{n}\sigma_{\epsilon^*}^2}{\frac{Tr(V^{-1})}{\mathbf{1}^t V^{-1}\mathbf{1}}n}. \tag{3.18}$$

El denominador en el lado derecho de esta ecuación proporciona la formula para el tamaño muestral efectivo a saber,

$$n^* = \frac{Tr(V^{-1})}{\mathbf{1}^t V^{-1}\mathbf{1}}n. \tag{3.19}$$

Si las $n$ observaciones son independientes la matriz de correlación queda $V = \mathbb{I}$, entonces $n^* = n$, y la VIF se convierte en $\frac{Tr(V^{-1})}{n} = 1$. Si la autocorrelación espacial perfecta prevalece, pues conceptualmente, $V^{-1} = k\mathbf{1}\mathbf{1}^t$, con $k \to \infty$, como la autocorrelación aumenta $n^* = 1$. Note que la cantidad $Tr(V^{-1})$ es un factor de inflación de la varianza para medir la multicolinealidad en modelos de regresión multiple.

La relación entre el factor de inflación de varianza y el tamaño muestral efectivo se discute a continuación.

### 3.3.1 Factor de Inflación de la Varianza y el Tamaño Muestral Efectivo

Sea $\mathbb{Y}_{n\times 1} = \mathbb{X}_{n\times p}\beta_{p\times 1}+\epsilon_{n\times 1}$ un modelo de regresión lineal multiple con $\epsilon_{n\times 1} \sim N_n(0, \sigma^2\mathbb{I})$. El estimador máximo verosímil de $\beta$ es $\widehat{\beta} = (\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\mathbb{Y}$.

Un método formal para detectar la presencia de multicolinealidad es por medio del Factor de Inflación de la Varianza. Estos factores miden cuan infladas están la varianzas de los coeficientes de regresión con respecto a cuando una de las variables independientes no se relacionan linealmente.

Para entender el significado de los factores de inflación de la varianza, comencemos con la presición de la estimación de los coeficientes de regresión mediante mínimos cuadrados con respecto a sus parámetros poblacionales. Se sabe que

$$\mathbb{V}ar(\widehat{\beta}) = \sigma^2(\mathbb{X}^t\mathbb{X})^{-1}. \tag{3.20}$$

Para reducir los errores de redondeo en el cálculo de $(\mathbb{X}^t\mathbb{X})^{-1}$, es conveniente transformar las variables por medio de la transformación de correlación. En el modelo transformado, los coeficientes $\widehat{\beta}_k$ son normalizados. La matriz de covarianza de los coeficientes de regresión estandarizados estimados son:

$$\mathbb{V}ar(\widehat{\beta}) = (\sigma^{'})^2 R^{-1}_{\mathbb{XX}}. \tag{3.21}$$

Donde $R^{-1}_{\mathbb{XX}}$ es la matriz de los coeficientes de correlación simple entre pares de las variables independientes, como se ilustra en las $p-1=2$ variables independientes y $(\sigma^{'})^2$ es la varianza del error para el modelo transformado. Note (3.21) que la variación de $\widehat{\beta}_k$, $k = 1,\ldots,p-1$ es igual al producto del término del error de la varianza $(\sigma^{'})^2$ y el $k$-ésimo elemento de la diagonal de la matriz $R^{-1}_{\mathbb{XX}}$. Este segundo factor se llama factor de inflación de la varianza VIF. Se puede demostrar que el factor de varianza es denotada por $(VIF)_k = (1-R^2_k)^{-1}$, $k = 1,\ldots,p-1$ donde $R^2_k$ es el coeficiente de determinación múltiple cuando $X_k$ es regresado sobre otra $p-2$ variables en el modelo, es decir, $R^2_k = R^2_{X_k|X_1,\ldots,X_{k-1},X_{k+1},\ldots,X_n}$ con $k = 1,\ldots,p$. Desde esto último se tiene:

$$\mathbb{V}ar(\beta_k) = (\sigma^{'})^2(VIF)_k = \frac{(\sigma^{'})^2}{1-R^2_k}. \tag{3.22}$$

Se puede mostrar que la esperanza de la suma del error cuadrático es:

$$\mathbb{E}\left[\sum_{k=1}^{p-1}(\widehat{\beta}_k-\beta_k)^2\right] = \sum_{k=1}^{p-1}\mathbb{V}ar(\beta_k) = (\sigma^{'})^2\sum_{k=1}^{p-1}(VIF)_k = (\sigma^{'})^2 Tr(R^{-1}_{\mathbb{XX}}). \tag{3.23}$$

Por lo tanto, los valores resultantes de $(VIF)_k$ son en promedio las diferencias entre el coeficiente de regresión estimado y los estandarizado.

Cuando no hay ninguna variable $X_i$ (i=1,...,p-1) que se relaciona linealmente con los demás en el modelo $R^2_k = 0$, por lo que $(VIF)_k = 1$ y

$$\mathbb{E}\left[\sum_{k=1}^{p-1}(\widehat{\beta}_k-\beta_k)^2\right] = (\sigma')^2(p-1) \tag{3.24}$$

Una relación de los resultados de (3.23) y (3.24) proporciona información útil sobre el efecto de la multicolinealidad en la suma de los errores al cuadrado:

$$\frac{(\sigma^{'})^2\sum_{k=1}^{p-1}(VIF)_k}{(\sigma^{'})^2(p-1)} = \frac{\sum_{k=1}^{p-1}(VIF)_k}{p-1} \tag{3.25}$$

Note que esta relación es la simple media de los factores $(VIF)_k$, puede ser denotada por:

$$(\overline{VIF}) = \frac{1}{p-1}\sum_{k=1}^{p-1}(VIF)_k = \frac{1}{p-1}Tr(R^{-1}_{\mathbb{XX}}) \tag{3.26}$$

Los valores medios de VIF que son considerablemente mayores que 1 son indicativos de graves problemas de multicolinealidad.

En el contexto de muestras georeferenciadas la varianza del término medio de la muestra para la variable $Y$ por la autocorrelación espacial se supone positiva. Para la matriz de covarianza $n \times n$ espacial $\sigma^2 V$, esta inflación de varianza está dada por $Tr(V)$, donde $Tr(\cdot)$, denota el operador matricial llamado traza. La información redundante aquí es cuantificada por $\mathbf{1}^t V^{-1}\mathbf{1}$, donde $^t$, denota el operador transpuesta. Si la autocorrelación espacial es cero, entonces $V = I$, y aquí $Tr(V^{-1}) = n$ y $\mathbf{1}^t V^{-1}\mathbf{1} = n$, donde $n$ es el tamaño de la muestra, es decir, el número de posiciones en estudio. Finalmente, el tamaño muestral efectivo está dado por $n* = n\frac{Tr(V^{-1})}{\mathbf{1}^t V^{-1}\mathbf{1}}$, lo cual es igual a $n$, cuando la correlación espacial es cero. Para una correlación espacial positiva, la matriz $V$ es progresivamente proporcional a $\mathbf{1}\mathbf{1}^t$. consecuentemente, $n^*$ converge a 1.

A continuación se presentan algunos resultados utilizando la propuesta de Griffith cuando se tiene estructura de correlación conocidas, como es el caso de la matriz de correlación intraclase y de un proceso autoregresivo de primer orden.

### 3.3.2 Tamaño Muestral Efectivo con estructura de Correlación Intra Clase y Autoregresiva

Para la validación de estas expresiones se realizarán algunos ejemplos. Se utilizarán matrices con formas particulares como son las matrices intraclases y AR(1). De esta forma se podrá visualizar el efecto que tiene la presencia de autocorrelación en el tamaño muestral efectivo. Como también, el número de datos independientes que se necesitan para estimar los parámetros de interés $\mu$ con una correlación espacial dada.

#### Matriz de Covarianza Intraclases

La matriz intraclase es muy utilizada en el análisis multivariante. En esta ocasión se cuantificará los efectos que tiene la matriz de correlación intraclase en la fórmula propuesta por Griffith. Como también, sus propiedades en función de su descomposición para la inversa y estimadores asociados. Para este caso se tiene un conjunto de datos con atributos georefenciados asociados la la variable $\mathbf{Y}$, $n \times 1$ que presentan una estructura de correlación intraclases, es decir; Sea $\mathbf{Y} \sim (\mu\mathbf{1}, \Sigma)$, donde $0 < \rho < 1$, su forma matricial es:

$$\Sigma(\rho) = \sigma^2 \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \rho & \dots & 1 \end{pmatrix}. \tag{3.27}$$

y la matriz de correlación tiene la siguiente forma

$$R(\rho) = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \rho & \dots & 1 \end{pmatrix}. \tag{3.28}$$

donde $R(\rho) = D^{-1/2}\Sigma D^{-1/2}$, $D = Diag(\Sigma)$. Esta matriz se puede descomponer de la siguiente manera

$$R(\rho) = \left[(1-\rho)\cdot \mathbb{I} + \rho \cdot \mathbf{1}\mathbf{1}^t\right]. \tag{3.29}$$

Consideremos la siguiente propiedad

**Propiedad 3.1** *Si* $1 \pm v^t \cdot A \cdot u \neq 0$, *entonces:*

$$(A \pm u \cdot v^t)^{-1} = A^{-1} \mp \frac{A^{-1}uv^tA^{-1}}{1 \pm v^t \cdot A \cdot u}. \tag{3.30}$$

Ahora, si se toma $A = (1-\rho)\cdot\mathbb{I}$, $u = \rho\cdot\mathbf{1}$ y $v = \mathbf{1}$. Y utilizando la propiedad anterior se tienen los siguientes resultados:

$$R^{-1}(\rho) = \frac{1}{(1-\rho)}\cdot\mathbb{I} - \frac{\rho}{(1-\rho)\cdot(1+\rho\cdot(n-1))}\mathbf{1}\mathbf{1}^t, \tag{3.31}$$

$$Tr(R^{-1}(\rho)) = \frac{n}{(1-\rho)} - \frac{n\cdot\rho}{(1-\rho)\cdot(1+\rho\cdot(n-1))} = \frac{n\cdot(1+\rho\cdot(n-2))}{(1-\rho)\cdot(1+\rho\cdot(n-1))}, \tag{3.32}$$

$$\mathbf{1}^t R^{-1}(\rho)\mathbf{1} = \frac{n}{(1-\rho)} - \frac{n^2\cdot\rho}{(1-\rho)\cdot(1+\rho\cdot(n-1))} = \frac{n}{1+\rho\cdot(n-1)}, \tag{3.33}$$

Utilizando la fórmula propuesta por Griffith en (3.19) el tamaño muestral efectivo queda como:

$$n^* = n\cdot\frac{(1+\rho(n-2))}{(1-\rho)}. \tag{3.34}$$

Otra característica importante de esta matriz es si consideramos $\mathbf{Y}_i \sim N(\mu\mathbf{1},\Sigma)$ una muestra de tamaño $n$ con $\Sigma_{intra} = (\sigma\cdot\rho)_{ij}$, $i,j = 1,\ldots,n$, los estimadores máximo verosímiles para $\sigma^2$ y $\rho$ (Ver Rencher, pág. 256) son :

$$\hat{\sigma} = \frac{1}{n}Tr[S] = \frac{1}{n}\sum_{i=1}^{n} S_{ii}^2. \tag{3.35}$$

$$\hat{\rho} = \frac{1}{n(n-1)\cdot\hat{\sigma}}\cdot 2\sum_{i>j}^{n} S_{ij}. \tag{3.36}$$

### Matriz de Covarianza AR(1)

Otra matriz que tiene una estructura conocida son las matrices con correlación autoregresivo de primer orden, las cuales vienen dada por $Y_t = \phi\cdot Y_{t-1} + \epsilon_t$ donde $|\phi| < 1$.
Sea $Y_t = \phi Y_{t-1} + \epsilon_t$, donde $\epsilon_t \sim (0,\sigma^2)$. Si el proceso $Y_t$ es invertible, se puede expresar este proceso en su forma causal

$$Y_t = \sum_{j=0}^{\infty}\phi^j\epsilon_{t-j}. \tag{3.37}$$

Calculando Esperanza, Varianza, Covarianza y correlación se obtiene

$$\text{(a) } \mathbb{E}[Y_t] = 0,\ \text{(b) } \mathbb{V}ar[Y_t] = \frac{\sigma^2}{1-\phi^2},\ \text{(c) } \mathbb{C}ov[Y_t,Y_{t+h}] = \frac{\sigma^2\phi^h}{1-\phi^2},\ \text{(d) } \mathbb{C}orr[Y_t,Y_{t+h}] = \phi^h. \tag{3.38}$$

Considere un conjunto de variables con atributos georefenciados con estructura de correlación $AR(1)$. Además, $Y \sim N_n(\mu\mathbf{1},\Sigma)$, con $\Sigma_{AR(1)} = \sigma\cdot\phi^{|i-j|}$ $i,j = 1,\ldots,n$ entonces la matriz de covarianza autoregresiva de primer orden es:

$$Corr(Y_t,Y_{t+h}) = R(\phi) = \begin{pmatrix} 1 & \phi & \ldots & \phi^{n-1} \\ \phi & 1 & \ldots & \phi^{n-2} \\ \ldots & \ldots & \ldots & \ldots \\ \phi^{n-1} & \phi^{n-2} & \ldots & 1 \end{pmatrix} \tag{3.39}$$

Se puede observar que $R^{-1}$ existe y su inversa pertenece a las matrices Toeplitz de tipo 2, la cual tiene la siguiente estructura

$$R^{-1}(\phi) = \frac{1}{(1-\phi^2)}\begin{pmatrix} 1 & -\phi & \ldots & 0 \\ -\phi & 1+\phi^2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

Matrices del Tipo 2 matrices también se producen como las inversas de otras matrices con patrones especiales que se presentan en otras aplicaciones estadísticas comunes.

$$Tr(R^{-1}(\phi)) = \frac{2+(n-2)\cdot(1+\phi^2)}{(1-\phi^2)}, \tag{3.40}$$

$$\mathbf{1}^t R^{-1}(\phi)\mathbf{1} = \frac{2+(n-2)\cdot(1-\phi)}{(1+\phi)}, \tag{3.41}$$

Finalmente, utilizando la fórmula anterior (3.19) se tiene

$$n^* = n\cdot\frac{(2+(n-2)\cdot(1+\phi)^2)}{(1-\phi)\cdot(2+(n-2)\cdot(1-\phi))}. \tag{3.42}$$

Como se muestra en los resultados anteriores dependiendo de la estructura de la matriz de covarianza el tamaño muestral efectivo cambia. Este problema se transforma en una relación inversa entre la función de correlación y el número de muestras, es decir, a menor correlación mayor es la muestra y a mayor correlación se necesita un tamaño muestral menor.

## 3.4 Tamaño Muestral Efectivo para Variables Espaciales

Por otro lado Vallejos (2010), basados en los resultados anteriores estudia y define el tamaño muestral efectivo, denotado por $ESS$ como el numerador de la expresión anterior propuesta por Griffith e inspecciona algunas características para el $ESS$ con matrices de patrones especiales. Más formalmente, se define $ESS$ como sigue,

**Definición 3.2** *Sea $Y$ un vector aleatorio con atributos georeferenciados de tamaño $n\times 1$ con valor esperado $\mu\mathbf{1}$ y matriz de correlación espacial $R$. Llamaremos la cantidad*

$$ESS = ESS(n,R) = \mathbf{1}^t R^{-1}\mathbf{1}. \tag{3.43}$$

*tamaño muestral efectivo (ESS)*

Por ejemplo, se puede verificar si $R = I$ entonces $ESS(n,I) = n$, es decir, el vector $Y_{n\times 1}$ es una muestra aleatoria e independiente. Bajo ciertas condiciones el tamaño muestral efectivo fluctúa entre $1 \leq ESS \leq n$. La cantidad $ESS$ entrega el número de observaciones no redundantes (independiente) de una muestra de tamaño $n$.

Considerando la definición (3.43) se puede definir una medida que refleje la reducción efectiva $RE$ del tamaño muestral $ESS$, solo basta dividir por $n$ y obtenemos:

$$\frac{1}{n} \leq \frac{\mathbf{1}^t R^{-1}\mathbf{1}}{n} \leq \frac{n}{n} = \frac{1}{n} \leq RE(n,R) \leq 1 \tag{3.44}$$

Se puede interpretar que para valores cercanos al cero existe una correlación espacial persistente que reduce efectivamente la muestra, en el caso que esta cantidad se acerque a uno, se puede decir, que no hay una correlación espacial persistente que reduzca efectivamente la muestra.

### 3.4.1 $ESS$ como Cantidad de Información de Fisher

Un resultado interesante a estudiar es tener una medida que nos entregue una cantidad de información contenida en una muestra. Esto es conocido como Cantidad de Información de Fisher. Sea $Y \sim N_n(\mu\mathbf{1}, \sigma^2 R)$ con función de densidad dada por:

$$f_Y(y) = \frac{1}{(2\pi\sigma^2)^{n/2}|R^{-1}|^{1/2}}\exp\left\{\frac{1}{2}\left((Y-\mu\mathbf{1})^t\frac{R^{-1}}{\sigma^2}(Y-\mu\mathbf{1})\right)\right\}. \tag{3.45}$$

Ahora bien, si consideremos la expresión para determinar la Cantidad de Información de Fisher

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log_e f(y,\theta)\right]. \tag{3.46}$$

y desarrollando se obtiene

$$\begin{aligned}\log_e f_Y(y) &= \frac{-n}{2}\left[log_e(2\pi) + log_e(\sigma^2)\right] - \frac{1}{2}\log_e|R^{-1}| - \frac{1}{2}\left((Y-\mu\mathbf{1})^t\frac{R^{-1}}{\sigma^2}(Y-\mu\mathbf{1})\right),\\ \log_e f_Y(y) &\propto -\frac{1}{2}\left((Y-\mu\mathbf{1})^t\frac{R^{-1}}{\sigma^2}(Y-\mu\mathbf{1})\right),\end{aligned}$$

Derivando una y otra vez, se tiene

$$\frac{\partial^2}{\partial\mu^2}\log_e f_Y(y) = \frac{-\mathbf{1}^tR^{-1}\mathbf{1}}{\sigma^2}. \tag{3.47}$$

Finalmente

$$-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log_e f(y,\theta)\right] = \frac{\mathbf{1}^tR^{-1}\mathbf{1}}{\sigma^2} = \frac{ESS(n,R)}{\sigma^2}. \tag{3.48}$$

La cantidad $ESS$ puede ser vista como una función de la cantidad de información. O bien, $ESS(n,\theta)$ es una cantidad que indica la reducción efectiva de la muestra de tamaño $n$. Se llama reducción efectiva a un subconjunto de elementos independientes provenientes de una muestra de tamaño $n$ correlacionada espacialmente.

### 3.4.2 Reducción Efectiva en el Tamaño Muestral

Para la validación de esta nueva cantidad, se puede inspeccionar los efectos en la reducción efectiva del tamaño muestral para diferentes tamaños muestrales y distintos valores de correlación entre los datos.

#### *ESS* para Matriz Intraclase

Utilizando los cálculos anteriores se puede establecer que el tamaño muestral efectivo para una muestra con estructura de correlación intraclases es:

$$ESS_{intra}(n,\rho) = \mathbf{1}^tR^{-1}(\rho)\mathbf{1} = \frac{n}{1+(n-1)\rho}. \tag{3.49}$$

Se puede observar que, cuando $\rho \to 0 \Rightarrow ESS_{intra}(n,\rho) \to n$. Por otro lado, si $\rho \to 1 \Rightarrow ESS_{intra}(n,\rho) \to 1$, o bien, $1 < ESS_{intra}(n,\rho) < n$.

El comportamiento del tamaño muestral efectivo para una estructura de correlación intraclase para distintos valores de $n = 20, 50, 100$ y $0 < \rho < 1$.

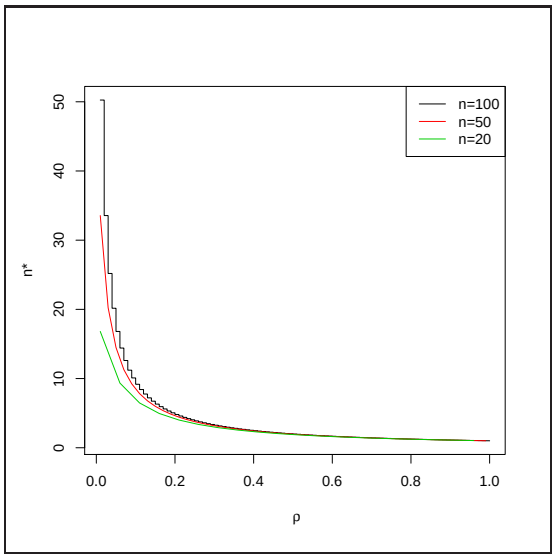


Figura 3.1: $ESS_{intra}(n, \rho)$ con $n = 20, 50, 100$ y $0 < \rho < 1$

Para una correlación intraclases, se puede ver que existe una reducción efectiva de la muestra en forma muy rápida, es decir, para valores de $\rho$ cercanos a 1, la reducción efectiva se acerca a 1.

### *ESS* para Matriz AR(1)

De igual forma, utilizando los cálculos anteriores se puede establecer que el tamaño muestral efectivo para una muestra con estructura de correlación AR(1), tiene la forma:

$$ESS_{AR(1)}(n, \phi) = \mathbf{1}^t R^{-1}(\phi)\mathbf{1} = \frac{2 + (n-2)(1-\phi)}{1+\phi}. \tag{3.50}$$

Se puede observar que, cuando $\phi \to 0 \Rightarrow ESS_{AR(1)}(n, \theta) \to n$. Por otro lado, si $\phi \to 1 \Rightarrow ESS_{AR(1)}(n, \phi) \to 1$, o bien, $1 < ESS_{AR(1)}(n, \phi) < n$.

El comportamiento del tamaño muestral efectivo para una estructura de correlación intraclases para distintos valores de $n = 20, 50, 100$ y $0 < \phi < 1$, se presenta gráficamente.

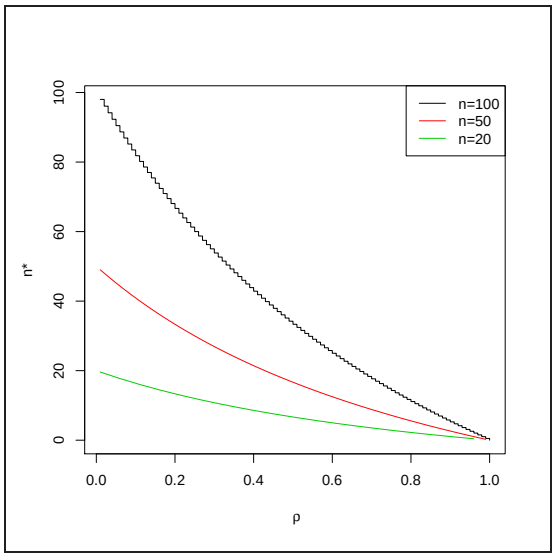


Figura 3.2: $ESS_{AR(1)}(n, \phi)$ con $n = 20, 50, 100$ y $0 < \phi < 1$

Para una correlación autoregresivo de primer orden, se puede ver que esta función reduce el tamaño muestral de manera menos severa, que en presencia de correlación intraclases, es decir, para valores de $\phi$ cercanos a 1, la reducción efectiva se acerca a 1 más lento.

Hasta ahora hemos visto que para matrices que poseen estructura de correlación conocidas, se tiene formulas explícitas para el tamaño muestral efectivo $ESS$. Sin embargo, en la practica necesitamos una muestra para estimar los parámetros de la matriz de correlación, de manera de obtener un tamaño muestral efectivo.

En la literatura hay expresiones conocidas en la estimación de estos parámetros $(\rho, \phi)$, como a su vez, la distribución asintótica de estos parámetros, asociados a la distribución normal. Es por eso, que funciones de estos parámetros, también tiene asociado una distribución asintótica normal, en la siguiente sección se inspeccionará esta propiedad utilizando el método Delta.

### 3.4.3 Transformaciones para estabilizar la varianza (Método Delta)

**Teorema 3.3** *Suponga que una secuencia de variables aleatorias reales $T_n$ que converge en probabilidad a otra variable aleatoria real $T$, cuando $n \to \infty$. Entonces $T_n \xrightarrow{d} T$, cuando $n \to \infty$.(Ver Mukhopadhay, 2000. página 256.)*

Ahora bien, supongamos que se tiene una secuencia de valores reales de los estadísticos $\{T_n; n \geq 1\}$, tal que, $n^{1/2}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)$, cuando $n \to \infty$. Entonces, desde el teorema de Mann-Wald 3.3 se puede concluir que:

$$\sqrt{n}\,[g(T_n) - g(\theta)] \xrightarrow{d} N\left(0, (g^{'}(\theta)\sigma)^2\right),\ n \to \infty \tag{3.51}$$

si $g(\cdot)$ es una función continua real y $g^{'}(\theta)$ es finita y no cero. Cuando $\sigma^2$ involucra a $\theta$, se puede determinar una función $g(\cdot)$ apropiada. Tal que, para $n$ grande la varianza aproximada de los estadísticos asociados a $g(T_n)$ se libera de los parámetros desconocidos $\theta$. Esta función $g(\cdot)$ es llamada una transformación para la estabilización de la varianza (Mukhopadhay, 2000. página 555).

Este resultado nos ayudará a encontrar una varianza para los tamaños muestrales efectivos con matrices intraclases y AR(1). La idea de estos resultados es presentar que el tamaño muestral efectivo puede ser vista como una variable aleatoria, entonces cuando se tiene una formula explícita se pueden encontrar expresiones para la media y varianza.

En este punto $ESS$ puede ser estudiada como una variable aleatoria, esto permite encontrar expresiones para la media y la varianza. Esta medidas nos permiten encontrar intervalos de confianzas asintóticos para $ESS$ y otras inferencias relacionadas.

### Matriz Intraclases

Utilizando los cálculos desarrollados anteriormente en (3.36) se tiene una expresión para estimar el coeficiente de correlación muestral para muestras multivariadas, denotado por $\widehat{\rho}_{n_1}$, es decir, $\widehat{\rho}_{n_1}$ es un estimador de $\rho$. Por otro, lado se sabe que

$$\sqrt{n_1}(\widehat{\rho}_{n_1} - \rho) \xrightarrow{d} N\left(0, (1-\rho^2)^2\right), n_1 \to \infty. \tag{3.52}$$

(la expresión (3.52) se cumple bajo cierta condiciones, dado que este resultado está asociado un proceso bi-dimensional) y

$$\widehat{\rho}_{n_1} \xrightarrow{d} N\left(\rho, \frac{\left(1-\rho^2\right)^2}{n_1}\right), n_1 \to \infty. \tag{3.53}$$

Las expresiones en (3.52) y (3.53) se puede encontrar en (Mukhopadhay, pág. 561). Estos elementos ayudarán a encontrar la distribución teórica de $ESS_{intra}(n,\rho)$. Por otro lado, sea $ESS_{intra}$, una función continua tal que, la derivada respecto a $\rho$ exista. Derivando $ESS_{intra}(n,\rho)$ se tiene:

$$\frac{\partial}{\partial\rho} ESS_{intra}(n,\rho) = \frac{\partial}{\partial\rho}\left(\frac{n}{1+(n-1)\rho}\right) = -n(1+(n-1)\rho)^{-2}(n-1) = \frac{-n(n-1)}{[1+(n-1)\rho]^2} \tag{3.54}$$

Sea $\widehat{ESS}_{intra}(n,\widehat{\rho}_{n_1})$ un estimador de $ESS_{intra}(n,\rho)$. Utilizando el resultado en (3.51) se puede determinar la distribución asintótica de $ESS_{intra}(n,\rho)$, entonces utilizando el método delta

$$\widehat{ESS}_{intra}(n,\widehat{\rho}_{n_1}) \xrightarrow{d} N\left(\mathbb{E}\left[\widehat{ESS}_{intra}(n,\widehat{\rho}_{n_1})\right], \mathbb{V}ar\left[\widehat{ESS}_{intra}(n,\widehat{\rho}_{n_1})\right]\right), n_1 \to \infty \tag{3.55}$$

$$\widehat{ESS}_{intra}(n,\widehat{\rho}_{n_1}) \xrightarrow{d} N\left(ESS_{intra}(n,\rho), \frac{(1-\rho^2)^2}{n_1}\left(\frac{\partial}{\partial\rho} ESS_{intra}(n,\rho)\right)^2\right), n_1 \to \infty \tag{3.56}$$

$$\widehat{ESS}_{intra}(n,\widehat{\rho}_{n_1}) \xrightarrow{d} N\left(\frac{n}{1+(n-1)\rho}, \frac{(1-\rho^2)^2}{n_1}\left[\frac{n(n-1)}{[1+(n-1)\rho]^2}\right]^2\right), n_1 \to \infty \tag{3.57}$$

Esto puede ser escrito como sigue.

$$\widehat{ESS}_{intra}(n,\widehat{\rho}_{n_1}) \xrightarrow{d} N\left(\mu(\rho), \nu^2(\rho)\right), n_1 \to \infty \tag{3.58}$$

$$\tag{3.59}$$

con

$$\mu(\rho) = \frac{n}{1+(n-1)\rho} \tag{3.60}$$

$$\nu(\rho) = \sqrt{\frac{(1-\rho^2)^2}{n_1}\left[\frac{n(n-1)}{[1+(n-1)\rho]^2}\right]^2} \tag{3.61}$$

Finalmente, un intervalo de confianza para $ESS_{intra}(n,\rho)$ con un nivel de confianza de $100(1-\alpha)\,\%$ es

$$I.C(ESS_{intra}(n,\rho))_{100(1-\alpha)\,\%} = \left[\widehat{ESS}_{intra}(n,\widehat{\rho}_{n_1}) \mp Z_{1-\alpha/2}\cdot\nu(\rho)\right] \tag{3.62}$$

Los resultados anteriores nos permiten establecer intervalos de confianza para el tamaño muestral efectivo con estructura analítica o expresiones explícitas cuando se tiene una estructura de correlación intraclase.

Matriz AR(1)

Para un modelo autoregresivo de primer orden de la forma $Y_t = \phi Y_{t-1} + \epsilon_t$, $\epsilon_t \sim N(0, \sigma_\epsilon^2)$

$$\sqrt{n}(\widehat{\phi}_{n_1} - \phi) \xrightarrow{d} N\left(0, 1-\phi^2\right), n_1 \to \infty, \tag{3.63}$$

$$\widehat{\phi}_{n_1} \xrightarrow{d} N\left(\phi, \frac{1-\phi^2}{n_1}\right), n_1 \to \infty. \tag{3.64}$$

En la literatura se pueden encontrar distintos métodos de estimación, entre ellos se pueden destacar el método de momentos de Yule Walker o estimación Máximo Verosímil (Ver, Box-Jenkins, 1976). Por otro lado, sea $ESS_{AR(1)}(n, \phi)$, una función continua tal que, la derivada respecto a $\phi$ existe:

$$\frac{\partial}{\partial \phi} ESS_{AR(1)}(n, \phi) = \frac{\partial}{\partial \phi}\left(\frac{2 + (n-2)(1-\phi)}{1+\phi}\right) = \frac{2-n}{1+\phi} - \frac{[2(n-2)(1-\phi)]}{(1+\phi)^2} = \frac{(3-\phi)(2-n)}{(1+\phi)^2}. \tag{3.65}$$

Sea $\widehat{ESS}_{intra}(n, \widehat{\phi}_{n_1})$ un estimador de $ESS_{AR(1)}(n, \phi)$. Realizando los mismos pasos que el resultado anterior se tiene:

$$\widehat{ESS}_{AR(1)}(n, \widehat{\phi}_{n_1}) \xrightarrow{d} N\left(ESS_{AR(1)}(n, \phi), \frac{1-\phi^2}{n_1}\left[\frac{\partial}{\partial \phi} ESS_{AR(1)}(n, \phi)\right]^2\right), n_1 \to \infty \tag{3.66}$$

$$\widehat{ESS}_{AR(1)}(n, \widehat{\phi}_{n_1}) \xrightarrow{d} N\left(\frac{2 + (n-2)(1-\phi)}{1+\phi}, \frac{1-\phi^2}{n_1}\left[\frac{(3-\phi)(2-n)}{(1+\phi)^2}\right]^2\right), n_1 \to \infty \tag{3.67}$$

Finalmente, un intervalo de confianza para $ESS_{AR(1)}(n, \phi)$ con un nivel de confianza de $100(1-\alpha)\,\%$ es

$$I.C(ESS_{AR(1)}(n, \phi))_{100(1-\alpha)\,\%} = \left[\widehat{ESS}_{AR(1)}(n, \widehat{\phi}_{n_1}) \mp Z_{1-\alpha/2} \cdot \nu(\phi)\right] \tag{3.68}$$

con

$$\nu(\phi) = \sqrt{\frac{1-\phi^2}{n_1}\left[\frac{(3-\phi)(2-n)}{(1+\phi)^2}\right]^2} \tag{3.69}$$

Cuando se tiene una estructura de correlación autoregresiva de primer orden los resultados anteriores nos permiten establecer intervalos de confianza para el tamaño muestral efectivo con una expresión explícita.

## 3.5 *ESS* con Modelos Paramétricos para el Semivariograma

### 3.5.1 *ESS* con Modelo Exponencial

Utilizando las definiciones 2.5.4 y 3.43, se puede definir el tamaño muestral efectivo cuando se tiene una estructura de correlación espacial exponencial $ESS_{Exp}$ de la siguiente manera:

**Definición 3.4** *Sea $\mathbb{Y}_{n\times 1} \sim N_n(\mu \mathbf{1}, \Sigma_{Exp}(\theta))$ con estructura de correlación Exponencial $R_{Exp}$ definida en (2.8), que contiene el vector de parámetros $\theta$, con $\theta = (\sigma^2, \tau^2, \phi)^t$. Se define el tamaño muestral con correlación espacial Exponencial, denotado por $ESS_{Exp}(n, \theta)$ y definido por:*

$$ESS_{Exp}(n, \theta) = \mathbf{1}^t R_{Exp}^{-1}(\theta)\mathbf{1}. \tag{3.70}$$

### 3.5.2 *ESS* con Modelo Esférico

De igual forma, utilizando las definiciones 2.5.3 y 3.43, se puede definir el tamaño muestral efectivo cuando se tiene una estructura de correlación espacial esférico $ESS_{Exp}$ de la siguiente manera:

**Definición 3.5** *Sea $\mathbb{Y}_{n\times 1} \sim N_n(\mu \mathbf{1}, \Sigma_{Esf}(\theta))$ con estructura de correlación Esférica $R_{Esf}$ definida en (2.8), que contiene el vector de parámetros $\theta$, con $\theta = (\sigma^2, \tau^2, \phi)^t$. Se define el tamaño muestral con correlación espacial Esférica, denotado por $ESS_{Esf}(n, \theta)$ y definido por:*

$$ESS_{Esf}(n, \theta) = \mathbf{1}^t R_{Esf}^{-1}(\theta)\mathbf{1}. \tag{3.71}$$

### 3.5.3 *ESS* con Modelo Mathérn

Utilizando las definiciones 2.5.6 y 3.43, se puede definir el tamaño muestral efectivo cuando se tiene una estructura de correlación espacial mathérn $ESS_{Exp}$ de la siguiente manera:

**Definición 3.6** *Sea $\mathbb{Y}_{n\times 1} \sim N_n(\mu \mathbf{1}, \Sigma_{Mat}(\theta))$ con estructura de correlación Mathérn $R_{Mat}$ definida en (2.8), que contiene el vector de parámetros $\theta$, con $\theta = (\sigma^2, \tau^2, \phi)^t$. Se define el tamaño muestral con correlación espacial Mathérn, denotado por $ESS_{Exp}(n, \theta)$ y definido por:*

$$ESS_{Mat}(n, \theta) = \mathbf{1}^t R_{Mat}^{-1}(\theta)\mathbf{1}. \tag{3.72}$$

### 3.5.4 *ESS* con Modelo Gaussiano

Utilizando las definiciones 2.5.5 y 3.43, se puede definir el tamaño muestral efectivo cuando se tiene una estructura de correlación espacial gaussiano $ESS_{Exp}$ de la siguiente manera:

**Definición 3.7** *Sea $\mathbb{Y}_{n\times 1} \sim N_n(\mu \mathbf{1}, \Sigma_{Gaus}(\theta))$ con estructura de correlación Gaussiano $R_{Gaus}$ definida en (2.8), que contiene el vector de parámetros $\theta$, con $\theta = (\sigma^2, \tau^2, \phi)^t$. Se define el tamaño muestral con correlación espacial Gaussiano, denotado por $ESS_{Exp}(n, \theta)$ y definido por:*

$$ESS_{Gaus}(n, \theta) = \mathbf{1}^t R_{Gaus}^{-1}(\theta)\mathbf{1}. \tag{3.73}$$

### 3.5.5 Ejemplos

A continuación se presentan unos ejemplos para ver el comportamiento del Tamaño Muestral Efectivo cuando existe correlación espacial. Se han realizado 4 modelos cada uno con distintos valores de $\theta$ y $n = 500$ fijo. Para cada uno de los tipos de correlación espacial se han considerado los siguientes modelos 1, 2, 3 y 4 con los siguientes valores: [Modelo N°1:] $\sigma^2$ =0.1, $\tau^2$ =0.1, $0 < \phi < 1$. [Modelo N°2:] $\sigma^2$ =0.9, $\tau^2$ =0.3, $0 < \phi < 1$. [Modelo N°3:] $\sigma^2$ =1.7, $\tau^2$ =0.5, $0 < \phi < 1$ y [Modelo N°4:] $\sigma^2$ =2.5, $\tau^2$ =0.7, $0 < \phi < 1$.

Las expresiones (3.70), (3.71), (3.72) y (3.73) tiene la misma definición descrita por Cressie y Griffith, es decir, estas cantidades representan el número de observaciones no correlacionadas que se necesitan para mantener la misma precisión que considerando todos los datos $n$. Es importante destacar que dependiendo de los valores que tome $\theta$ es el tamaño muestral efectivo resultante. La Figura (3.3) presenta los siguientes resultados, el objetivo de realizar los 4 modelos con los mismos valores para cada modelo es que permite la comparación entre ambos y destacar que cada uno tiene un comportamiento distinto respecto su forma. Una característica que tienen los 4 casos es que el Tamaño Muestral Efectivo aumenta a medida $\phi$ es cercano a cero y el Tamaño Muestral Efectivo disminuye a medida que $\phi$ es cercano a uno. Se podría decir, que $\phi$ es un parámetro influyente en el Tamaño Muestral Efectivo.

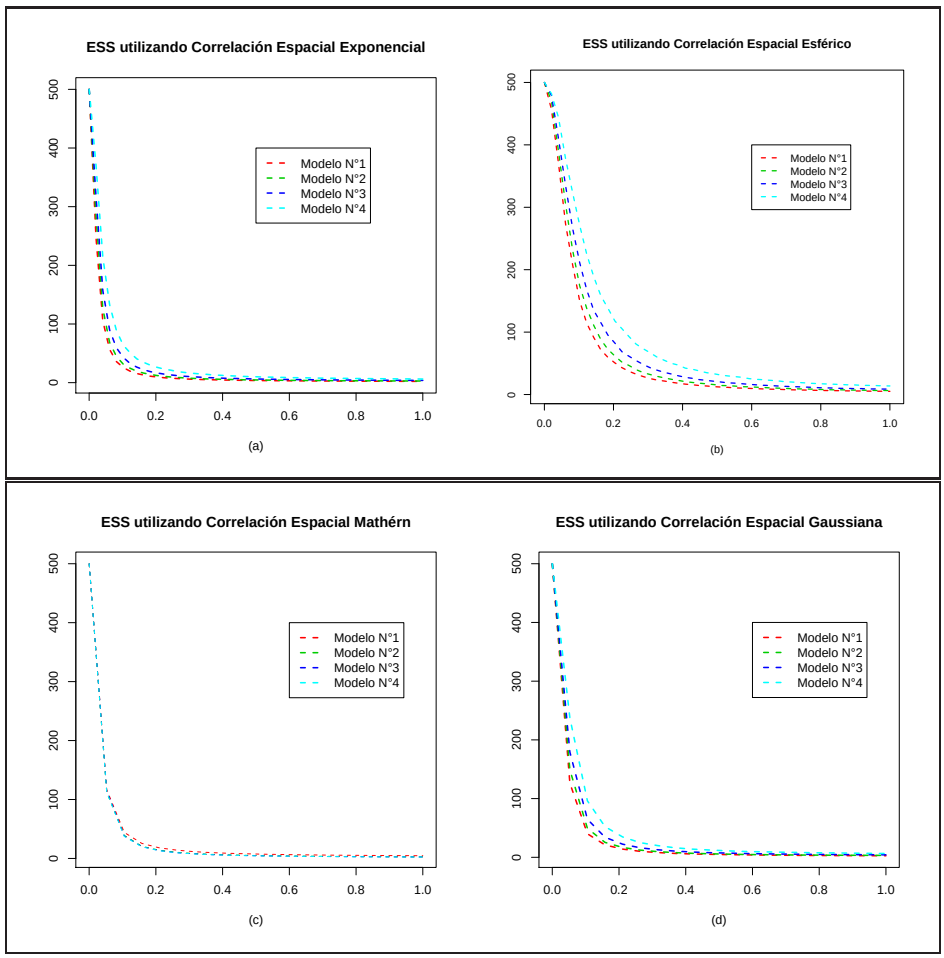


Figura 3.3: Tamaño Muestral Efectivo con diferentes correlación espacial (a) Exponencial, (b) Esférica, (c) Mathérn y (d) Gaussiana.

## 3.6 Comentarios

El objetivo de este capítulo es presentar los impactos de la correlación espacial en el tamaño muestral efectivo. En este sentido, el estudio de las secciones 3.2, 3.3 y 3.4 muestran los efectos de tener una muestra correlacionada espacialmente al momento de querer estimar la media y varianza muestral de la variable de interés. Claramente, la presencia de correlación espacial en la muestra debe tener un tratamiento especial. Al existir correlación espacial en un conjunto de variables espaciales se puede decir en cierto sentido que existe una redundancia en la información obtenida. Al existir información redundante en una muestra espacial, se podría pensar que quizás no sea necesario ocupar toda la muestra, es decir, que existe un subconjunto de elementos que puede mantener las mismas propiedades que la muestra original, con respecto al estimador y la precisión. Se han presentado tres cantidades (3.13), (3.19) y (3.43) que pueden cuantificar el número de observaciones de observaciones independientes necesarias para mantener la misma precisión que la muestra original. Las Figuras 3.1 y 3.2 muestran los efectos del tamaño muestral efectivo para distintos valores de correlación espacial intraclase y autoregresiva de primer orden respectivamente. Una de las características que tiene el tamaño muestral efectivo $ESS$, es cuando la matriz de correlaciones espacial tiene patrones conocidos y es posible encontrar expresiones analíticas para $ESS$. Esta característica nos permite utilizar propiedades asintóticas, como el Método Delta y establecer la distribución del tamaño muestral efectivo $ESS$, cuando la estructura de correlación espacial es intraclase (3.57) y cuando la estructura de correlación espacial es autoregresivo de primer orden (3.67). No obstante, cuando la estructura de correlación espacial es como en (3.71), (3.70), (3.73) y (3.72) determinar la distribución de estas cantidades se complica, debido

a que no hay una expresión analítica que resolver, en este sentido determinar la distribución del tamaño muestral efectivo cuando la correlación espacial es modelada mediante un semivariograma es un problema abierto.

Capítulo 4

# Muestreo Espacial

## 4.1 Introducción

El objetivo de este Capítulo es presentar de una manera muy sintetizada el estado del arte, con respecto a las técnicas muestreo espacial y destacar como estos métodos son extensiones del muestreo clásico vistos por Chocran, Thompson, Kish por nombrar algunos autores. La diferencia es que ahora se agrega una información adicional y es que los datos tienen asociada una referencia geográfica, es decir, la unidad de muestreo tiene una ubicación en el espacio. En el contexto de este trabajo de Tesis de Tamaño Muestral Efectivo en el Modelamiento de Variables Espaciales, toma un rol relevante, dado que luego de determinar un tamaño muestra efectivo (óptimo), la pregunta siguiente es como seleccionar estos datos (unidades muestrales). Este capítulo presenta una breve reseña de algunos esquemas de muestreos específicos para variables aleatorias espaciales. El objetivo del muestreo espacial es tener la libertad de elegir $n$ puntos cualquiera dentro de una región y luego medir en esos puntos ver (Ripley, 1981), (Müller, 2007), (Haining et al, 1990, 2004), (Griffith, 2005, 2008) y (de Gruijter y Brus, 2006).

El muestreo es un método de selección de unidades experimentales que están asociados a la realización de una variable aleatoria. Generalmente los datos se obtienen de una encuesta (instrumento de medición) que busca medir la variable de interés. Dependiendo de cómo este dispuesta la población al momento de seleccionar un subconjunto de ella (muestra), es el esquema de muestreo a utilizar. El muestreo espacial mantiene el mismo principio, sin embargo tiene asociada una dimensión espacial, ya sea, uni o multidimensional. Las mayoría de las veces, el muestreo espacial es la realización de un proceso estocástico en 3D. Por ejemplo, que se quiere medir el ingreso medio de las viviendas en cierta ciudad, para cada vivienda seleccionada se tiene una posición latitud, longitud y el ingreso total de la vivienda. Ver Figura (4.1)

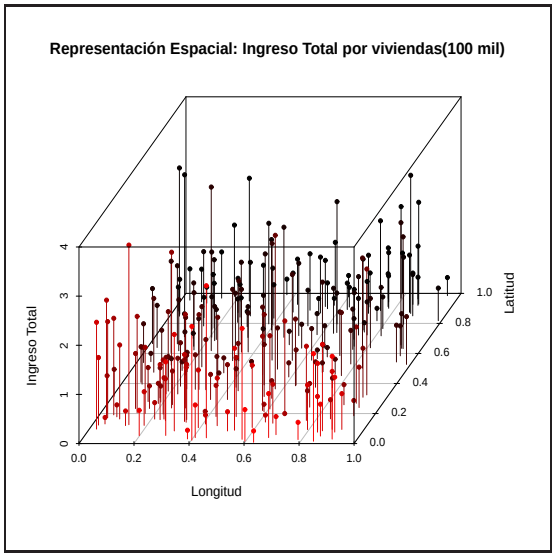


Figura 4.1: Representación espacial 3D de los ingresos por viviendas en ciento de miles.

## 4.2 Propósitos y Realización del Muestreo

El propósito de realizar muestreo espacial es el de realizar inferencias sobre una población donde cada unidad experimental tiene una referencia geográfica o georefernciada, la cual, está inserta en una población de interés. El muestreo se utiliza para evitar llevar a cabo un censo. Una de las principales razones pueden ser que la población puede ser tan grande que un censo sería físicamente imposible o poco práctico.

En otras situaciones, no es el tamaño de la población o el coste de la adquisición de los datos requeridos en el muestreo, pero el nivel de precisión, es una cantidad de interés y requiere tener cierto tratamiento. El error de muestreo respecto a un estimador tiene propiedades de interés como de medir el nivel de precisión. Esto se puede ver en (Haining, 1990, 2004) donde propone antes de realizar inferencias acerca de una población geográfica mediante muestreo se requiere una serie de decisiones que se deben tomar al momento de planificar el diseño de la encuesta. Estas decisiones se toman en relación a las siguientes preguntas:

($i$) ¿Qué va a ser estimado (estimador)?

($ii$) ¿Qué tamaño de la muestra ($n$) es necesaria para alcanzar el nivel deseado de precisión (error de muestreo)?

($iii$) Dado que el muestreo es espacial, ¿qué lugares dentro de los $n$ deben ser seleccionados para la muestra (Métodos de Muestreo espacial)?

($iv$) ¿Qué método se debe utilizar para calcular la cantidad de interés (Métodos de inferencia basado en el diseño o método basado en el modelo)?

($v$) ¿Qué medida de precisión (entre la estimación y el atributo de interés en la población) en el diseño de la muestra buscan minimizar (sesgo)?

Las respuestas a estas preguntas no son necesariamente independientes uno de otro. Por lo general es necesario usar criterios económicos y estadísticos en el diseño de una muestra. (Cressie, 1993) revisa brevemente a los intentos de formalizar dicha toma de decisiones.

## 4.3 Métodos de Muestreo e Inferencia Estadística

Por otro lado (de Gruijter y Brus, 2006) plantean que la necesidad de establecer un método; con *método* se refieren a una combinación de técnicas para la selección de unidades de muestreo y un método de inferencia estadística para la estimación una media espacial o la predicción de cierta cantidad en el espacio. Un método basado en el diseño se define como un método en el que las unidades de muestreo son seleccionados por muestreo probabilístico (ver Cochran, 1977) y en el que la inferencia estadística se basa en el diseño del muestreo, es decir, el diseño basado en la inferencia, véase la tabla (4.1). Un método basado en el modelo se define como un método que se basa en la inferencia estadística del modelo estocástico y no hay requisitos para la selección de las unidades de muestreo.

### 4.3.1 Método de Muestreo e Inferencia Estadística

Existen tres posibles métodos en la selección de la unidad de muestreo entre ellas se puede distinguir: Muestreo de conveniencia, el Muestreo intencional y Muestreo probabilístico o muestreo aleatorio. El concepto de muestreo de conveniencia se explica por sí mismo. Un ejemplo obvio es cuando el muestreo se limita a los bordes de las carreteras u otros lugares de fácil acceso. La ventaja de esta modalidad es que ahorra tiempo y costes. La desventaja es que las propiedades estadísticas son inferiores a los de los otros métodos. El muestreo intencional trata de seleccionar las unidades de muestreo para un fin determinado de manera que sirva mejor. Un ejemplo es el método de *estudio libre* en la asignación de las clases de suelos,

| Tipo de Método | Método de Selección | Método de Inferencia |
|---|---|---|
| Método Basado en el Diseño | Muestreo Probabilístico | Basado en el Diseño |
| Método Basado en el Modelo | Muestreo Intencionado | Basado en el Modelo |

Tabla 4.1: Definición del método de diseño basado en modelos y basado en una combinación de un método de selección de unidades de muestreo y un método de inferencia estadística.

donde inspector selecciona los puntos de muestreo que espera que sean más informativos con respecto a la delimitación de las clase de suelos. En este ejemplo, los lugares son seleccionados de una manera subjetiva, utilizando la experiencia, las características del paisaje visible e hipótesis pedogenéticos [1] , de modo que el inspector espera que la información de las observaciones sean más útiles.

En el muestreo probabilístico nos referimos a la metodología usada en (Cochran, 1977), (Kish, 1965), (Hansen et al, 1953), (Groves, 1928, 2004), entre otros autores. El objetivo principal del muestreo probabilístico es que la selección de la muestra sea independiente y que cada unidad de muestreo tenga una probabilidad de selección.

## 4.4 Enfoques del Muestreo Espacial

### 4.4.1 Basado en el Diseño

La inferencia basada en el diseño es la teoría clásica de muestreo. En el muestreo espacial los puntos en la población son los valores de la región como un conjunto de valores desconocidos que al margen de cualquier error de medición es un valor fijo. La aleatoriedad entra a través del proceso de selección de los lugares de muestreo. En el caso de una población discreta, la media tiene propiedades importantes:

$$\frac{1}{n}\sum_{i=1}^{n} Y(s_i) \tag{4.1}$$

donde $n$ es el número de unidades seleccionadas extraidas de la población, por lo que (4.1) es la media muestral de la población. Si $Y(s_i)$ es binario pendiendo si en la posición $s_i$ es una categoría o no, entonces (4.1) es la proporción de la población con algún atributo específico. En el caso de una población continua digamos una región $A$ de la zona $|A|$, entonces (4.1) se sustituye por la integral:

$$\frac{1}{|A|}\int Y(x)dx \tag{4.2}$$

El muestreo basado en el diseño está orientado en los estimadores (4.1) o (4.2) que incluyen el peso de cada observaciones en las probabilidades de ser incluido en la muestra. En los diferentes planes de muestreo la estrategia del muestreo depende de la estructura de la variación espacial de la población.

### 4.4.2 Basado en el Modelo

la inferencia basada en el modelo también tiene opiniones de muestreo espacial (intencional o convivencia) en población de interés o valores en una región de estudio, y su objetivo es la realización algún modelo estocástico. La fuente de aleatoriedad que está presente en una muestra se deriva de un modelo estocástico.

[1] Rondon y Elizalde (1992) proponen la siguiente definición de proceso pedogenéticos: Es toda acción que se produce en el cuerpo natural del suelo como un todo, o en algunos de sus componentes, por intercambios de materia y energía entre sus propios componentes y con su medio ambiente (determinado por sus límites) y que, con el tiempo, provoca cambios en la composición, en las propiedades físicas, químicas y biológicas, mineralógicas y/o estructurales, que pueden ser observados y/o medidos in-situ o en muestras aisladas (ex-situ)

Una estrategia de muestreo basado en modelos son los modelos utilizados para predecir cierta cantidad de interés en donde los modelos predictivos de (4.1) sólo se debe utilizar cuando el conocimiento disponible está disponible sobre la estructura de la población subyacente.

Aquí no hay necesidad de aleatoriedad en el plan de muestreo. Mientras que en el enfoque basado en el diseño la superficie no cambia nunca y la evaluación de una estrategia de muestreo debe considerar la adopción de repetidas muestras de probabilidad, en el enfoque basado en el modelo de cada realización produce una nueva superficie de valores y el plan de muestreo que se podrían adoptar en cada caso (de Gruijter y Brus, 2006).

## 4.5 Decisiones importantes del Diseño

### 4.5.1 Elección entre la inferencia basada en el diseño y basada en modelos

Antes de decidir sobre los diseños de muestreo la elección debe hacerse entre la inferencia basada en el diseño y el modelo de base, ya que la inferencia basada en el diseño necesita un muestreo de probabilidad, mientras que para la inferencia basada en el modelo el muestreo en general es sub-óptima. Esto está más allá del alcance de esta sección para discutir este tema en detalle, por lo que sólo se da aquí un esbozo. Un amplio debate se presenta en (de Gruijter y Brus, 2006). El *ideal* de las circunstancias para la aplicación del enfoque basado en el diseño son los siguientes:

- El resultado requiere una estimación de la distribución de frecuencias de la variable de destino en el universo como un todo, o un parámetro de esta distribución, tales como la media, la desviación estándar o un cuantil.
- Un tamaño de muestra mínimo, digamos, 5 o 10 unidades pueden permitir crear una función de la variación espacial, para tener por lo menos una idea aproximada del error de muestreo.
- En la práctica es factible probar en distintos lugares seleccionados al azar.
- Es importante obtener una estimación objetiva en el sentido, como media de todas las muestras posibles de un diseño aplicado, la estimación debe representar el valor real del parámetro objetivo.
- Es importante obtener una estimación objetiva de la incertidumbre de la estimación.

Alrededor de este *ideal* hay una serie de circunstancias en las que el enfoque basado en el diseño sigue siendo preferible al enfoque basado en el modelo. El *ideal* de las circunstancias para la aplicación del enfoque basado en el modelo son los siguientes:

- El resultado deseado es una predicción de los valores en los puntos individuales en el espacio al igual que con las previstas en una distribución de valores en el universo, similar a la cartografía.
- Se debe seleccionar un tamaño de muestra razonable, en función de la variación espacial. El modelo implica por lo general los supuestos de estacionariedad y un variograma, que debe ser estimado a partir de unos 100 a 150 puntos de muestreo.
- La variación espacial está descrita por un modelo.
- Existen fuertes autocorrelaciones en el universo.

Al igual que antes, en torno a este *ideal* hay una serie de circunstancias en las que el enfoque basado en el modelo sigue siendo preferible al enfoque basado en el diseño.

## 4.6 Panorámica en los Tipos de Diseños de Muestreo

Los tipos de diseños que presentaremos en esta sección pueden ser divididos dentro de cinco grupos principales:

1. Tipos de diseños básicos;
2. Tipos de diseños compuestos;
3. Tipos de diseños espacial;
4. Tipos de diseños en dos fases;
5. Tipos de diseños secuenciales.

Tipos de diseños básicos son:

- **Muestreo Aleatorio Simple:** las posiciones se han extraído del universo, al azar e independiente entre sí.

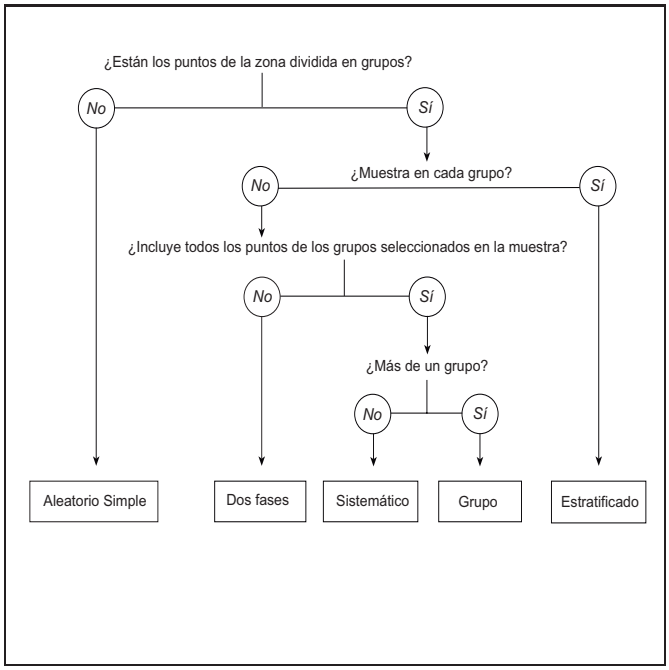


Figura 4.2: Similitudes y diferencias entre los tipos básicos de diseño

- **Muestreo Aleatorio Simple Estratificado:** las posiciones del universo son divididos dentro de grupos, aquí llamados *estratos*, y en cada estrato aplicar muestreo aleatorio simple.
- **Muestreo Aleatorio en Dos Etapas:** las posiciones en el universo son divididas dentro de grupos, aquí llamadas *unidades primarias de muestreo* (UPM), y se eligen mediante un muestreo aleatorio simple en la primera etapa para cada UPM seleccionada.
- **Muestreo Aleatorio de Grupos/cluster/Conglomerado:** Las posiciones en el universo son divididas dentro de grupos, aquí llamados *cluster*, después de un número de conglomerados ($> 1$) se seleccionan al azar, y todos los lugares de los conglomerados seleccionados se incluyen en la muestra. En el contexto espacial, los grupos son generalmente definidos de manera que formen patrones regulares espaciales, por ejemplo, lugares, a medio camino en una línea.

- **Muestreo Aleatorio Sistemático:** Similar a muestreo aleatorio de cluster, excepto que solamente un cluster es seleccionado. Otra vez, en el contexto espacial estos cluster únicos es típicamente definido como una forma regular espacial de patrones de posiciones, por ejemplo una grilla cuadrada.

- **Muestreo de probabilidades proporcional al tamaño (PPS):** Las unidades de muestreo se seleccionan con probabilidad proporcional a su tamaño o de una variable auxiliar que se correlaciona con la variable de destino.

Las diferencias y similitudes entre estos tipos se ilustran en el árbol lógico de la Figura (4.3), a excepción del muestreo (PPS), que puede ser considerado como una variedad de muestreo aleatorio simple desigual en lugar de las probabilidades de selección iguales.

Los **Tipos diseños compuestos** son combinaciones o estructuras anidadas de los tipos básicos de diseño, ellos representan los métodos de muestreo más avanzados. Por ejemplo, en la segunda etapa de muestreo aleatorio bietápico se podría aplicar muestreo aleatorio por conglomerados en lugar de muestreo aleatorio simple.

Los **Tipos de diseños espaciales** se presentan sobre la base de las coordenadas espaciales de los puntos de muestreo que están disponible en cierta zona. Puede ser que el tipo de diseño de muestreo básico o un diseño compuesto se podría hacer a partir de una lista tipo de marco de muestreo, con todos los puntos de muestreo posible en cualquier orden, independientemente de su posición en el espacio. (y aplicar variedades de muestreo espacial, ya sea, el muestreo aleatorio por conglomerados y muestreo aleatorio sistemático, con sus patrones de punto regular donde obviamente también se utilizarían las coordenadas, pero sólo en la definición de los grupos). Al igual que los tipos de diseño compuesto, los tipos espaciales de diseño representan los métodos de muestreo más avanzadas.

Los **Tipos de diseños en dos fases** son los métodos de muestreo que exploran la correlación entre una medida variable auxiliar y la variable objetivo. En la primera fase, una muestra relativamente grande se toma, en el que sólo la variable auxiliar se mide. En la segunda fase, una sub-muestra se toma de la muestra general, y la variable objetivo se mide sólo en esta sub-muestra. Si la variable auxiliar es cuantitativa, entonces un *estimador de regresión* se utiliza para estimar medias o fracciones. En el caso de una variable cualitativa auxiliar se puede utilizar post-estratificación.

Los **Tipos de diseño secuencial** procede tomando muestras de una a la vez o lote por lote, durante la cuál se calcula una estadística para determinar si debe continuar o no.

### 4.6.1 Eligiendo una Estrategia Basada en el Diseño

Las estrategias de muestreo consta de tres componentes principales: tipo de diseño, los atributos del tipo de diseño y el estimador. Con pocas excepciones, estos componentes tienen que ser elegidos por este orden, porque los atributos dependen del tipo, y el estimador depende del tipo y los atributos ver Figura 4.3.

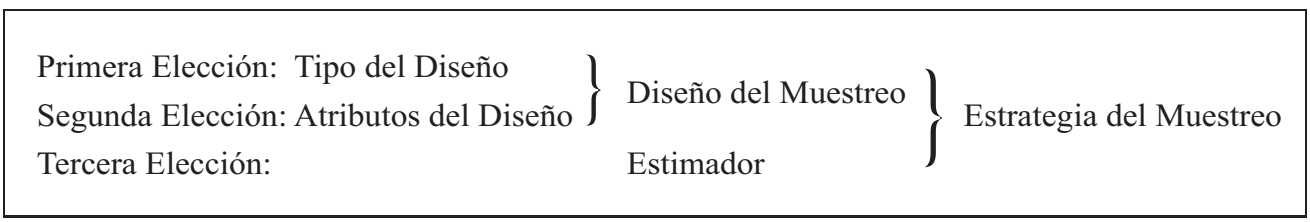


Figura 4.3: Opciones principales para decidir sobre una estrategia de muestreo basado en el diseño.

La elección de un tipo de diseño implica muchos aspectos. A fin de estructurar el proceso de diseño

en una forma manejable, se ha condensado y esquematizado estas consideraciones en el árbol de decisiones que se presenta en las Figuras. 4.4, 4.5 y 4.6.

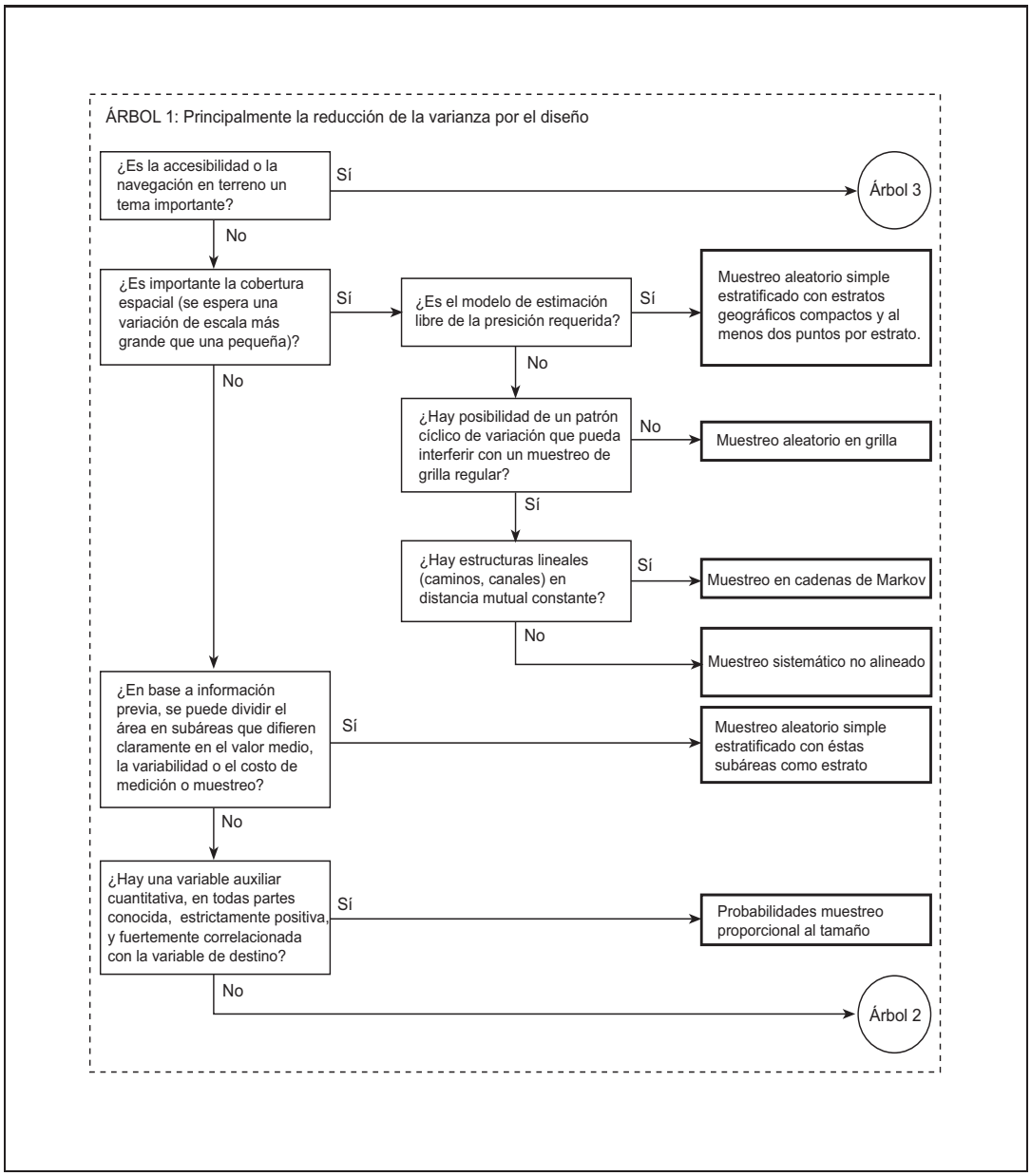


Figura 4.4: Árbol de decisiones para ayudar a la elección de un tipo de diseño para cantidades globales en el espacio.

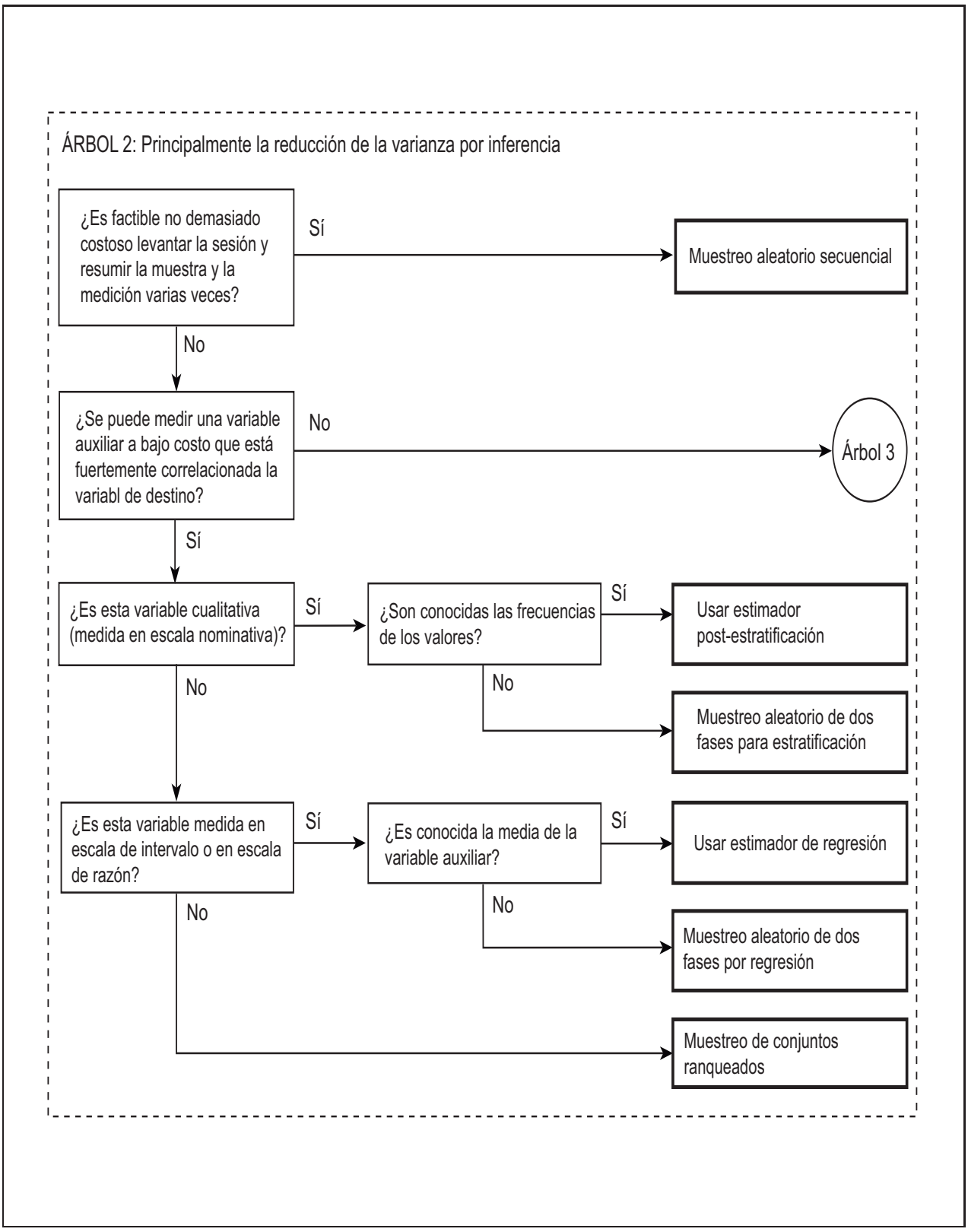


Figura 4.5: Continuación Figura 4.4

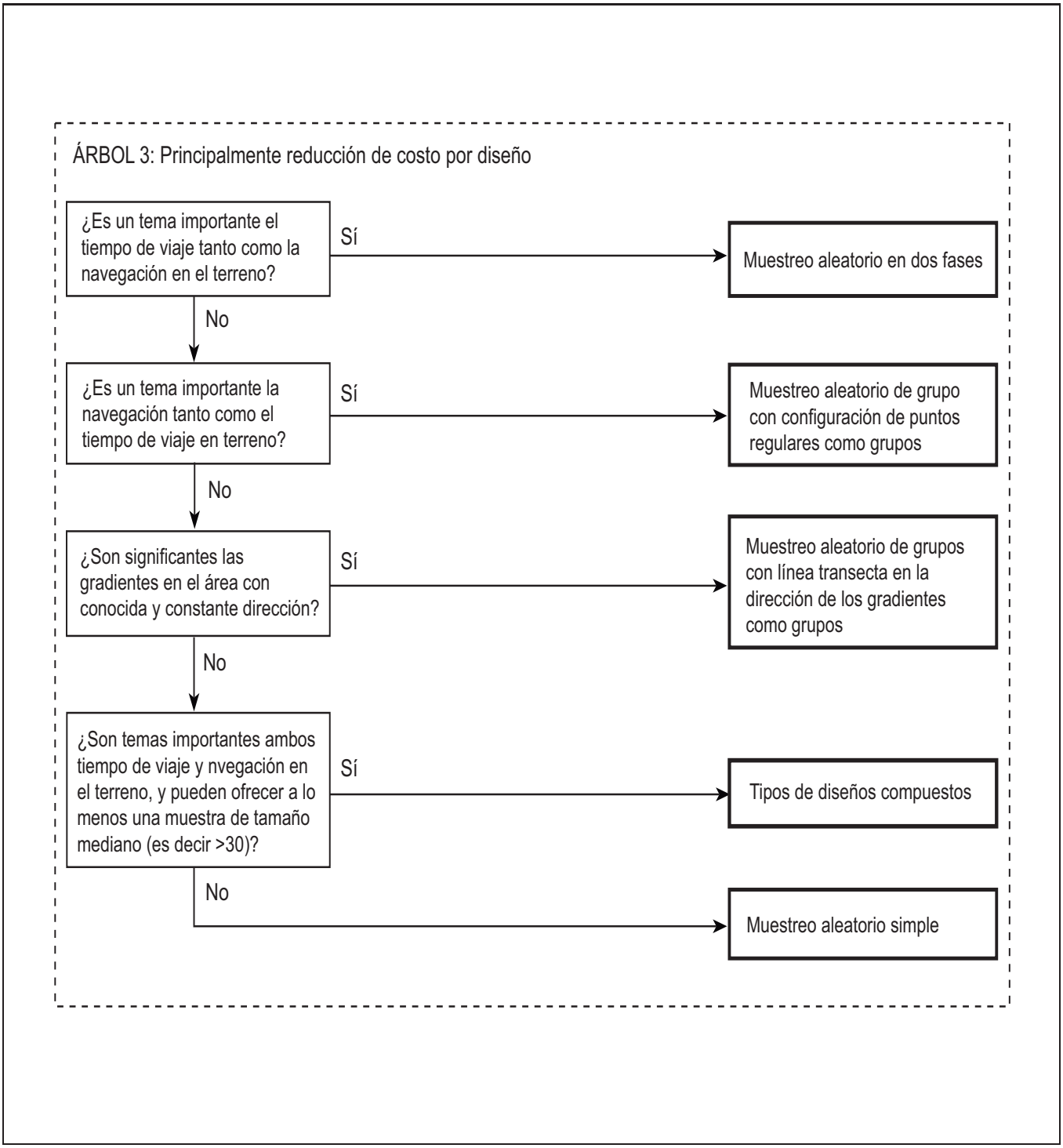


Figura 4.6: Continuación Figura 4.5

Cabe mencionar que estas figuras es el resultado de muchas simplificaciones, por lo que los resultados deben ser interpretados como sugerencias más que conclusiones. En segundo lugar, se han mantenidos algunas de las preguntas más o menos vaga, ya que la estructura lógica subyacente es inherentemente vaga. En tercer lugar, mientras que este árbol supone respuestas claras (sí o no). Si uno tiene dudas considerables acerca de la respuesta a una pregunta, le sugerimos que las dos ramas del árbol se siguen.

## 4.7 Esquemas de Muestreo Espacial

Hay varios esquemas de muestreos para la organización de $n$ puntos en el espacio. Sean $\{s_1, s_2, s_3, \ldots, s_n\}$ un conjunto de posiciones en el espacio. Los esquemas de muestreo espacial son:

### 4.7.1 Muestreo Basado en el Diseño

1. **Muestreo Aleatorio Simple:** Todos los sitios de muestreo se seleccionan con probabilidad igual e independientemente el uno del otro.

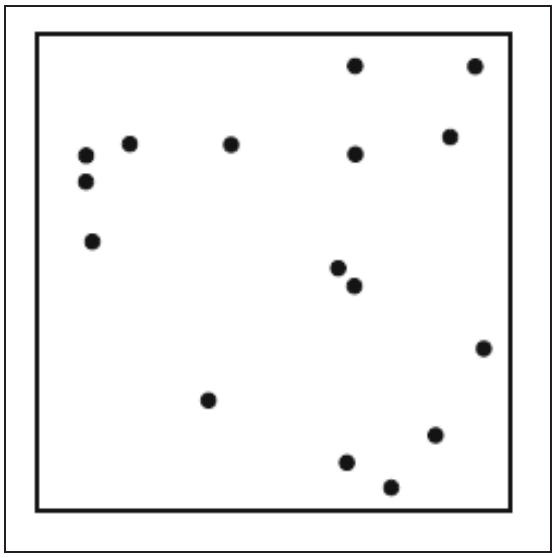

Figura 4.7: Ejemplo teórico de una muestra aleatoria simple

2. **Muestre Aleatorio Simple Estratificado:** El área se divide en sub-áreas, llamadas *estratos*, en cada uno de los que se aplica muestreo aleatorio simple con un tamaño de muestra elegida de antemano.

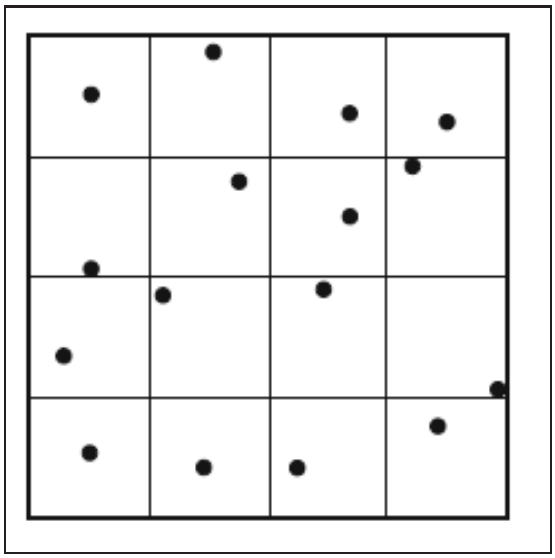

Figura 4.8: Ejemplo teórico de una muestra aleatoria simple estratificado

3. **Muestreo Muestreo en Dos Etapas:** Al igual que con muestreo aleatorio estratificado simple, el área se divide en una serie de sub-zonas. El muestreo es entonces restringido a un número de sub-áreas seleccionados al azar, en este caso las llamadas unidades primarias.

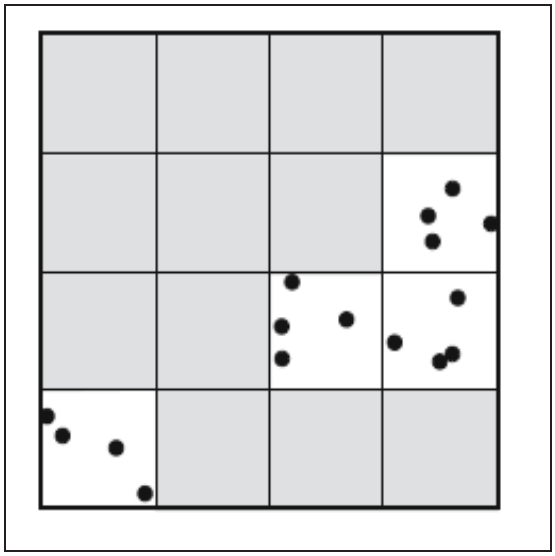

Figura 4.9: Ejemplo teórico de una muestra en dos etapas.

4. **Muestreo por Conglomerados:** Los grupos seleccionados son un conjunto predefinidos, en cada grupo se puede aplicar un muestreo aleatorio simple, muestreo aleatorio estratificado simple y muestreo aleatorio de dos etapas.

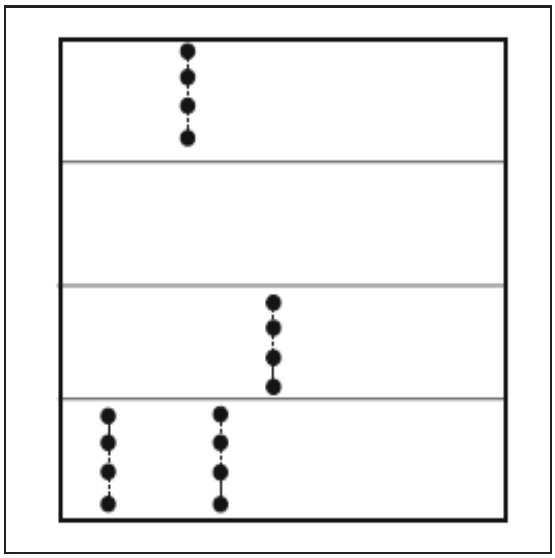

Figura 4.10: Ejemplo teórico de una muestra en grupos.

5. **Muestreo Aleatorio Sistemático:** en este caso el muestreo aleatorio sistemático es un caso especial de muestreo aleatorio por conglomerados.

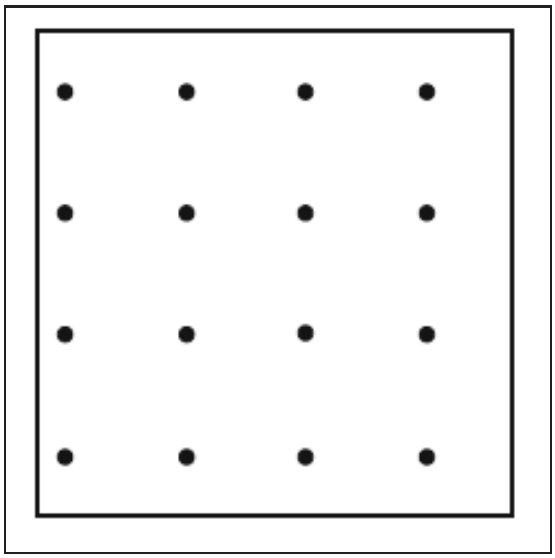

Figura 4.11: Ejemplo teórico de una muestra sistemática.

### 4.7.2 Muestreo Basado en el Modelo

1. **Muestreo en Rejilla Centrada:** Tres son los aspectos que se deberá decidir: la forma de las celdas de la grilla, el tamaño de las celdas de la grilla, y la dirección de la red. Con respecto a la forma, hay tres opciones: cuadradas, triangulares o hexagonales.

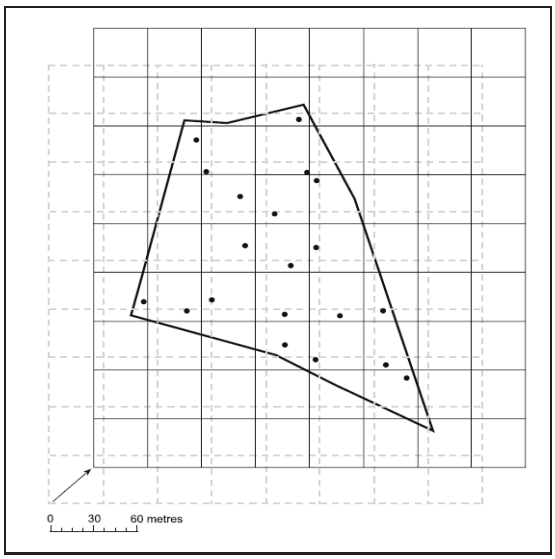


Figura 4.12: Ejemplo teórico de una muestra en rejilla centrada.

2. **Muestreo Geoestadístico:**


Figura 4.13: Ejemplo teórico de una muestra geoestadística.

3. **Muestreo Estratificado de Teselación Hexagonal:** Para aplicar este tipo de muestreo se debe considerar los siguientes elementos; el radio $r_h$, de un hexágono deseado puede ser aproximado calculando la cantidad $\sqrt{\frac{2}{3\sqrt{3}}\frac{area}{n}}$, donde *area* denota el área de estudio. El valor producido por este cálculo es en realidad un límite superior para el radio deseado.

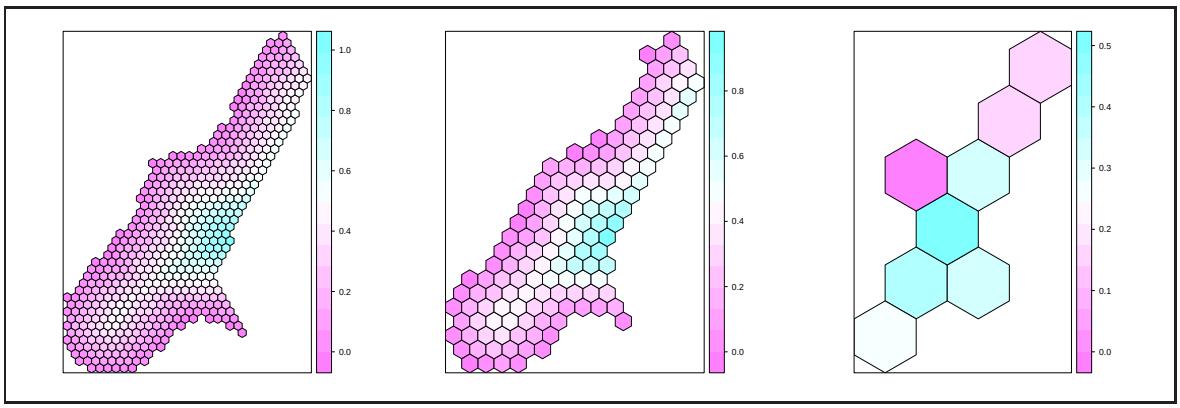

Figura 4.14: Código Fuente en R, extraído de Bivand et al. (2008). Applied Spatial Data Analysis Spatial with R. Pág 137

## 4.8 Comentarios

Este Capítulo muestra varias perspectivas presentadas por varios autores con respecto al muestreo espacial. Los cálculos asociados a las diferentes planes de muestreo, se pueden encontrar en algunos de los siguiente trabajos:

En Ripley, B, D. (1981). *Spatial Statistics*. Presenta varios esquemas de muestreo, en el se estudian las características de implementar un tipo muestreo, en el se realizan y presentan los cálculos asociados al estimador muestral de $\mu$ y varianza del estimador muestral o error muestral.

Otro ejemplo se puede encontrar en Griffith, D. (2008). *Geographic sampling of urban soils for contaminant mapping: how many samples and from where*. Este paper aborda los temas de eficiencia asociadas con estos resultados de muestreo espacial basados en modelos. En él se resumen los resultados de una colección de datos de actividad de recolección de (soil samples collected from across Syracuse, NY). Así como un conjunto de muestreo y de experimentos de simulación para los principios de diseño experimental enunciado por Overton y Stehman (in Communications in Statistics: Theory and Methods, 22, 2641-2660). Las directrices que se sugieren en cuanto al tamaño apropiado de la muestra (es decir, la cantidad) y la red de muestreo (es decir, dónde). Además, presenta el método de Teselación de muestreo aleatorio estratificado: Particiones Hexagonal de la Superficie como un método eficiente de selección de muestras.

Por otro lado, Haining, R. (2003). En sus libros *Spatial Data Analysis: Theory and Practice* y Haining, R. (1990). *Spatial Data Analysis in the social and environmental sciences*. Aquí se presenta herramientas de análisis geoespacial, modelos estadísticos espaciales. Como también, Muestreo Espacial. Presenta varios ejemplos y motivos para utilizar muestreo espacial, como también entrega referencia de cuando, como y donde realizar muestreo espacial. Analiza también el cálculo del tamaño muestral para datos georefenciados.

En el libro presentado por de Gruijter,J; Bruc, D; Bierkens, M; Knotters, M. (2006). *Sampling for Natural Resource Monitoring*. Este libro presenta una extensión de los métodos clásicos de muestreo vistas por Chocran, Thompson entre otros, aplicados al muestreo espacial, en el se presentan las diferencias de aplicar métodos basados en muestreo y en modelos. Presenta la elección de una estrategia basada en el diseño de las cantidades globales en el espacio y otras técnicas de muestreo complejos. En cada tipo de muestreo se presentan los algoritmos de aplicación y los resultados asociados al estimador de interés, al igual de varios ejemplos y fórmulas asociadas a cada tipo de muestreo.

Bivand, R;Pebesma, E; Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R. Springer*. Presenta algunas técnicas de muestreo y como manipular datos en el software de libre distribución R, con ejemplos y rutinas para replicar. Herramientas visuales de datos espaciales. Importación y Exportación de datos espaciales. Análisis de Patrones de puntos Espaciales. Creación de mapas y cartografía entre otros tópicos, asociados al análisis espacial de datos.

Si se desea profundizar más sobre las distintas técnicas, se pueden consultar la referencias bibliográficas mencionadas.

Capítulo 5

# Simulación del Tamaño Muestral Efectivo ESS

## 5.1 Introducción

En los Capítulos anteriores se ha planteado la estructura que subyace a los datos con atributos georeferenciados, al igual que sus propiedades. Así mismo, modelos paramétricos que pueden representar la información contenida en la correlación espacial y algunos métodos de estimación de los parametros.

Seguidamente, se ha estudiado algunos aspectos de como la literatura ha presentado la correlación espacial como un elemento importante en el cálculo del tamaño muestral y como esta información está contenida en algunas expresiones analíticas para determinar el tamaño muestral efectivo. Se han estudiado aspectos asintóticos del tamaño muestral efectivo $ESS$ cuando se tiene una estructura de correlación con patrones conocidos, como es el caso de estructuras de correlación intraclase y correlación autoregresivo de primer orden. Además, se estudiarán los modelos Esféricos, Exponencial, Gaussiano y Mathérn. Esto se realizará vía simulación para ver el comportamiento del cálculo del tamaño muestral efectivo.

## 5.2 Algoritmos para simular ESS

Sea $\mathbb{Y} = \mu\mathbf{1} + \epsilon^*$, $\epsilon^* \sim N(0, \Sigma)$, el modelo inicial a simular donde $\mu$ es la media general del proceso y $\Sigma$ es la matriz de correlación espacial. Dependiendo de la estructura que tenga $\Sigma$ es la correlación espacial que se va a usar. Por ejemplo:

1. si $\Sigma = \sigma^2\Sigma(\rho)$ la correlación espacial es intraclase, es decir, $R(\rho)$ ver (3.49),
2. si $\Sigma = \sigma^2\Sigma(\phi)$ la correlación espacial es autoregresiva de primer orden, es decir, $R(\phi)$ ver (3.50),
3. si $\Sigma = \Sigma_{Exp}(\theta)$ la correlación espacial es exponencial, es decir, $R_{Exp}(\theta)$ ver (3.70),
4. si $\Sigma = \Sigma_{Esf}(\theta)$ la correlación espacial es esférica, es decir, $R_{Esf}(\theta)$ ver (3.71),
5. si $\Sigma = \Sigma_{Mat}(\theta)$ la correlación espacial es de Mathérn, es decir, $R_{Math}(\theta)$ver (3.72),
6. si $\Sigma = \Sigma_{Gaus}(\theta)$ la correlación espacial es gaussiana, es decir, $R_{Gaus}(\theta)$ ver (3.73).

### 5.2.1 Algoritmo N°1

Para los casos 1 y 2 se utiliza el siguiente algoritmo:

**Paso 1:** *Establecer la matriz $\Sigma$, fijando $\sigma^2$, $\mu$ y $\rho$*

**Paso 2:** *Generar una variable aleatoria $Y_{n\times 1} \sim N_{n\times 1}(\mu 1, \boldsymbol{R})$*

**Paso 3:** *Estimar $\rho$ y ESS*

**Paso 4:** *Repetir los pasos 1-3 K veces con* $K = 2000$

### 5.2.2 Algoritmo N°2

Por otro lado, para los casos del 3 al 6 se realizará utilizando el siguiente algoritmo:

**Paso 1:** *Fijar las localizaciones* $\{s_1, s_2, \ldots, s_n\}$ *en el espacio e incorporarlas en* $\gamma(h_{ij}) = f(\|h\| = \|s_i - s_j\|)$,

**Paso 2:** *Seleccionar Modelo* $\gamma(h_{ij}) = f(\|h\| = \|s_i - s_j\|)$ *Exponencial, Gaussiano, Esférico,etc. Introducir valores iniciales a* $\theta$, *donde* $\theta = (\sigma^2, \tau^2, \phi)$

**Paso 3:** *Calcular matriz de correlación* $R(\theta)$ *(Tomando en cuenta los pasos 1 y 2)*

- *Simular* $Y_{n\times 1} \sim N(\mathbf{1}\mu, \Sigma(\theta))$ $Y_{n\times 1} = \{Y(s_1), Y(s_2), \ldots, Y(s_n)\}$
- *Estimar* $\hat{\theta} = (\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi})$ *mediante estimación REML*
- *Obtener* $\hat{R}(\hat{\theta})$

**Paso 4:** *Calcular* $ESS = \mathbf{1}^t \hat{R}(\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi})^{-1} \mathbf{1}$

**Paso 5:** *repetir Paso 1- 4 K veces con* $K = 2000$

Estos algoritmos son distintos porque en los casos donde la correlación espacial es intraclase y autoregresiva de primer orden, se tienen expresiones analíticas que simplifican los cálculos. En los otros casos no se tiene una expresión analítica.

## 5.3 Tamaño Muestral Efectivo con distintas Correlaciones Espaciales

### 5.3.1 Correlación Intraclase $ESS_{intra}(n, \rho)$

Ahora se analizará mediante simulación el cálculo del tamaño muestral efectivo, cuando la muestra con atributo georefenciados tienen una estructura de correlación intraclase ver (3.49). Para el cálculo del tamaño muestral se consideraron tres valores de correlaciones intraclases; $\rho_1$ =0.8, $\rho_2$ =0.3, $\rho_3$ =0.01, $\mu = 2$ y $\sigma^2 = 1$. Utilizando el algoritmo (5.2.1) se obtienen los siguientes resultados.

| $\widehat{ESS}_{intra}$ c/ | $K$ | Media | Des.Est | Mediana | Mín | Máx | Rango | Sesgo | Curtosis | Error Est |
|---|---|---|---|---|---|---|---|---|---|---|
| $\rho_1$ =0.8 | 2000 | 1.25 | 0.01 | 1.25 | 1.21 | 1.3 | 0.08 | 0.19 | 0.17 | 0 |
| $\rho_2$ =0.3 | 2000 | 3.32 | 0.1 | 3.32 | 3.01 | 3.67 | 0.67 | 0.17 | 0.03 | 0 |
| $\rho_3$ =0.01 | 2000 | 83.69 | 3.68 | 83.55 | 71.59 | 96.95 | 25.36 | 0.15 | -0.1 | 0.08 |

Tabla 5.1: Estadísticas Descriptivas para $\widehat{ESS}_{intra}(n, \widehat{\rho}_K)$ con $n = 500$, $K = 2000$, $\rho_1$ =0.8, $\rho_2$ =0.3, $\rho_3$ =0.01, $\mu = 2$ y $\sigma^2 = 1$

En base a las 2000 simulaciones se puede observar que la media muestral de $\widehat{ESS}_{intra}(n, \widehat{\rho}_{1_K})$ es 1.25 unidades experimentales, con una desviación estándar respecto de su media es 0.01. El valor de la mediana es 1.25. Los valores extremos que puede alcanzar $\widehat{ESS}_{intra}(n, \widehat{\rho}_{1_K})$ son 1.22 y 1.28 como mínimo y máximo respectivamente. El rango de $\widehat{ESS}_{intra}(n, \widehat{\rho}_{1_K})$ es 0.06. La curtosis es -0.36 la cual esta cercana a cero, lo que evidencia una distribución simétrica y casi igualmente apuntada que una distribución normal. Es evidente que la interpretación para los otros resultados es la misma. Además, se puede ver que a medida que la correlación intraclse disminuye el tamaño muestra efectivo aumenta.

### 5.3.2 Correlación Autoregresiva de Primer Orden (AR(1)) $ESS_{AR(1)}(n, \phi)$

Otro ejemplo para el cálculo del tamaño muestral efectivo se tiene cuando la muestra con atributo georefenciados tienen una estructura de correlación autoregresiva de primer orden ver (3.50). Para el cálculo del tamaño muestral se consideraron tres correlaciones; $\phi_1$ =0.7, $\phi_2$ =0.3, $\phi_3$ =0.1, $\mu = 2$ y $\sigma = 1$. Aplicando el algoritmo (5.2.1) se obtienen los siguientes resultados.

| $\widehat{ESS}_{AR1}$ c/ | $K$ | Media | Des.Est | Mediana | Mín | Máx | Rango | Sesgo | Curtosis | Error Est |
|---|---|---|---|---|---|---|---|---|---|---|
| $\phi_1$ =0.7 | 2000 | 90.69 | 11.05 | 90.05 | 54.54 | 148.22 | 93.68 | 0.41 | 0.62 | 0.25 |
| $\phi_2$ =0.3 | 2000 | 272.8 | 25.77 | 272.12 | 200.03 | 381.86 | 181.83 | 0.26 | 0.13 | 0.58 |
| $\phi_3$ =0.1 | 2000 | 413.03 | 36.53 | 411.24 | 311.25 | 540.53 | 229.27 | 0.26 | 0.09 | 0.82 |

Tabla 5.2: Estadísticas Descriptivas para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_K)$, con $\phi_1$ =0.7, $\phi_2$ =0.3, $\phi_3$ =0.1, $\mu = 2$ y $\sigma = 1$, $K = 2000$ y $n = 500$

En base a las 2000 simulaciones se puede obtener un coeficiente $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{1_K})$ promedio de 90.69, con una desviación promedio respecto de su media de 11.05. Se puede observar que los distintos valores de $\hat{\rho}$ fluctúa entre 54.54 y 148.22 como mínimo y máximo respectivamente, dando un rango de 93.68. La mediana esta cercana al promedio muestral, lo que nos da información sobre una distribución simétrica. Es evidente que la interpretación para los otros resultados es la misma. De igual forma que en la simulación anterior a medida que el parametro $\phi$ se acerca a cero, el tamaño muestral efectivo aumenta.

### 5.3.3 Correlación Espacial $ESS(n, \theta)$

Finalmente, para cálculo del tamaño muestral efectivo cuando existe correlación espacial, ya sea, Exponencial, Esférica y Mathérn, con vector de parámetros $\theta = (\tau^2, \sigma^2, \phi)^t$ los cuales serán estimados mediante Máxima Verosimilitud Restringida (REML). Consideremos una muestra de tamaño $n = 500$ y los parámetros del vector $\hat{\theta}$; $\sigma^2 = 1$, $\tau^2$ =0.5, $\phi$ =0.25 y $\mu = 100$. Utilizando el algoritmo (5.2.2) se obtienen los siguientes resultados.

| | $K$ | Media | Des.Est | Mediana | Mín | Máx | Rango | Sesgo | Curtosis | Error Est |
|---|---|---|---|---|---|---|---|---|---|---|
| $\widehat{ESS}_{Exp}(n, \hat{\theta})$ | 1000 | 16.63 | 11.45 | 7.99 | 2.16 | 112.46 | 110.29 | 2.55 | 11.23 | 0.36 |
| $\widehat{ESS}_{Esf}(n, \hat{\theta})$ | 1000 | 122.74 | 33.94 | 121.91 | 28.48 | 268.88 | 240.4 | 0.22 | 1.24 | 1.07 |
| $\widehat{ESS}_{Mat}(n, \hat{\theta})$ | 1000 | 47.56 | 33.49 | 39.32 | 2.1 | 234.14 | 232.04 | 1.76 | 4.63 | 1.06 |

Tabla 5.3: Estadísticas Descriptivas para cada $\widehat{ESS}_{esp}(n, \widehat{\theta})$ con distintos modelos, $n = 500$ y $K = 1000$

La tabla (5.3), muestra los resultados de las 1000 simulaciones y se puede observar que la cantidad $\widehat{ESS}_{Exp}(n, \hat{\theta})$ presenta una media de 16.63 unidades experimentales independientes. Que a su vez tiene asociada una dispersión de 11.45 unidades independientes. Los valores de $\widehat{ESS}_{Exp}(n, \hat{\theta})$ fluctúan entre 2.16 y 112.46, como mínimo y máximo respectivamente. El rango es de esta cantidad es de 110.29. O bien, la reducción efectiva en el tamaño muestral efectivo en presencia de correlación espacial Exponencial varía entre 0.43 % y 22.49 %. Estas 1000 realizaciones dan evidencia sobre el comportamiento del tamaño muestral efectivo y algunas características generales de la distribución que podría tener, para esto se han realizado un histograma.Ademas, se ha modelado este comportamiento con modelos probabilísticos que tengan características de asimetría y se puedan asociar a esta distribución muestral. Es evidente que la interpretación para los otros resultados es la misma.

## 5.4 Gráficos

Esta sección presenta el comportamiento del tamaño muestral efectivo gráficamente

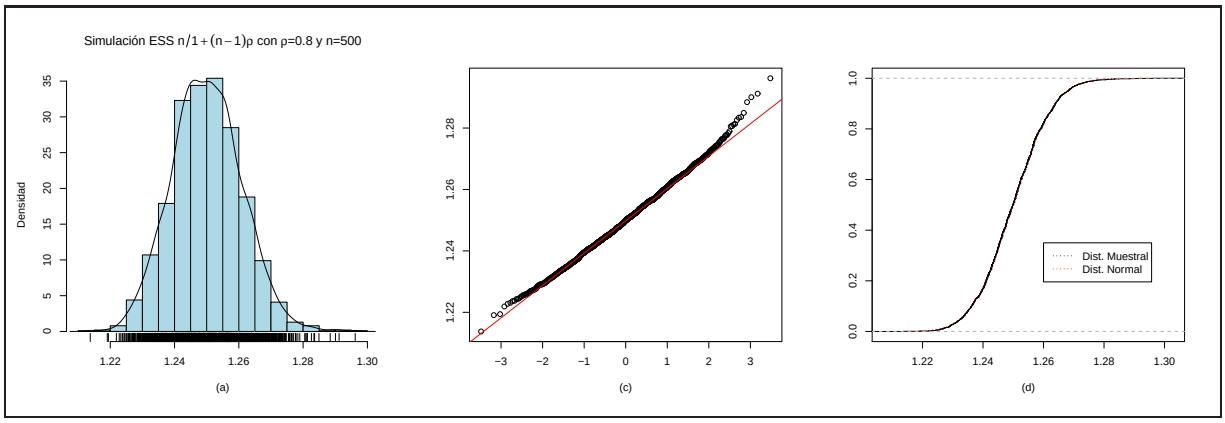


Figura 5.1: (a) Histograma para $\widehat{ESS}_{intra}(n, \hat{\rho}_{1_K})$, $\rho_1$=0.8 con $K = 2000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{intra}(n, \hat{\rho}_{1_K})$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{intra}(n, \hat{\rho}_{1_K})$ versus la distribución teórica Normal.

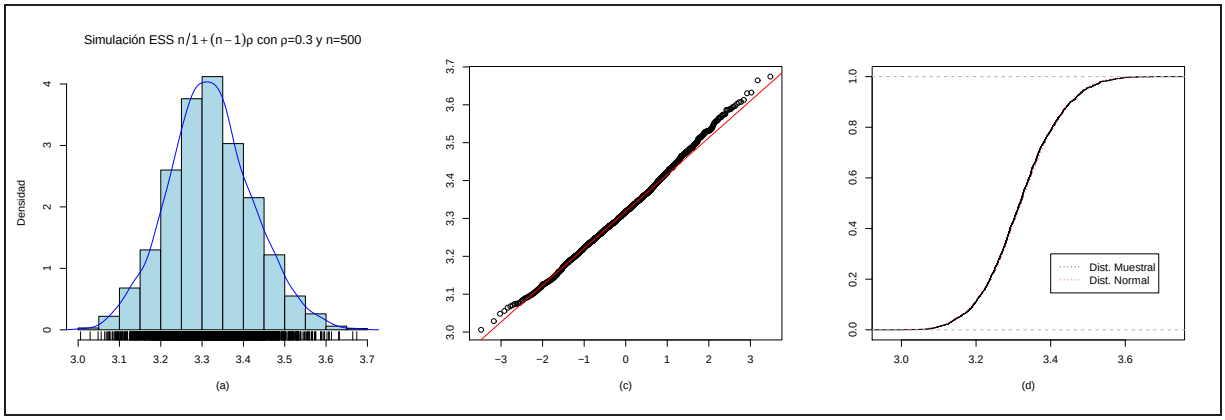


Figura 5.2: (a) Histograma para $\widehat{ESS}_{intra}(n, \hat{\rho}_{2_K})$, $\rho_2$=0.3 con $K = 2000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{intra}(n, \hat{\rho}_{2_K})$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{intra}(n, \hat{\rho}_{2_K})$ versus la distribución teórica Normal.

En la Figura (5.1) el histograma representa a $\widehat{ESS}_{intra}(n, \widehat{\rho}_{1_K})$ como una curva unimodal y evidencia una distribución simétrica. Los gráficos QQ-lines presenta la relación entre los cuantiles teóricos de la distribución normal y los cuantiles muestrales, donde se puede ver que hay una fuerte relación entre estos dos cuantiles evidenciando que la muestra sigue una ley normal. Por otro lado, si comparamos la función de distribución teórica con la muestral se puede ver cierta semejanza entre la muestral y la teórica. Las Figuras (5.2) y (5.3) presenta comportamiento similar.

En la Figura (5.4) el histograma presenta una curva unimodal y evidencia una distribución simétrica de $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{1_K})$. Las Figuras (5.5) y (5.6) presentan comportamientos similares.

En la figura (5.7), se representa el comportamiento de $\widehat{ESS}_{Exp}(n, \hat{\theta})$ mediante el histograma, el cual, presenta una curva unimodal y evidencia una distribución con sesgo positivo. Por otro lado, el Box-plot gráfica la asimetría de la distribución de las simulaciones para $\widehat{ESS}_{Exp}$. El comportamiento de esta nueva variable aleatoria la cual se compara con dos modelos probabilísticos, con la distribución Normal y Log-

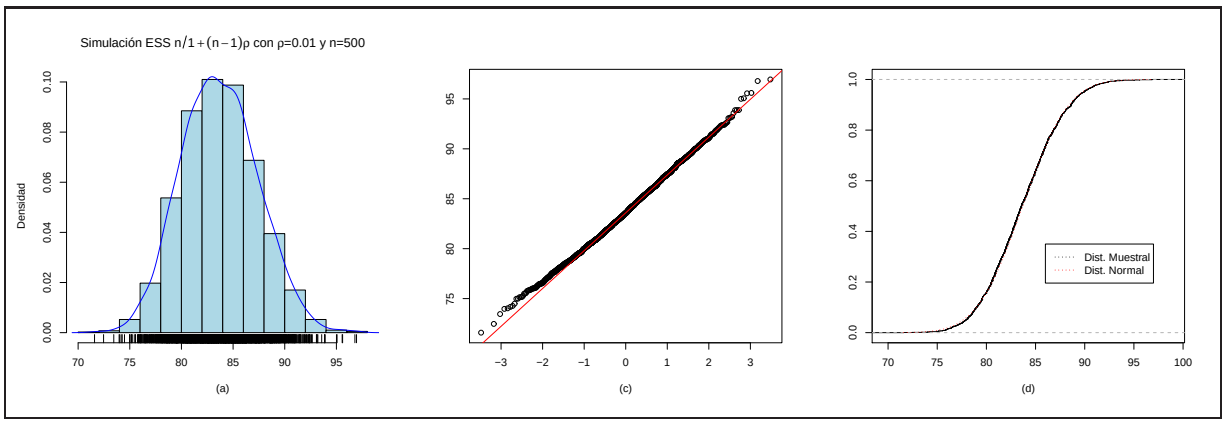


Figura 5.3: (a) Histograma para $\widehat{ESS}_{intra}(n, \hat{\rho}_{3_K})$, $\rho_1$=0.01 con $K = 2000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{intra}(n, \hat{\rho}_{3_K})$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{intra}(n, \hat{\rho}_{3_K})$ versus la distribución teórica Normal.

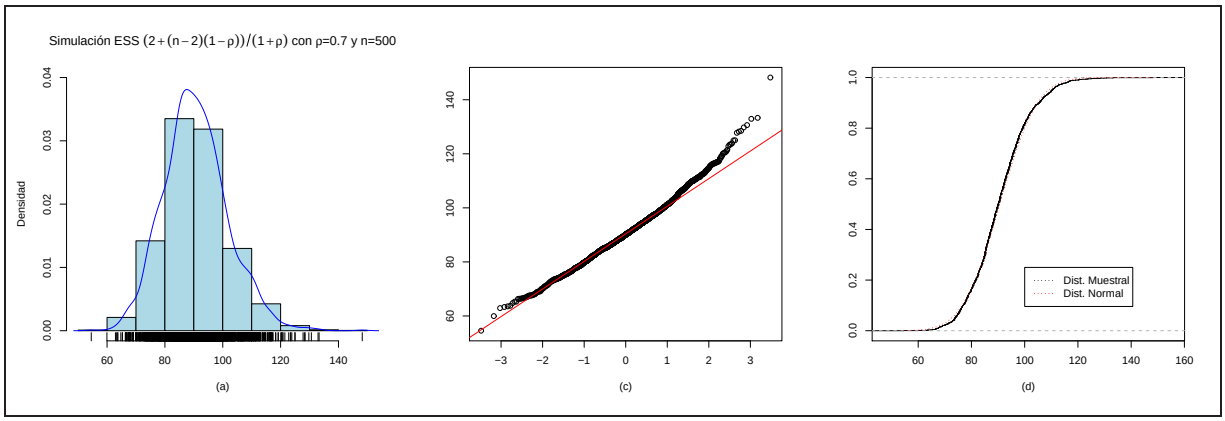


Figura 5.4: (a) Histograma para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{1_K})$, $\phi_1 = 0.7$ con $K = 2000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{1_K})$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{1_K})$ versus la distribución teórica Normal.

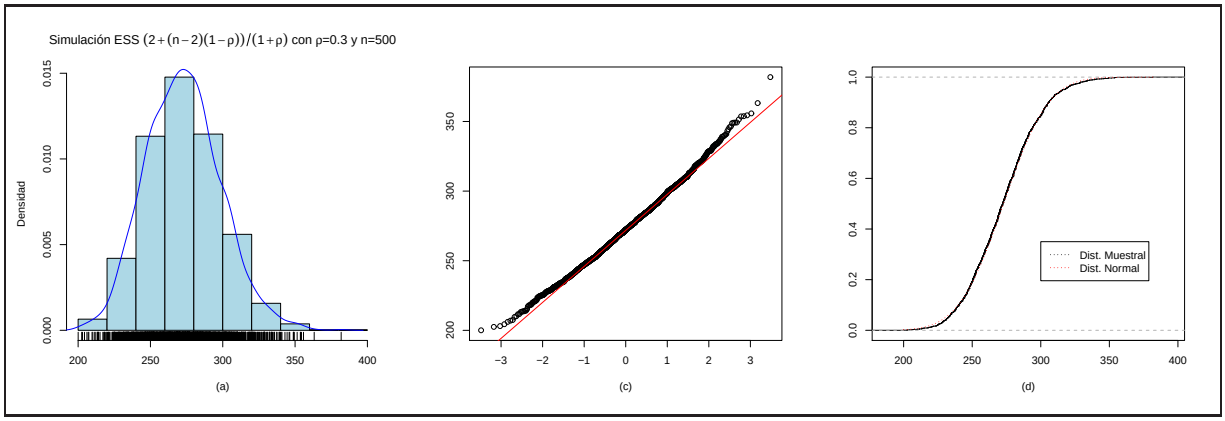


Figura 5.5: (a) Histograma para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{2_K})$, $\phi_2 = 0.3$ con $K = 2000$ y $n = 500$ , (c) QQ-Normal para para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{2_K})$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{2_K})$ versus la distribución teórica Normal.

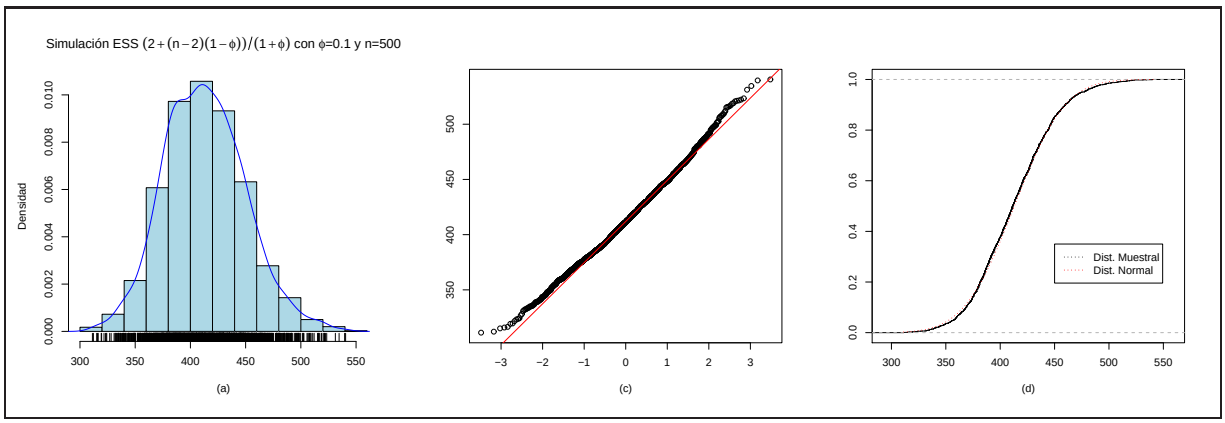


Figura 5.6: (a) Histograma para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{3_K})$, $\phi_3 = 0.1$ con $K = 2000$ y $n = 500$ , (c) QQ-Normal para para $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{3_K})$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{AR(1)}(n, \hat{\phi}_{3_K})$ versus la distribución teórica Normal.

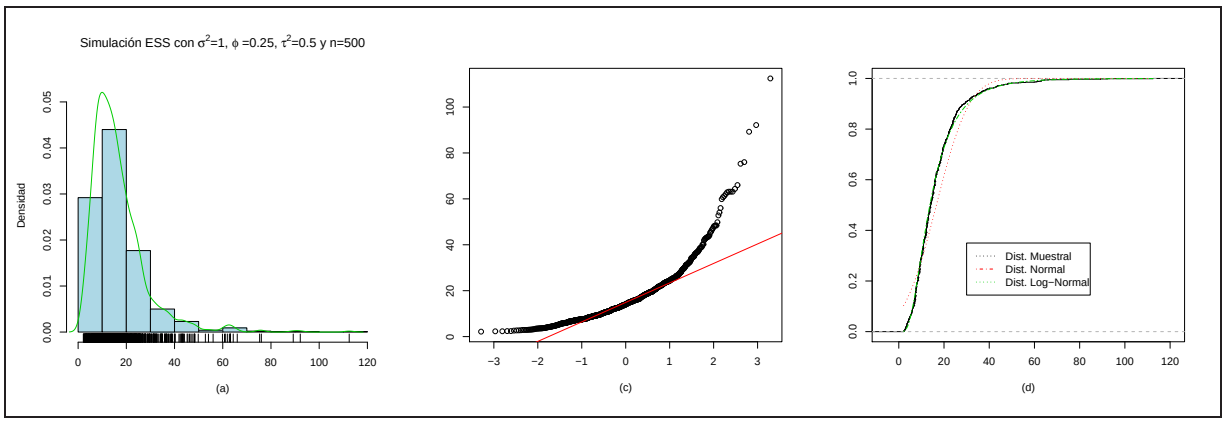


Figura 5.7: (a) Histograma para $\widehat{ESS}_{Exp}$, con $K = 1000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{Exp}$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{Exp}$ versus la distribución teórica Normal

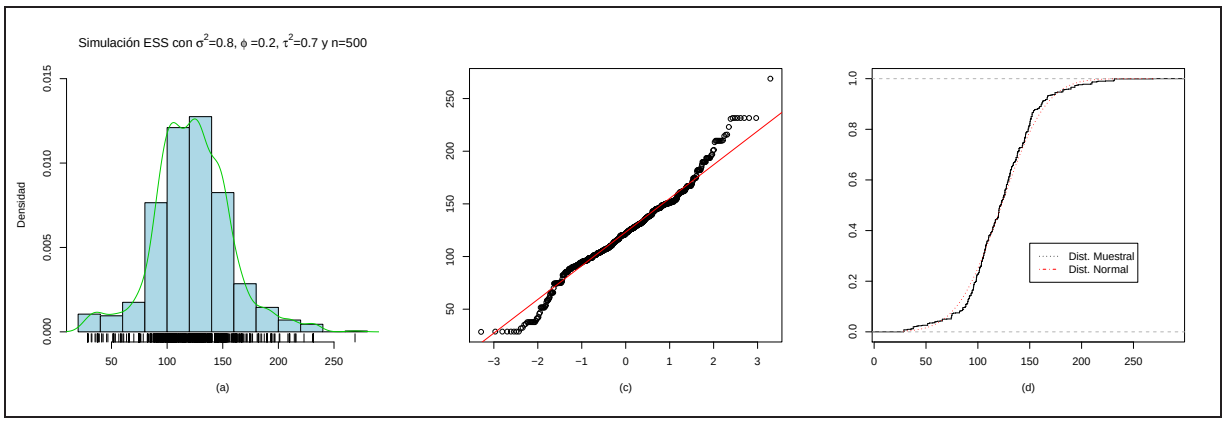


Figura 5.8: (a) Histograma para $\widehat{ESS}_{Esf}$, con $K = 1000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{Esf}$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{Esf}$ versus la distribución teórica Normal.

Normal. Los gráficos QQ-lines presenta la relación entre los cuantiles teóricos de la distribución normal y los cuantiles muestrales, donde se puede ver que los cuantiles muestrales no se asemejan a la distribución teórica

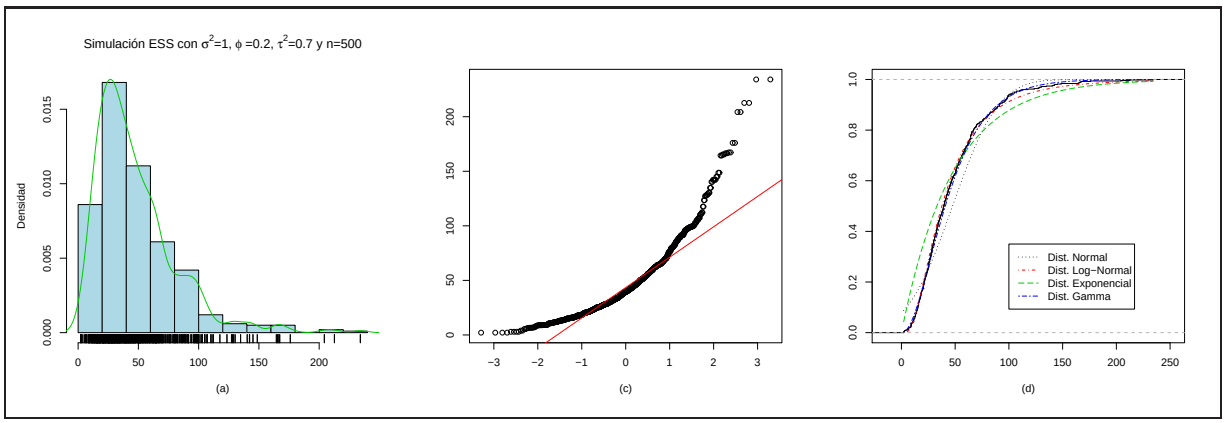


Figura 5.9: (a) Histograma para $\widehat{ESS}_{Math}$, con $K = 1000$ y $n = 500$, (c) QQ-Normal para para $\widehat{ESS}_{Math}$ , (d) Comparación entre la distribución muestral de $\widehat{ESS}_{Math}$ versus la distribución teórica Normal.

normal. Por otro lado, si comparamos la función de distribución teórica con la muestral con la distribución Log-Normal se puede ver cierta semejanza entre la muestral y la modelada.

En la Figura (5.8) el histograma presenta una curva unimodal y evidencia una distribución con sesgo positivo. Por otro lado, el Box-plot gráfica la asimetría de la distribución de las simulaciones para $\widehat{ESS}_{Esf}$. Los gráficos QQ-lines presenta la relación entre los cuantiles teóricos de la distribución normal y los cuantiles muestrales, donde se puede ver que los cuantiles muestrales no se asemejan a la distribución teórica normal. Por otro lado, si comparamos la función de distribución teórica con la muestral se puede ver cierta semejanza entre la muestral y la teórica.

En la Figura (5.9) El histograma presenta una curva unimodal y evidencia una distribución con sesgo positivo. Por otro lado, el Box-plot gráfica la asimetría de la distribución de las simulaciones para $\widehat{ESS}_{Math}$. Los gráficos QQ-lines presenta la relación entre los cuantiles teóricos de la distribución normal y los cuantiles muestrales, donde se puede ver que los cuantiles muestrales no se asemejan a la distribución teórica normal. Por otro lado, si comparamos la función de distribución teórica con la muestral se puede ver cierta semejanza entre la muestral y la teórica.

## 5.5 Comentarios

En este Capítulo se demostró computacionalmente la capacidad de reducción que tiene el tamaño muestral efectivo $ESS(\theta, n)$, cuando la muestra tiene correlación espacia. Comprobando que $ESS \leq n$. Por definición la cantidad $ESS(\theta, n)$ es el número de observaciones independientes que depende del grado de correlación espacial. Estos impactos de correlación espacial fueron representados por modelos paramétricos del semivariograma que a su vez fueron estimados mediante $REML$ y repetido 1000 veces con una nueva muestra georeferenciada independientes una de otra y estimar los parámetros, por ende, estimar la matriz de correlación espacial y calcular $ESS(\theta, n)$. Se ha podido representar la distribución muestral de $\widehat{ESS}(\widehat{\theta}, n)$ para cada uno de los modelos como son: Exponencial, Esférico, Gaussiano y Mathérn. Esta situación plantea la pregunta o querer resolver la problemática de determinar la distribución de probabilidad de $\widehat{ESS}(\widehat{\theta}, n) = \mathbf{1}^t R^{-1}(\widehat{\theta})\mathbf{1}$ lo cual en este momento es un problema abierto, ya que, consideremos $R(\widehat{\theta})$ es una matriz de covarianza espacial estocástica de dimensión $n \times n$ y determinar $R^{-1}(\widehat{\theta})$. Otra opción sería inspeccionar los elementos $i, j$-ésimo de la matriz de covarianza $R^{-1}(\theta)_{ij}$ y luego determinar la distribución $\mathbf{1}^t R^{-1}(\widehat{\theta})\mathbf{1}$. O bien, calcular $\mathbb{E}[\widehat{ESS}] = \mathbf{1}^t \mathbb{E}\left[R^{-1}(\widehat{\theta})\right]\mathbf{1}$ es un problema abierto.

# Parte III

# Estimación del Tamaño Muestral Efectivo en Jóvenes Pinos Radiata

Capítulo 6

# Aplicación

## 6.1 Introducción

A continuación se realizará la estimación del tamaño muestral efectivo a variables espaciales obtenidos del sector forestal asociada a la producción de jóvenes pinos radiata, como son el área basal ($m^2$) y la altura ($m$). Para ello se utilizarán todas las herramientas metodológicas mencionadas anteriormente. El objetivo de este capítulo es presentar una pauta para estimar el tamaño muestral efectivo $ESS$. En la sección 6.2, se presenta el Método de Aplicación de $ESS$. Luego en la Sección 6.3, se realizará una pequeña contextualización de los datos a utilizar. En la Sección 6.4, se realizará un análisis de datos exploratorio describiendo algunas estadísticas de interés. En la sección 6.5, se calculará el Tamaño Muestral Efectivo. En la Sección 6.5.2 se realizará la Selección de Muestras, para cada variable en estudio, con una breve descripción estadística. Finalmente, en la sección 6.6, se expondrán algunos comentarios sobre los resultados más importantes.

## 6.2 Método de Aplicación para ESS

En esta sección se detallan los pasos a seguir en la estimación del tamaño muestral efectivo. Considere la siguiente estructura:

**Paso 1:** *Asumamos que nuestro modelo es* $Y(s_i) = \mu \mathbf{1} + \epsilon(s_i)$

**Paso 2:** *Ajustar alguno de estos modelos con correlación Espacial:*

- *Modelo Exponencial,*
- *Modelo Gaussiano,*
- *Modelo Esférico,*
- *Modelo Mathérn.*

**Paso 3:** *Calcular* $\widehat{ESS}(\widehat{\theta}, n) = \mathbf{1}^t \Sigma^{-1}(\widehat{\theta}_{REML})\mathbf{1}$.

**Paso 4:** *Realizar selección de las unidades muestrales georefenciadas mediante muestreo aleatorio simple.*

**Paso 5:** *Estimar estadísticas de interés (Media, desviación estándar, mínimo, máximo, rango, sesgo, curtosis y error estándar (se)).*

Previo a la aplicación de $ESS$ se presenta el contexto de donde fueron obtenido estos datos; variables en estudio, su ubicación geográfica, área de cobertura y número de observaciones.

## 6.3 Variables de producción en plantaciones de jóvenes Pinos Radiata

Las variables espaciales en estudio son:

- $Y_1(s)$ : Área Basal en $m^2$,
- $Y_2(s)$ : Altura en $m$.

Los datos espaciales se obtuvieron mediante la tecnología de posicionamiento global GPS. El sitio de estudio tiene una superficie de 1.244,43 hectáreas, está ubicado en el sector del ≪Escuadrón≫, al sur de Concepción en la porción sur de Chile ($36^o\ 54'S$,$73^o\ 54'O$) y pertenece a la empresa Forestal MININCO S.A.

El interés aquí es el área que contiene pinos jóvenes (es decir, cuatro años de edad) de las plantaciones de Pinos Radiata D. Con una densidad media de 1.600 árboles por hectárea.

La geomorfología general del sitio corresponde a un rango lineal de las montañas, *la Cordillera de la Costa*, que está compuesto principalmente por una topografía abrupta y llega a una elevación de 500 $msnm$.

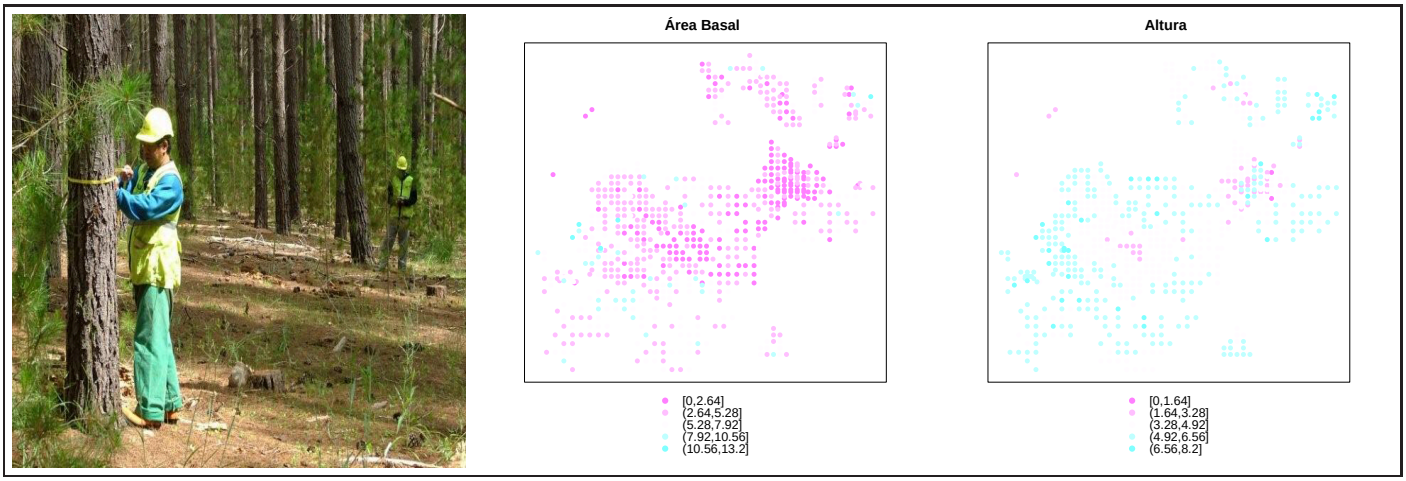


Figura 6.1: Técnico Forestal realizando medición Área Basal $m^2$ y Altura $m$ de Pinos Jóvenes Radiata y su distribución en el espacio.

El paisaje está dominado por perfiles topográficos escarpados de pequeñas dimensiones y el sitio es representativo de las plantaciones de pinos radiata existentes en las condiciones de montaña cerca de la costa en la parte centro-sur de Chile.

En la Figura (6.1), en la izquierda se muestran como los técnicos forestales realizan la medición de los pinos radiata, al lado derecho, como están distribuidas espacialmente las 688 observaciones para el área basal en $m^2$ y la altura en $m$ respectivamente.

## 6.4 Análisis de datos Exploratorio Espacial

En esta sección se realizarán estadísticas descriptivas a las unidades pertenecientes al Área Basal $m^2$ y Altura medida en metros($m$). También, se utilizarán herramientas gráficas para representar el comporta-

miento espacial.

| | $n$ | Media | Desv. Est | Mediana | Mínimo | Máximo | Rango | Sesgo | Curtosis | Error Estándar |
|---|---|---|---|---|---|---|---|---|---|---|
| $Y_1$ | 688 | 4.32 | 2.1 | 4.1 | 0 | 13.2 | 13.2 | 0.64 | 0.32 | 0.08 |
| $Y_2$ | 688 | 4.8 | 1.04 | 4.8 | 0 | 8.2 | 8.2 | -0.06 | 0.28 | 0.04 |

Tabla 6.1: Estadísticas Descriptivas para Área Basal $m^2$ y Altura $m$

En base a las 688 observaciones se puede obtener una área basal media de 4.32 $m^2$, con una dispersión respecto de su media de 2.1 $m^2$. El 50 % de los datos acumulan hasta 4.1 $m^2$ y desde ahí el otro 50 % hacia arriba. Los valores extremos fluctúan entre 0 y 13.2 como mínimo y máximo respectivamente. Se puede apreciar que el área basal tiene un pequeño sesgo a la izquierda.

Por otro lado, la altura media es de 4.8 $m$, con una dispersión respecto de su media de 1.04 $m$. El 50 % de los datos acumulan hasta 4.8 $m$ y desde ahí el otro 50 % hacia arriba. Los valores extremos fluctúan entre 0 y 8.2 como mínimo y máximo respectivamente. Se puede apreciar que la altura tiene una distribución simétrica.

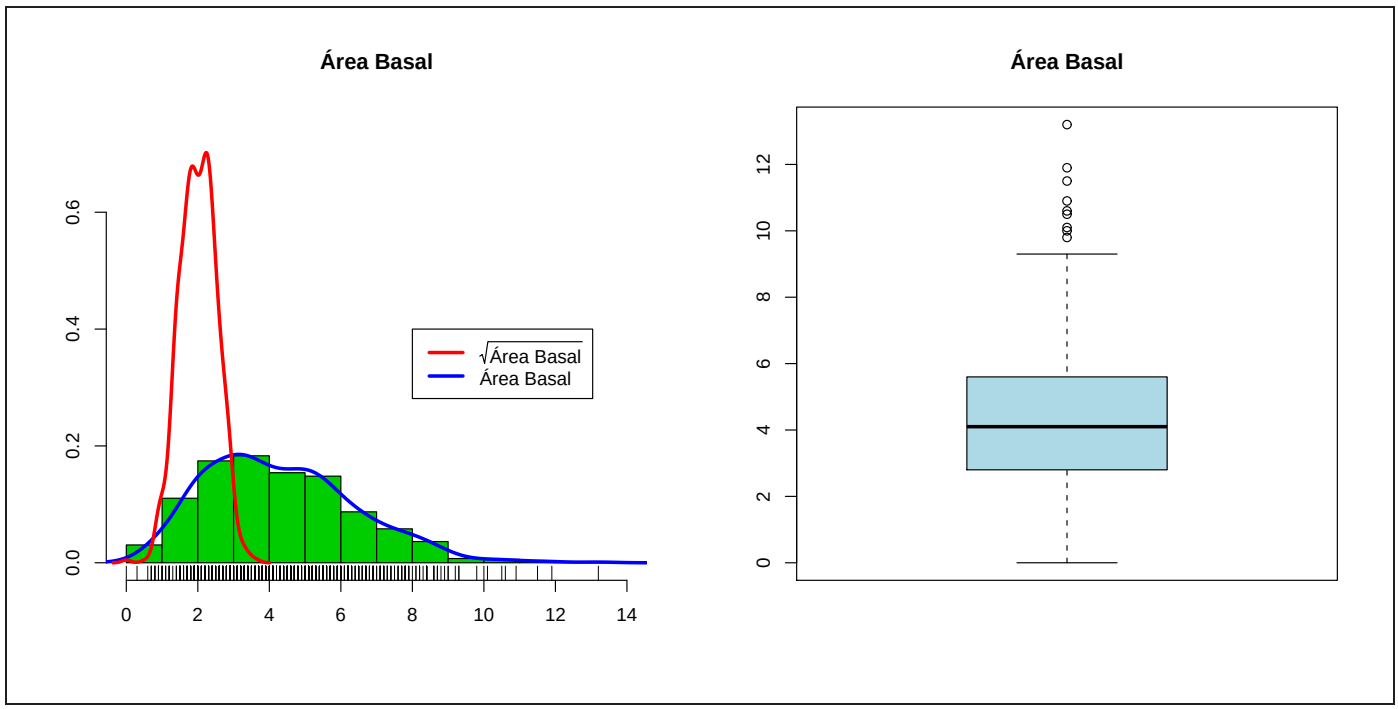


Figura 6.2: Histogramas y boxplot para Área Basal.

En la figura 6.2 se muestran la distribución del área basal está presenta una forma asimétrica. La linea roja corresponde a una transformación ($\sqrt{\cdot}$) que se realizo al área basal, donde claramente se puede observar una distribución simétrica. En cambio, la distribución del Altura presenta una forma simétrica y no se le ha hecho ninguna transformación.

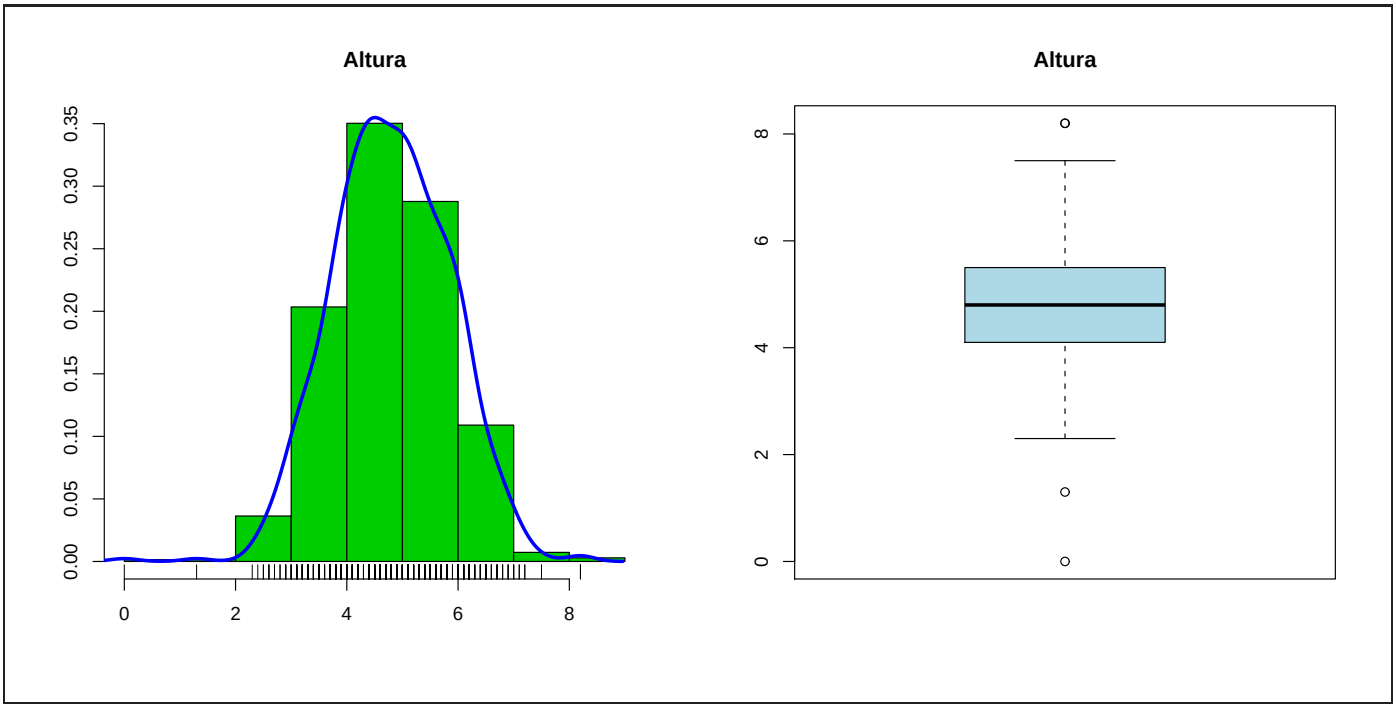


Figura 6.3: Histogramas y boxplot para Altura.

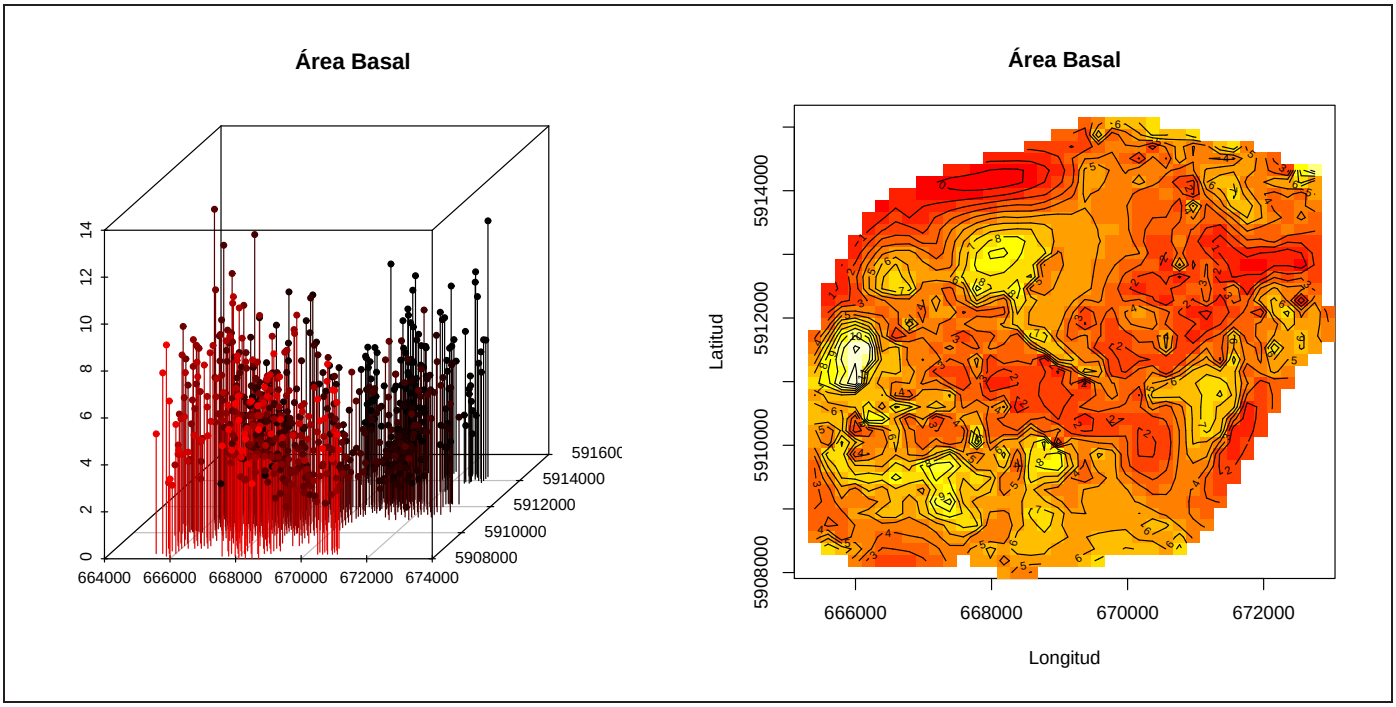


Figura 6.4: Puntos 3D y Mapa de la distribución de la Área Basal.

Las Figuras 6.4 y 6.5 muestran la región en donde se observó la mayor cantidad respecto del Área Basal $m^2$ y la Altura en $m$, con un color amarillo y con los valores que indica el contorno en la región.

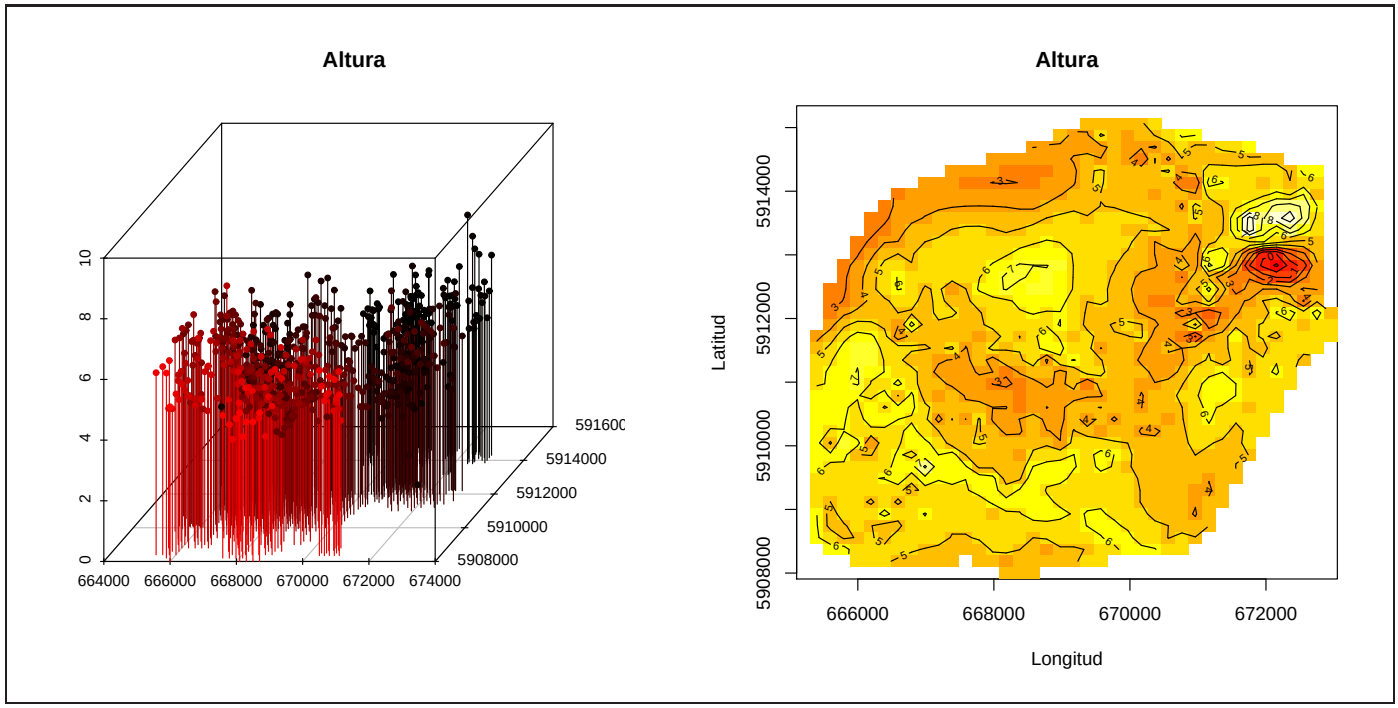


Figura 6.5: Puntos 3D y Mapa de la distribución de la Altura.

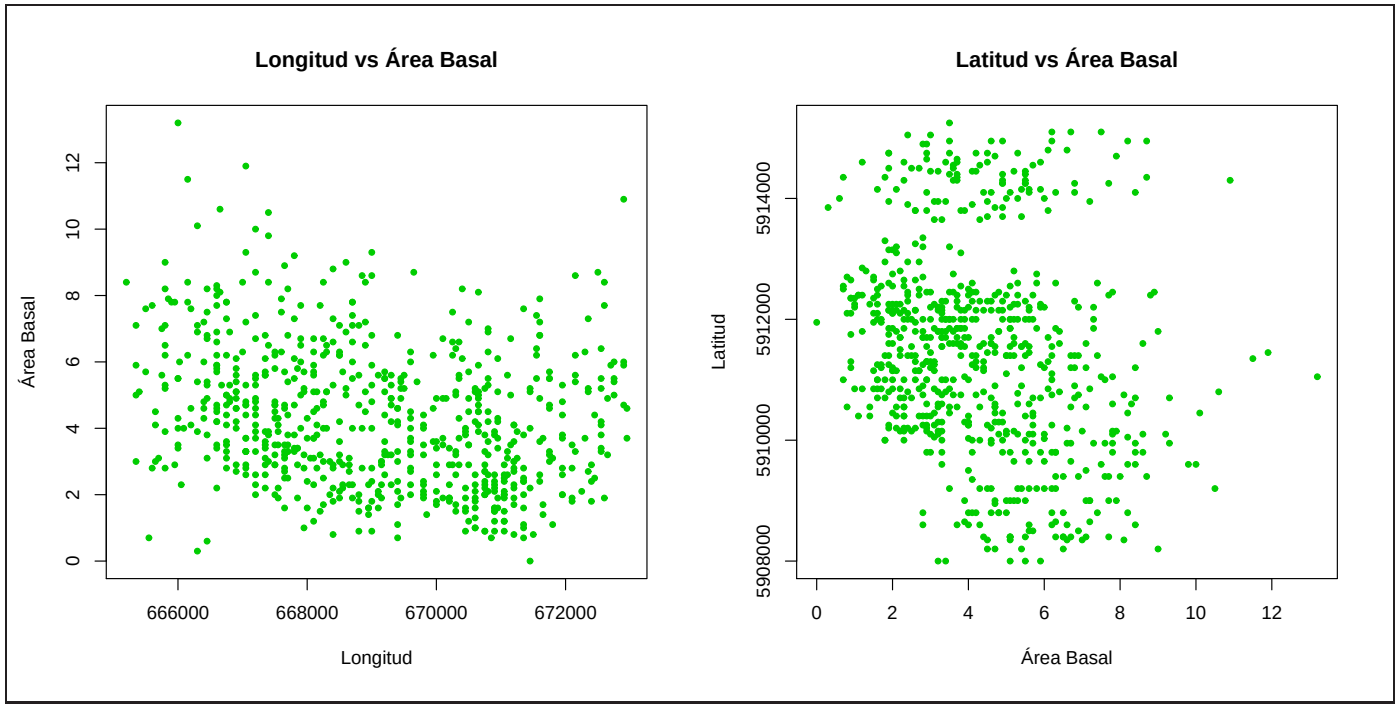


Figura 6.6: Dispersión Longitud-Latitud para Área Basal y Altura.

Las Figuras 6.6 y 6.7 muestran la dispersión entre latitud-longitud para el área basal $m^2$ y la altura $m$, en donde se puede observar que en ambos casos no hay algún tipo de relación lineal, es decir, no hay una relación entre la latitud-longitud para el área basal y altura.

La Figura 6.8 muestra los datos desde perspectiva, se realizó una interpolación entre los datos de la área basal y altura con sus correspondientes coordenadas. Aquí se aprecia claramente la existencia de una

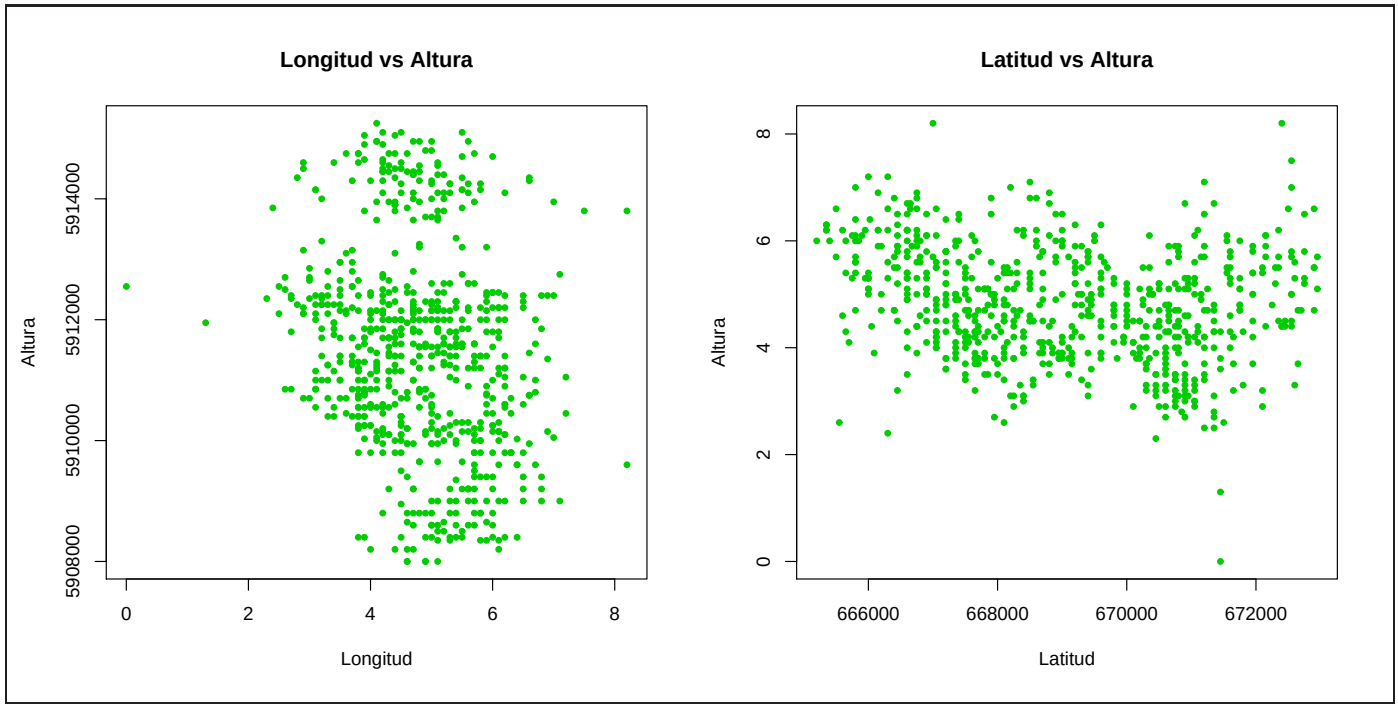


Figura 6.7: Dispersión Longitud-Latitud para Área Basal y Altura.

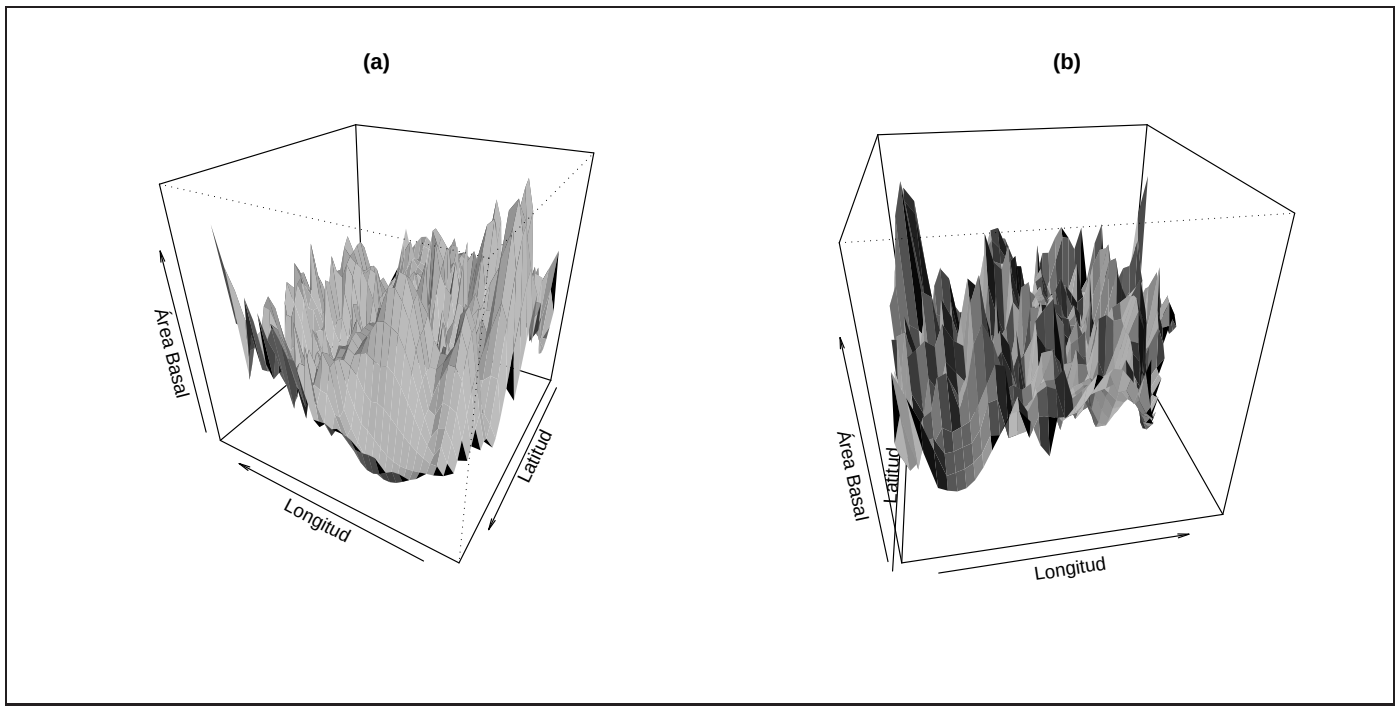


Figura 6.8: Distribución espacial de Área Basal y Altura desde distintos ángulos de visión.

zona de extracción del recurso forestal en comparación con la totalidad del área en estudio.

## 6.5 Estimación Tamaño Muestral Efectivo

En la sección anterior se han podido destacar las principales características de área basal y la altura de los pinus radiata, en esta sección se realizara la estimación del tamaño muestral efectivo.

### 6.5.1 Selección de Modelos, Estimación de Parámetros y Cálculo del Tamaño Muestral Efectivo

Para las variables espaciales área basal y altura se tienen los siguientes resultados:

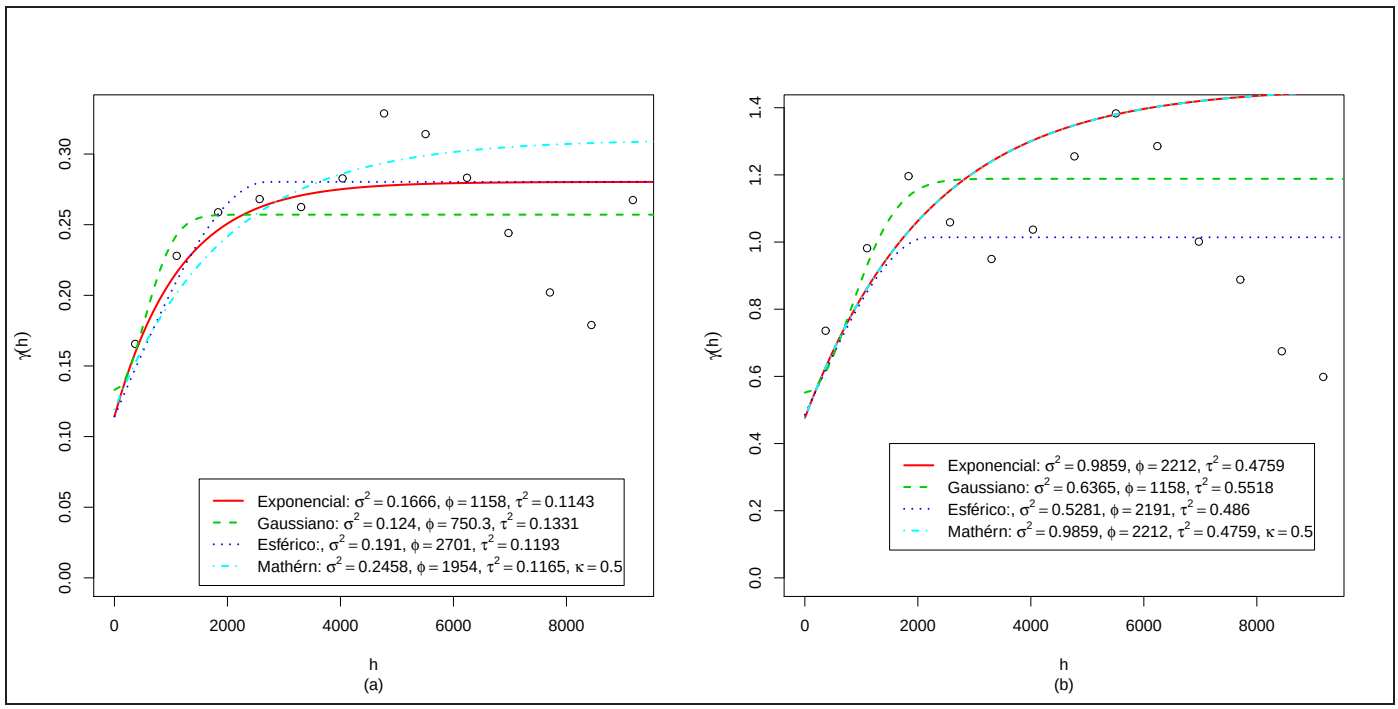


Figura 6.9: Semivariogramas Muestral para Área Basal $m^2$ (a) y Altura $m$ (b) estimado mediante REML.

| | | | Parámetros Estimados | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Datos | Modelos | $n$ | $\hat{\tau}^2$ | $\hat{\sigma}^2$ | $\hat{\phi}$ | $Log.L$ | $n.par$ | $AIC$ | $(\overline{VIF})$ | $ESS$ |
| $\sqrt{\text{Área Basal}}$ | Exponencial | 688 | 0.1143 | 0.1666 | 1158 | -355.5 | 4 | 719 | 1.8076 | 15.7602 |
| | Esférico | 688 | 0.1193 | 0.1910 | 2701 | -353.4 | 4 | 714.7 | 2.2238 | 22.4249 |
| | Gaussiano | 688 | 0.1331 | 0.1240 | 750.3 | -355.1 | 4 | 715 | 1.7917 | 51.6829 |
| | Mathérn ($k$ =0.5) | 688 | 0.1165 | 0.2458 | 1954 | -355.1 | 4 | 718.2 | 2.2238 | 4.8983 |
| Altura | Exponencial | 688 | 0.4759 | 0.9859 | 2212 | -829.1 | 4 | 1666 | 2.4507 | 5.7720 |
| | Esférico | 688 | 0.4860 | 0.5281 | 2191 | -823.7 | 4 | 1655 | 1.7144 | 34.97011 |
| | Mathérn ($k$ =0.5) | 688 | 0.4760 | 0.9859 | 2212 | -829.1 | 4 | 1666 | 2.4507 | 5.7720 |
| | Gaussiano | 688 | 0.5518 | 0.6365 | 1158 | -829.1 | 4 | 1648 | 2.0130 | 26.4955 |

Tabla 6.2: Estimación Tamaño Muestral Efectivo vía $REML$

Para la estimación del semivariograma se han propuesto varios candidatos de modelos, se elegirá sólo el que logré minimizar mejor los errores, o bien, el criterio de Akaike (AIC) menor. La estimación de los parámetros fue mediante estimación máxima verosimilitud restringida $REML$ y con la obtención de estos parámetros se calcula el tamaño muestral efectivo para cada una de las variables espaciales en estudio. La siguiente tabla presenta el resumen de los resultados:

La Tabla (6.2), presenta los parámetros estimados para cada modelo. Se puede observar que el modelo que presenta menor $AIC$ para el área basal en $m^2$ es el Esférico, obteniendo un tamaño muestral efectivo estimado de 23 unidades. De igual forma se observa que el modelo que presenta menos AIC para la Altura en $m$ es el modelo Gaussiano con un tamaño muestral efectivo estimado de 27 unidades. La interpretación de estos resultados nos indican que de las 688 unidades con correlación espacial para el área basal en $m^2$ y altura en $m$, se necesitan respectivamente 23 y 27 unidades independiente para tener la misma precisión que con 688 unidades.

Lo anterior se puede resumir con la siguiente notación (ver sección 3.5) $\widehat{ESS}_{Esf}(n, \hat{\theta}_1) = 23$ unidades y $\widehat{ESS}_{Gauss}(n, \hat{\theta}_2) = 27$ unidades con $\hat{\theta} = (\hat{\sigma^2}, \hat{\tau^2}, \hat{\phi})^t$, como las estimaciones del tamaño muestral efectivo para las variables espaciales Área Basal $m^2$ y Altura $m$ respectivamente.

### 6.5.2 Muestreo y estimación de $ESS$ para Área Basal $m^2$ y Altura $m$

En el Capítulo 4, se realizó un breve referencia de las técnicas de muestreo espacial, donde se puedo concluir que dependiendo de los objetivo de la investigación, como este dispuesta la población en estudio y la información que se desea extraer, es el método muestral espacial a utilizar. Para este caso, se opto por el muestreo espacial aleatorio simple en la selección de las unidades de muestreo georeferenciadas, por ende, los resultados que se presentan a continuación están bajo esa configuración. Para cada una de las variables en estudio Área Basal y Altura, se presentan la distribución espacial de las unidades de muestreo georefenciada en el área de estudio, es decir, de las 688 unidades muestrales, seleccionamos 23 y 27 unidades respectivamente. También, se realizo un test de hipótesis, para comprobar que esta nueva muestra es representativa de la población inicial, con intervalos de confianza para la media del 95 %. Las siguientes figuras presenta como se distribuye espacialmente la nueva muestra calculada con $ESS$:

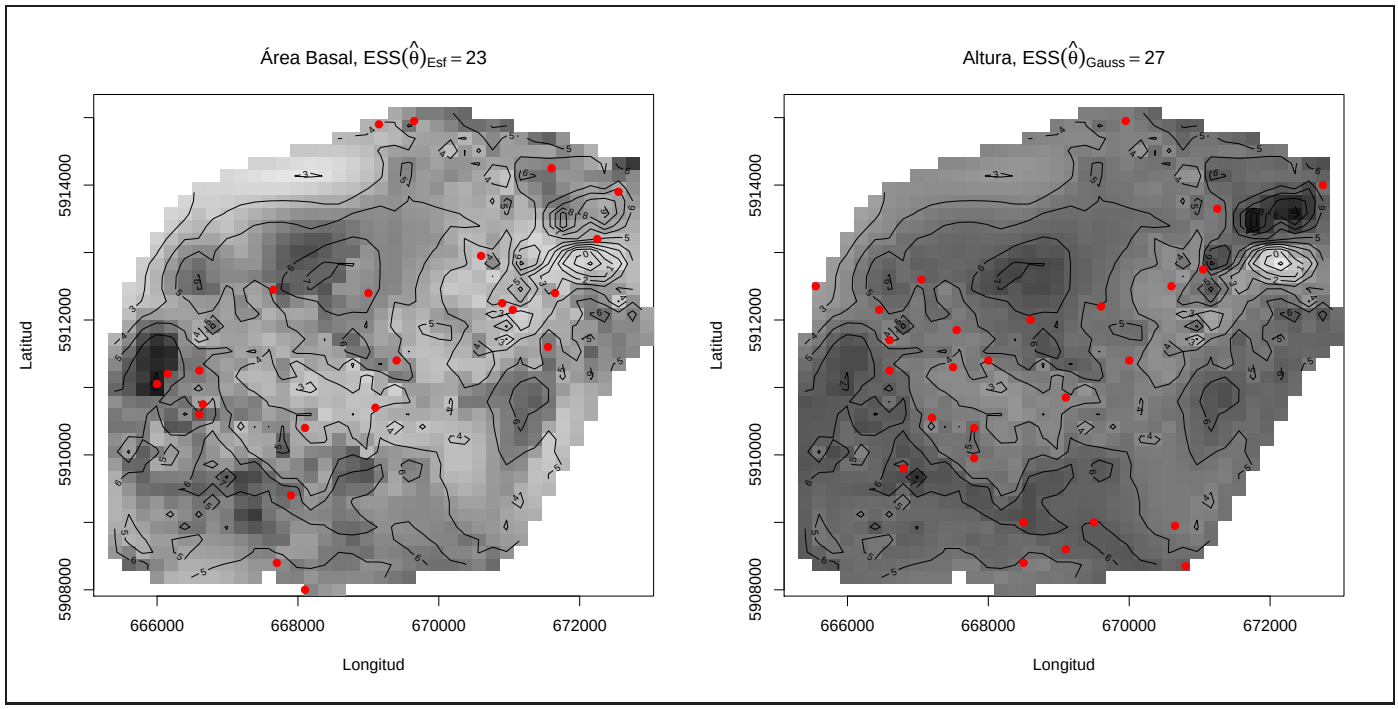


Figura 6.10: De las 688 unidades georeferenciadas, graficamos la distribución espacial de $\widehat{ESS}_{Esf} = 23$ para el Área Basal $m^2$ y $\widehat{ESS}_{Gauss} = 27$ para la Altura $m$, bajo un muestreo espacial aleatorio simple.

La Figura (6.10) muestra la distribución espacial para los tamaños muestrales efectivos de $\widehat{ESS}_{Esf}$ y $\widehat{ESS}_{Gauss}$. Para esta selección se cálculo las siguientes estadísticas descriptivas de interés:

En base a las cantidades $\widehat{ESS}_{Esf} = 23$ y $\widehat{ESS}_{Gauss} = 27$ se realizó un muestreo aleatorio simple para seleccionar la nueva muestra. A continuación se presenta una tabla con el resumen de algunas estadísticas de interés.

| | $\widehat{ESS}(n, \hat{\theta})$ | Media | de.est | Mediana | Mín | Máx | Rango | Sesgo | Curtosis | se |
|---|---|---|---|---|---|---|---|---|---|---|
| $Y_1(s)$ | 23 | 5.15 | 2.09 | 5 | 2.3 | 9.3 | 7 | 0.52 | -0.85 | 0.44 |
| $Y_2(s)$ | 27 | 4.46 | 1.37 | 4.8 | 2.5 | 6.8 | 6.8 | -1.18 | 2.05 | 0.26 |

Tabla 6.3: Estadísticas Descriptivas para $\widehat{ESS}_{Esf_1}$ y $\widehat{ESS}_{Esf_2}$

La media muestral de la variable espacial $Y_1(s)$ : área basal es 5.15 $m^2$ con una dispersión respecto de su media de 2.09 $m$. El área basal fluctúa entre 2.3 y 9.3 $m^2$ como mínimo y máximo respectivamente. De igual forma, la media muestral de la variable espacial $Y_2(s)$ : altura es 4.46 $m$ con una dispersión respecto de su media de 1.37 $m$. La altura fluctúa entre 2.5 y 6.8 $m$ como mínimo y máximo respectivamente.

Otro resultado de interés es construir un intervalo de confianza del $100(1-\alpha)\,\%$ para el área basal y la altura (asumiendo normalidad):

$$I.C[\mu_i(s)]_{100(1-\alpha)\,\%} = \left[\overline{Y}_i(s) \mp Z_{\alpha/2} \cdot \frac{\sigma_i}{\sqrt{\widehat{ESS}_i(n, \hat{\theta})}}\right] \text{ con } i = 1 \text{ área basal, } i = 2 \text{ altura.} \tag{6.1}$$

donde $\mu_i$ es la media poblacional de la variable $i-$ ésima, $\overline{Y}_i(s)$ es la media muestral de la $i-$ ésima variable, $Z_{\alpha/2}$ es el percentil $\alpha-$ ésimo de la distribución normal estándar, $\sigma_i$ es la desviación estándar poblacional de la variable $i-$ ésima, obtenida de la Tabla 6.1 y $\widehat{ESS}_{Esf_i}(n, \hat{\theta})$ los tamaños muestrales efectivos $i-$ ésimo. Considerando estos elementos los resultados son los siguientes:

| $i$ | $\overline{Y}_i(s)$ | $\widehat{ESS}_i$ | $\sigma_i$ | $Z_{0.975}$ | Lim. Inf | Lim. Sup |
|---|---|---|---|---|---|---|
| 1 | 5.15 | 23 | 2.10 | 1.96 | 4.2917 | 6.0082 |
| 2 | 4.46 | 27 | 1.04 | 1.96 | 4.0677 | 4.8522 |

Tabla 6.4: Intervalos de confianza del 95 % para $Y_1(s)$ e $Y_2(s)$.

Con un nivel de confianza del 95 % se puede afirmar que el verdadero valor de la media poblacional $\mu(s)$, para el área basal $m^2$ se encuentra entre 4.29 y 6.0 $m^2$. Para altura $m$ se puede afirmar con un nivel de confianza del 95 % el verdadero valor se encuentra entre 4.06 y 4.85 $m$.

## 6.6 Comentarios

En este capítulo se ha analizado el impacto de la correlación espacial en las variables espaciales Área Basal ($m^2$) y Altura ($m$) de los Jóvenes Pinos Radiata de forma sistemática y progresivamente, realizando una caracterización descriptiva de las variables espaciales, luego la modelación y el ajuste del semivariograma, hasta llegar a la estimación del tamaño muestral efectivo. Los resultados en la simulación y la aplicación a estas dos variables espaciales evidencian que la correlación espacial influye sobre la reducción efectiva de una muestra georeferenciada. Se contaba con una base de datos que contenían 688 registros de variables con atributos georeferenciados, utilizando la cantidad (3.43) se obtuvieron $ESS_1 = 17$ y $ESS_2 = 35$ para el Área Basal y Altura respectivamente

Capítulo 7

# Conclusión

El desarrollo de este trabajo de investigación está en función de poder responder a los objetivos generales y específicos, los capítulos tiene un orden que está orientado para poder entender el cálculo del Tamaño Muestral Efectivo en el Modelamiento de Variables Espaciales en este sentido se puede observar que desde el capítulo **Introducción** el objetivo es presentar algunos aspectos de la teoría clásica de muestreo, se presenta el caso cuando se quiere calcular del tamaño muestral para una muestra aleatoria simple. Este resultado nos permite comparar la teoría clásica con la nueva metodología propuesta en este trabajo. En la teoría clásica de muestreo, los elementos que influyen en el cálculo del tamaño muestral son la dispersión, la precision que se quiera obtener para estimar la media y el nivel de significancia. Por otro lado, un escenario distinto pero que tiene la misma idea para calcular el tamaño muestral es cuando las unidades experimentales están georeferenciadas y la correlación espacial es un elemento importante para el cálculo del tamaño muestral efectivo para variables espaciales.

Bajo esta premisa es necesario definir la estructura que subyace los datos en el espacio. La **Estadística Espacial** es la base para estimar el tamaño muestral efectivo, debido a que necesitamos una función que logre capturar la similitud de los datos distribuidos espacialmente, mediante la estructura de correlación espacial. El tamaño muestral efectivo depende de la matriz de correlación espacial y dependiendo del tipo de correlación espacial modelada es el tamaño muestral efectivo a obtener.

La literatura ha tratado de dar respuesta al **Tamaño Muestral Efectivo**, por ejemplos las propuestas vistas en la sección 3.2, 3.3 y 3.4 muestran el impacto de tener una muestra correlacionada espacialmente cuando se querer estimar la media y varianza muestral de la variable de interés. Evidentemente, la presencia de la correlación espacial en la muestra tiene que tener un tratamiento especial, una medida que se introduce es el factor de inflación de varianza esta cantidad que nos permite cuantificar el aporte que tiene cada variable respecto la variabilidad total. La definiciones para cada unas de las propuestas (3.13), (3.19) y (3.43) es la de cuantificar el número de observaciones independientes necesarias para mantener la misma precisión que la muestra original. Un resultado importante surge cuando el tamaño muestral efectivo $ESS$ tiene una forma explícita o analítica, utilizando propiedades asintóticas como el Método Delta se puede establecer la distribución del tamaño muestral efectivo $ESS$, cuando la estructura de correlación espacial es intraclase (3.57) y cuando la estructura de correlación espacial es autoregresivo de primer orden (3.67) y construir intervalos de confianzas en cada caso.

En el contexto de este trabajo de Tesis el Tamaño Muestral Efectivo en el Modelamiento de Variables Espaciales, toma un rol relevante dado que luego de determinar un tamaño muestra efectivo (óptimo), la pregunta siguiente es ¿cómo seleccionar estos datos?

El capítulo de **Muestreo Espacial** presenta de una manera muy sintetizada el estado del arte de las técnicas muestreo espacial y destacar como estos métodos son extensiones del muestreo clásico vistos por Chocran, Thompson, Kish por nombrar algunos autores, manteniendo algunas diferencias. Este Capítulo muestra varias perspectivas presentadas por varios autores con respecto al muestreo espacial y los cálculos asociados a las diferentes planes de muestreo. Los autores donde se puede profundizar esta metodología son: Griffith, D. (2008), Haining,R. (2003) y Gruijter,J; Bruc, D; Bierkens, M; Knotters, M. (2006) Cuando la estructura de correlación espacial es definida como (3.71), (3.70), (3.73) y (3.72) determinar la función de

distribución de estas cantidades (como en el caso (3.57) y (3.67)) se complica debido a que no se tiene una expresión analítica, entonces determinar la distribución del tamaño muestral efectivo cuando la correlación espacial es modelada mediante un semivariograma es aún un problema abierto. Sin embargo es posible entender el comportamiento del tamaño muestral para las cantidades (3.1), (3.2), (3.71), (3.70), (3.73) y (3.72) mediante **Simulación** para representar el comportamiento distribucional de cada cantidad $ESS$.

En este Capítulo 5 se demostró computacionalmente la capacidad de reducción que tiene el tamaño muestral efectivo $ESS(\theta, n)$. Como también, poder comprobar que $ESS \leq n$. Por definición la cantidad $ESS(\theta, n)$ es el número de observaciones independientes que depende de cierto grado de correlación espacial. Estos impactos de correlación espacial fueron representados por modelos paramétricos del semivariograma que a su vez fueron estimados mediante $REML$. Se Repitió 1000 veces este escenario con muestras georeferenciadas independientes y se estimaron los parámetros de la matriz de correlación espacial y luego calcular $ESS(\theta, n)$. En este capítulo se pudo representar la distribución muestral de $\widehat{ESS}(\widehat{\theta}, n)$ para cada uno de los modelos como son: Intraclase, Autoregresivo de primer orden, Exponencial, Esférico, Gaussiano y Mathérn.

Finalmente en el capítulo **Aplicación** se ha analizado el impacto de la correlación espacial de las variables espaciales Área Basal ($m^2$) y Altura ($m$) de Jóvenes Pinos Radiata de forma sistemática y progresivamente, realizando una caracterización descriptiva de las variables espaciales, luego la modelación y el ajuste del semivariograma, hasta llegar a la estimación del tamaño muestral efectivo. Los resultados en la simulación y la aplicación a estas dos variables espaciales evidencian que la correlación espacial influye sobre la reducción efectiva en una muestra georeferenciada. Se contaba con una base de datos que contenían 688 registros de variables con atributos georeferenciados, utilizando la cantidad (3.43) se obtuvieron $ESS_1 = 23$ y $ESS_2 = 27$ para el Área Basal y Altura respectivamente

Hay varias situaciones que pueden ser motivo nuevas investigaciones, un ejemplo es cuando se quiere determinar la función de distribución de $\widehat{ESS}(\widehat{\theta}, n) = \mathbf{1}^t\Sigma^{-1}(\widehat{\theta})\mathbf{1}$ cuando la correlación espacial es Exponencial, Esférico, Gaussiano y Mathérn, ya que no se tiene una expresión analítica y determinar la función de distribución es un problema abierto. Esto se puede ver ya que, $\Sigma(\widehat{\theta})$ es una matriz de covarianza espacial estocástica de dimensión $n \times n$ y necesitamos determinar $\Sigma^{-1}(\widehat{\theta})$ lo aún más se complica. Una opción sería inspeccionar los elementos $i, j$-ésimo de la matriz de covarianza $\Sigma^{-1}(\theta)_{ij}$ y luego determinar la distribución $\mathbf{1}^t\Sigma^{-1}(\widehat{\theta})\mathbf{1}$. O bien, calcular $\mathbb{E}[\widehat{ESS}] = \mathbf{1}^t\mathbb{E}\left[\Sigma^{-1}(\widehat{\theta})\right]\mathbf{1}$ es un problema abierto.

Capítulo 8

# Bibliografía


[**1** ] Anselin, L; J. Rey, S. (2010). Perspectives on Spatial Data Analysis. Springer, New York.

[**2** ] Banerjee, S; Carin, B, P; Gelfand, A, E. (2004). Hierarchical Modelling and Analysis for Spatial Data. Chapman & Hall CRE

[**3** ] Bivand, R;Pebesma, E; Gómez-Rubio, V. (2008). Applied Spatial Data Analysis with R. Springer, New York.

[**4** ] Brockwell, P; Davis, R. (1991). Times Series: Theory and Methods, 2da Edition. Springer-Verlag, New York.

[**5** ] Box, G, E; Jenkins, G. (1976). Times Series Analysis: Forecasting and Control. Revised Edition. Holden Pay, Inc.

[**6** ] Cochran, W, G. (1977). Sampling Techniques.Third Edition. John Wiley & Sons. Inc.

[**7** ] de Gruijter,J; Bruc, D; Bierkens, M; Knotters, M. (2006). Sampling for Natural Resource Monitoring. Springer-Verlag Berlin Heidelberg.

[**8** ] Cressie, N. (1993). Statistics for spatial data.New York : John Wiley, 1993 pág,13-15.

[**9** ] Díaz, Martín. (2002). Geoestadística Aplicada. Instituto de Geofísica, UNAM. Cuba.

[**10** ] Gaetan, C; Guyon, X. (2010). Spatial Statistics and Modeling. Springer, New York.

[**11** ] Griffith, D. (2005). Effective geographic sample size in the presence of spatial autocorrelation. Annals of the Association of American Geographers 95, 740-760.

[**12** ] Griffith, D. (2008). Geographic sampling of urban soils for contaminant mapping: how many samples and from where, Environ Geochem Health, pág 495-507.

[**13** ] Groves, R. (1989, 2004). Survey Errors and Survey Cost. John Wiley & Sons, Inc. Hoboken. New Jersey.

[**14** ] Haining, R. (1990). Spatial Data Analysis in the social and environmental sciences. Cambridge University Press.

[**15** ] Haining, R. (2004).Spatial Data Analysis: Theory and Practice. Cambridge University Press.

[**16** ] Hansen, M; Hurwitz, W; Madow, W.(1953). Sample Survey Method and Theory. Volume II-Theory. John Wiley & Sons, Inc. New York.

[**17** ] Henao, R. Introducción a la Geoestadística: Teoría y Aplicación. Universidad Nacional de Colombia. Facultad de Ciencias, Departamento de Estadística. Bogotá.

[**18** ] Kish, L. (1965). Survey Sampling. John Wiley & Sons, Inc. New York.

[**19** ] Levy, P; Lemeshow, S. (1999). Sampling of Populations: Methods and Applications. Third Edition. Wiley & Sons, Inc. Canada.

[**20** ] Mukhophadhay, N. (2000). Probability and Statistical Inference. Statistics: Textbooks and Monograph.

[**21** ] M*ü*ller, W. (1998, 2000, 2007). Collecting Spatial Data, Third Edition, Springer, New York

[**22** ] Neter, J; Wasserman, W; Kutner, M. (1983). Applied Linear Regression Models. Richard R Irwin, Inc. pág, 390-393.

[**23** ] Pateiro-López, B; Rodríguez-Casal, A. (2008). Generalizing the Convex Hull of a Sample: The R Package alphahull. Journal of Statistical Software. pág, 1-28.

[**24** ] Rao, R,C. (2002). Linear Statistical Inference and Its applications. Second Edition. Wiley Series in Probability and Statistics.

[**25** ] Rao, P. (2000). Sampling Methodologies with Applications. Chapman & Hall/CRC.US.

[**26** ] Reimann, C; Filzmoser,P; Garrett, R; Dutter. (2008). Statistical Data Analysis Explained: Applied Environmental Statistics with R. John Wiley & Sons ltd.

[**27** ] Rencher, A, C. (2002). Methods of Multivariate Analysis. Second Edition. Brigham Young University. John Wiley & Sons, Inc.

[**28** ] Ripley, B, D. (1981). Spatial Statistics. University of London: John Wiley & Sons, Inc. New Jersey.

[**29** ] Schabenberger, O; Gotway, C, A. (2005). Statistical Methods for Spatial Data Analysis. Chapman & Hall CRE

[**30** ] Stein, M. (1999). Interpolation of Spatial Data: Some Theory for Kriging. Springer Verlag. pág. 17.

[**31** ] Vallejos, R;Torres, M. (2008).Herramientas Estadísticas para la Predicción de Variables Georefenciadas. Memoria para Optar al Título de Estadístico. Universidad Catolica de Valparaíso. Chile. pág,9-40.

Apéndice A

# Anexo N°1: Rutinas R - Simulación

```
#---------------Cargar liberias-----------------------------------------------#
library(MASS)
library(geoR)
library(gstat)
library(akima)
library(nlme)
library(nFactors)
library(ruf)
library(scatterplot3d)
library(sp)
#-----------------------------------------------------------------------------#
#--------------Esquema de muestreo aleatorio Simple---------------------------#
#-----------------------------------------------------------------------------#
#-------------------------SIMULACION 1----------------------------------------#
#----------------------Modelo Exponencial-------------------------------------#
Corr_Exp<-function(n)
{
n<-500
sigma<-1
phi<-0.25   #rango
ngt<-0.5    #nugget
sim1 <- grf(n, grid = "irreg",mean = 100, nugget = ngt,cov.pars = c(sigma,phi)
,cov.model = "exponential")
#windows()
#points(sim1)
#sim1$data
#hist(sim1$coords[1:100])
#windows()
#hist(sim1$data)
#windows()
#plot(sim1)
#windows()
#plot(hclust(dist(sim1$coords)))
#scatterplot3d(s1, s2,datos1, type = "h", angle = 55)
geo.data.exp<- as.geodata(sim1,coords.col=1:2, data.col=3)
rml <-likfit(geo.data.exp,geo.data.exp$coords,data=geo.data.exp$data,
ini=c(0.5,0.5),lik.method="REML",cov.model = "exponential")
#rml
#par1<-rml$tausq
#par2<-rml$sigmasq
```

```
par3<-rml$phi
par4<-rml$nugget
s1<-as.matrix(sim1$coords[,1])
s2<-as.matrix(sim1$coords[,2])
spatDat <- data.frame(s1,s2)
cs1Exp <- corExp(c(par3,par4), form = ~ s1 + s2, nugget = TRUE)
cs1Exp <- Initialize(cs1Exp, spatDat)
MExp1<-as.matrix(corMatrix(cs1Exp))
uno<-matrix(1,n,1)
ESS<-t(uno)%*%solve(MExp1)%*%(uno)
ESS
}
set.seed(101)
n<-500
nsim<-578
AC<-matrix(0,nsim,1)
for(i in 1:nsim)
{
AC[i]<-Corr_Exp(n)
}
#-------------------------------------------------------------------------------#
#----------------------Guarda los Resultados en una hoja de excell-------------#
write.table(AC, file = "D:/DOC JUAN CARLOS/Año_2010/Tesis Magister 2010/Simulacion
 Mod Exponencial/ess_exp_sim_final.txt", append =FALSE, col.names =TRUE,sep="\t")
#-----------------------Abre hoja txt-------------------------------------#
D1<-read.table('D:/DOC JUAN CARLOS/Año_2010/Tesis Magister 2010/Simulacion
 Mod Exponencial/ess_exp_sim.txt')
D1<-as.matrix(D1)
#-------------------------------------------------------------------------------#
AC<-D1[,2]
nsim<-length(AC)
#------------------ANALISIS DESCRIPTIVO-----------------------------------#
describe(AC)
windows()
hist(AC,xlab="(a)",ylim=c(0,0.055),ylab="Densidad",main=expression(paste('Simulación
ESS con ', sigma^2,"=1, ", phi ," =0.25, ",tau^2,"=0.5 y ",n,'=500')),probability=TRUE
,col="light blue")
lines(density(AC),col=3)
rug(AC)
windows()
boxplot(AC,xlab="(b)",ylab="",main="",col="light grey")
windows()
qqnorm(AC,xlab="(c)",ylab="",main="")
qqline(AC, col = 2)
ajuste.lognormal<-fitdistr(log(AC),"Normal")
mean.log<-ajuste.lognormal$estimate[1]
sd.log<-ajuste.lognormal$estimate[2]
windows()
x<-seq(min(AC),max(AC),by=(max(AC)-min(AC))/nsim)
plot(ecdf(AC), do.points=FALSE, verticals=TRUE,xlab="(d)",ylab="",main="",col="1")
lines(x, pnorm(x, mean=mean(AC), sd=sqrt(var(AC))), lty=3,col=2)
lines(x, plnorm(x, meanlog=mean.log, sdlog=sd.log), lty=6,col=3)
```

```
legend(30,0.35, c("Dist. Muestral","Dist. Normal","Dist. Log-Normal"), lty=3:4,col=1:3)
#-------------------------SIMULACION 2----------------------------------------#
#----------------------Modelo Gaussiano-------------------------------------#
Corr_Gaus<-function(n)
{
#n<-500
sigma<-0.1
phi<-0.1   #rango
ngt<-0.1  #nugget  tau
sim2 <- grf(n, grid = "irreg",mean = 100, nugget = ngt,cov.pars = c(sigma,phi),
cov.model = "gaussian")
#windows()
#points(sim2)
#sim2$data
#hist(sim2$coords[1:100])
#windows()
#hist(sim2$data)
#windows()
#plot(sim2)
#sim2$coords
geo.data.gaus<- as.geodata(sim2,coords.col=1:2, data.col=3)
rml1 <-likfit(geo.data.gaus,geo.data.gaus$coords,data=geo.data.gaus$data,
ini.cov.pars=c(0.01,0.01)
,lik.method="ML",cov.model = "gaussian")
#par1<-rml1$tausq
#par2<-rml1$sigmasq
par3<-rml1$phi
par4<-rml1$nugget
s1<-as.matrix(sim2$coords[,1])
s2<-as.matrix(sim2$coords[,2])
spatDat <- data.frame(s1,s2)
Cgaus <- corGaus(c(par3,par4), form = ~ s1+s2, nugget = TRUE)
Cgaus <- Initialize(Cgaus, spatDat)
MGaus<-as.matrix(corMatrix(Cgaus))
#det(MGaus)
uno<-matrix(1,n,1)
ESS<-t(uno)%*%solve(MGaus)%*%(uno)
ESS
}
set.seed(5001)
n<-500
nsim<-100
AC1<-matrix(0,nsim,1)
for(i in 1:nsim)
{
AC1[i]<-Corr_Gaus(n)
}
#-------------------------------------------------------------------------#
#----------------------Guarda los Resultados en una hoja de excell-------------#
write.table(acm_rho, file = "D:/ess_gaus_sim1.txt", append =FALSE,
col.names = TRUE,sep="\t")
#----------------------Abre hoja txt-----------------------------------#
```

```
#D1<-read.table('D:/DOC JUAN CARLOS/Año_2010/Tesis Magister 2010/Simulacion
Mod Exponencial/ess_exp_sim.txt')
#D1<-as.matrix(D1)
#-------------------------------------------------------------------------------#
#AC<-D1[,1]
#------------------ANALISIS DESCRIPTIVO-----------------------------------------#
describe(AC1)
windows()
hist(AC1[1:93],xlab="",ylab="Densidad",ylim=c(0,10),main="",probability=TRUE,
col="light blue")
lines(density(AC1[1:93]),col=3)
rug(AC1[1:93])
windows()
boxplot(AC1[1:93],xlab="",ylab="",main="",col="light grey")
windows()
qqnorm(AC1[1:93],xlab="",ylab="",main="")
qqline(AC1[1:93], col = 2)
windows()
x<-seq(min(AC1[1:93]),max(AC1[1:93]),by=(max(AC1[1:93])-min(AC1[1:93]))
/length(AC1[1:93]))
plot(ecdf(AC1[1:93]), do.points=FALSE, verticals=TRUE,xlab="",ylab="",
main="",col="black")
lines(x, pnorm(x, mean=mean(AC1[1:93]), sd=sqrt(var(AC1[1:93]))), lty=3,col="red")
#-------------------------SIMULACION 3---------------------------------------#
#----------------------Modelo Esferico-----------------------------------------#
Corr_Esf<-function(n)
{
#n<-500
sigma<-0.8
phi<-0.2   #rango
ngt<-0.7    #nugget
sim3 <- grf(n, grid = "irreg",mean = 100, nugget = ngt,cov.pars = c(sigma,phi),
cov.model ="spherical")
#points(sim3)
#sim3$data
#hist(sim3$coords[1:100])
#hist(sim3$data)
#plot(sim3)
#sim3$coords
geo.data.sph<- as.geodata(sim3,coords.col=1:2, data.col=3)
rml2 <-likfit(geo.data.sph,geo.data.sph$coords,data=geo.data.sph$data,
ini.cov.pars=c(0.8,0.1),
lik.method="REML",cov.model ="spherical")
#rml2
#par1<-rml2$tausq
#par2<-rml2$sigmasq
par3<-rml2$phi
par4<-rml2$nugget
s1<-as.matrix(sim3$coords[,1])
s2<-as.matrix(sim3$coords[,2])
spatDat <- data.frame(s1,s2)
Csph <- corSpher(c(par3,par4), form = ~ s1 + s2, nugget = TRUE)
```

```
Csph <- Initialize(Csph, spatDat)
MSph<-as.matrix(corMatrix(Csph))
uno<-matrix(1,n,1)
ESS<-t(uno)%*%solve(MSph)%*%(uno)
ESS
}
set.seed(1001)
n<-500
nsim<-451
AC2<-matrix(0,nsim,1)
for(i in 1:nsim)
{
AC2[i]<-Corr_Esf(n)
}
#-------------------------------------------------------------------------------#
#----------------------Guarda los Resultados en una hoja de excell-------------#
write.table(AC2, file = "D:/DOC JUAN CARLOS/Año_2010/Tesis Magister 2010/Simulacion
 Mod Esferico/ess_esf_sim_final.txt", append =FALSE, col.names = TRUE,sep="\t")
#-----------------------Abre hoja txt-------------------------------------#
D2<-read.table('D:/DOC JUAN CARLOS/Año_2010/Tesis Magister 2010/Simulacion
Mod Esferico/ess_esf_sim1.txt')
D2<-as.matrix(D2)
#-------------------------------------------------------------------------------#
AC2<-D2[,2]
#------------------ANALISIS DESCRIPTIVO-----------------------------------------#
describe(AC2)
windows()
hist(AC2,xlab="(a)",ylab="Densidad",ylim=c(0,0.015),main=expression
(paste('Simulación ESS
con ', sigma^2,"=0.8, ", phi ," =0.2, ",tau^2,"=0.7 y ",n,'=500')),
probability=TRUE,col="light blue")
lines(density(AC2),col=3)
rug(AC2)
windows()
boxplot(AC2,xlab="(b)",ylab="",main="",col="light grey")
windows()
qqnorm(AC2,xlab="(c)",ylab="",main="")
qqline(AC2, col = 2)
windows()
x<-seq(min(AC2),max(AC2),by=(max(AC2)-min(AC2))/nsim)
plot(ecdf(AC2), do.points=FALSE, verticals=TRUE,xlab="(d)",ylab="",
main="",col="black")
lines(x, pnorm(x, mean=mean(AC2), sd=sqrt(var(AC2))), lty=3,col="red")
legend(150,0.35, c("Dist. Muestral","Dist. Normal"), lty=3:4,col=c("black","red")
#-------------------------SIMULACION 4-----------------------------------------#
#----------------------Modelo Mathern-----------------------------------------#
Corr_Math<-function(n)
{
#n<-100
sigma<-1
phi<-0.2 #rango
ngt<-0.7    #nugget
```

```
sim3 <- grf(n, grid = "irreg",mean = 100,kappa = 0.2, nugget = ngt,
cov.pars = c(sigma,phi),cov.model ="matern")
#points(sim3)
#sim3$data
#hist(sim3$coords[1:100])
#hist(sim3$data)
#plot(sim3)
#sim3$coords
geo.data.mt<- as.geodata(sim3,coords.col=1:2, data.col=3)
rml2 <-likfit(geo.data.mt,geo.data.mt$coords,data=geo.data.mt$data,
fix.kappa =
TRUE,ini.cov.pars=c(1,0.1),lik.method="REML",cov.model ="matern")
#rml2
#par1<-rml2$tausq
#par2<-rml2$sigmasq
par3<-rml2$phi
par4<-rml2$nugget
par5<-rml2$kappa
MT<-vmatcov(geo.data.mt,c(par3,0.2))
#det(MT)
uno<-matrix(1,n,1)
ESS<-t(uno)%\*%\solve(MT)%*%(uno)
ESS
}
set.seed(1001)
n<-500
nsim<-500
AC3<-matrix(0,nsim,1)
for(i in 1:nsim)
{
AC3[i]<-Corr_Math(n)
}
#-------------------------------------------------------------------------------#
#----------------------Guarda los Resultados en una hoja de excell-------------#
write.table(AC3, file = "D:/ess_math_sim2.txt", append =FALSE, col.names =
TRUE,sep="\t")
#-----------------------Abre hoja txt-------------------------------------#
D3<-read.table("D:/DOC JUAN CARLOS/Año_2010/Tesis Magister 2010/Simulacion
 Mod Matheron/ess_math.txt")
D3<-as.matrix(D3)
#-------------------------------------------------------------------------------#
AC3<-D3
#-----------------ANALISIS DESCRIPTIVO-------------------------------------#
describe(AC3)
windows()
hist(AC3,xlab="(a)",ylab="Densidad",ylim=c(0,0.017),main=expression
(paste('Simulación ESS con ', sigma^2,"=1, ", phi ," =0.2, ",tau^2,"=0.7
 y ",n,'=500')),
probability=TRUE,col="light blue")
lines(density(AC3),col=3)
rug(AC3)
windows()
```

```
boxplot(AC3,xlab="(b)",ylab="",main="",col="light grey")
windows()
qqnorm(AC3,xlab="(c)",ylab="",main="")
qqline(AC3, col = 2)
ajuste.lognormal<-fitdistr(log(AC3),"Normal")
mean.log<-ajuste.lognormal$estimate[1]
sd.log<-ajuste.lognormal$estimate[2]
sh<-fitdistr(AC3,"gamma")$estimate[1]
rte<-fitdistr(AC3,"gamma")$estimate[2]
#-------------------------------------------------------------------------------#
#------------------Fitting distribution with R-------------------------------#
#-------Librerias--------------#
library(MASS)
library(vcd)
library(tseries)
library(nortest)
library(fBasics)
library(SuppDists)
nsim<-length(AC3)
#-------------------------------------------------------------------------------#
q1<-nigFit(D3, alpha =3.019021, beta = 2.994079, delta = 7.113855,
 mu =-7.446310, doplot = TRUE)
#-------------------------------------------------------------------------------#
#Model: Normal Inverse Gaussian Distribution
#Estimated Parameter(s):
#    alpha      beta     delta        mu
# 3.855976  3.830922  6.338983 -7.774242
#-------------------------------------------------------------------------------#
x<-seq(min(AC3),max(AC3),by=(max(AC3)-min(AC3))/nsim)
plot(ecdf(AC3), do.points=FALSE, verticals=TRUE,xlab="(d)",ylab="",
main="",col="black")
#lines(x, pnorm(x, mean=mean(AC3), sd=sqrt(var(AC3))), lty=3,col=1)
#lines(x, plnorm(x, meanlog=mean.log, sdlog=sd.log), lty=4,col=2)
lines(x, pinvGauss(x, nu=7.113855, lambda=-7.446310, lower.tail=TRUE,
 log.p=FALSE),col=7)
#lines(x, pgamma(x,shape=sh, scale=1/rte), lty=6,col=4)
legend(100,0.35, c("Dist. Normal","Dist. Log-Normal","Dist. Exponencial",
"Dist. Gamma"), lty=3:6,col=1:4)
#-------------------------------------------------------------------------------#
a<-fitdistr(AC3,"gamma")
BIC(logLik(a))
AIC(a)
#--------------------------#
#     shape         rate
#  2.283382421   0.048014445
# (0.095487825) (0.002243787)
#-----------------------------#
ks.test(AC3,"pgamma",2.283382421,0.048014445,exact=F)
#-------------------------------------------------------------------------------#
b<-fitdistr(AC3,"lognormal")
BIC(logLik(b))
AIC(b)
```

```
#-------------------------#
#   meanlog       sdlog
#  3.62727461   0.72089467
# (0.02279669) (0.01611969)
#-------------------------#
ks.test(AC3,"plnorm",3.62727461,0.72089467)
#-------------------------------------------------------------------------------#
c<-fitdistr(AC3,"normal")
BIC(logLik(c))
AIC(c)
#-------------------------------------------------------------------------------#
d<-fitdistr(AC3,"exponential")
BIC(logLik(d))
AIC(d)
#-------------------------------------------------------------------------------#
e<-fitdistr(AC3,"weibull",start=list(shape=1.53374243,scale=53.13771967))
BIC(logLik(e))
AIC(e)
ks.test(AC3,"pweibull", shape=1.53374243,scale=53.13771967)
#-------------------------------------------------------------------------------#
g<-fitdistr(AC3,"t",df=999)
BIC(logLik(g))
AIC(g)
#-------------------------#
#     m            s
#  47.5043092   33.3629824
# ( 1.0560830) ( 0.7488322)
#-------------------------#
ks.test(AC3,"pt",47.5043092,scale=33.3629824)
#-------------------------------------------------------------------------------#
windows()
x<-seq(min(AC3),max(AC3),by=(max(AC3)-min(AC3))/nsim)
plot(ecdf(AC3), do.points=FALSE, verticals=TRUE,xlab="(d)",ylab="",
main="",col="black")
lines(x, pnorm(x, mean=mean(AC3), sd=sqrt(var(AC3))), lty=3,col=1)
lines(x, plnorm(x, meanlog=mean.log, sdlog=sd.log), lty=4,col=2)
lines(x, pexp(x, rate=length(AC3)/sum(AC3)), lty=5,col=3)
lines(x, pgamma(x,shape=sh, scale=1/rte), lty=6,col=4)
legend(100,0.35, c("Dist. Normal","Dist. Log-Normal","Dist. Exponencial",
"Dist. Gamma"), lty=3:6,col=1:4)
#-------------------------------------------------------------------------------#
x <- AC3
n <- length(x)
nllhood = function(lambda) {
 -1 * (n * log(lambda) - lambda * sum(x))
}
fit <- nlminb(length(AC3)/sum(AC3), nllhood)
fit
library(stats4) ## loading package stats4
ll<-function(lambda,alfa) {n<-200
x<-x.gam
-n*alfa*log(lambda)+n*log(gamma(alfa))-(alfa-
```

```
1)*sum(log(x))+lambda*sum(x)} ## -log-likelihood function
est<-mle(minuslog=ll, start=list(lambda=2,alfa=1))
summary(est)
fitdistr(AC3,"gamma")
#-----------------------------------------------------------------------#
```

Apéndice B

# Anexo N°2: Rutinas R - Modelación Datos

```
#----------------Cargar liberias-----------------------------------------------#
library(MASS)
library(geoR)
library(gstat)
library(akima)
library(nlme)
library(nFactors)
library(ruf)
library(scatterplot3d)
library(sp)
library(ape)
#------------------------------------------------------------------------------#
#----------------------Abre hoja txt-------------------------------------------#
datos1<-read.table('D:/DOC JUAN CARLOS/Año_2010/Tesis Magister 2010/Datos para
tesis/Datos_arboles.txt',col.names=c('x','y','abasal'))

coord<-read.table('D:/DOC JUAN CARLOS/Año_2010/Tesis Magister 2010/Datos para
tesis/coordenadas.txt',col.names=c('x','y'))

bsl<-read.table('D:/DOC JUAN CARLOS/Año_2010/Tesis Magister 2010/Datos para
tesis/abasal.txt',col.names=c('abasal'))
#------------------------------------------------------------------------------#
md<-lm(abasal~x+y,data=datos1)
summary(md)
md1<-lm(abasal~1,data=datos1)
summary(md1)
describe(bsl)
#------------------------------------------------------------------------------#
dist.datos<-1/(dist(coord))
diag(dist.datos) <- 0
ozone.dists.bin <- (dist.datos > 0 & dist.datos <= .75)
Moran.I(as.matrix(bsl[,1]),dist.datos)
#------------------------------------------------------------------------------#
library(MASS)
library(vcd)
library(tseries)
library(nortest)
library(fBasics)
```

```
library(SuppDists)
#-------------------------------------------------------------------------------#
shapiro.test(bsl[,1])
jarque.bera.test(bsl[,1])
sf.test(bsl[,1])
ad.test(bsl[,1])
cvm.test(bsl[,1])
lillie.test(bsl[,1])
#ks.test(bsl[,1],"pnormal", mean(bsl[,1]),sqrt(var(bsl[,1])))
windows()
qqnorm(bsl[,1],xlab="",ylab="",main="")
qqline(bsl[,1], col = 2)
#-------------------------------------------------------------------------------#
coordinates(datos1) <- ~ x + y;
str(datos1);
summary(datos1);
#-------------------------------------------------------------------------------#
windows()
spplot(datos1,main="Área Basal")
#-------------------------------------------------------------------------------#
xy1<-spsample(datos1, 600, "random");
windows()
plot(xy1);
#-------------------------------------------------------------------------------#
xy2<-spsample(datos1, 100, "regular");
windows()
plot(xy2);
#-------------------------------------------------------------------------------#
xy3<-spsample(datos1, 100, "stratified");
windows()
plot(xy3);
#-------------------------------------------------------------------------------#
xy4<-spsample(datos1, 100, "nonaligned");
windows()
plot(xy4);
#-------------------------------------------------------------------------------#
xy5<-spsample(datos1, 100, "hexagonal");
windows()
plot(xy5);
#-------------------------------------------------------------------------------#
vm <- voronoi.mosaic(coord)
plot(vm,xlab="",ylab="",main="",sub="")
par(new=T)
plot(coord,col='blue',xlab="",ylab="",main="",sub="")
#-------------------------------------------------------------------------------#
##Define the bounding box and make the call
n<-length(coord[,1])
xmin<-min(coord[,1])
xmax<-max(coord[,1])
ymin<-min(coord[,2])
ymax<-max(coord[,2])
b.box <- c(xmin,ymin,xmax,ymax)
```

```
out <- hexGrid(20, b.box)
##Plot using lapply
windows()
plot(out$hex.centroids, pch=19, cex=0.5,
ylab="Northing", xlab="Easting")
lapply(out$hex.polygons, polygon, col="blue")
#-------------------------------------------------------------------------------#
##Now color hexagons based on value
my.col.ramp <- function(z){
zlim <- range(z)
zlen <- zlim[2]-zlim[1]+1
colorlut <- heat.colors(as.integer(zlen))
col <- colorlut[z-zlim[1]+1]
col
}
#-------------------------------------------------------------------------------#
windows()
n <- nrow(out$hex.centroids)
col <- my.col.ramp(t(bsl))
plot(out$hex.centroids, typ="n", ylab="", xlab="Easting")
for(i in 1:n){
polygon(out$hex.polygons[[i]], col=col[i], border=col[i])
}
## End(Not run)
#-------------------------------------------------------------------------------#
windows()
Z<-cbind(coord[,1],coord[,2],datos1[,3])
scatterplot3d(Z,pch=20,angle=50,xlab="",ylab="",zlab="",main="Área Basal",type="h",
highlight.3d=TRUE)
#-------------------------------------------------------------------------------#
windows()
hist(datos1[,3], probability=TRUE,xlab="",ylab="",main="Área Basal",sub="",col=3)
lines(density(datos1[,3]),col=4)
rug(datos1[,3])
#-------------------------------------------------------------------------------#
windows()
boxplot(datos1[,3],xlab="",ylab="",main="",sub="",col="Lightblue")
#-------------------------------------------------------------------------------#
windows()
qqnorm(datos1[,3],xlab="",ylab="",main="")
qqline(datos1[,3], col = 2)
#-------------------------------------------------------------------------------#
hist(D1[,4], probability=TRUE,xlab="",ylab="",main="",sub="",col=5)
lines(density(D1[,4]),col=4)
rug(D1[,4])
#-------------------------------------------------------------------------------#
geo.data<- as.geodata(datos1,coords.col=1:2, data.col=3)
windows()
plot(geo.data)
#-------------------------------------------------------------------------------#
windows()
plot(coord[,1],coord[,2],xlab="",ylab="",main="",type="p",lty=8,pch=20,col=3)
```

```
#-------------------------------------------------------------------------------#
windows()
plot(coord[,1],datos1[,3],xlab="Longitud",ylab="Área Basal",main="Longitud vs
Área Basal",type="p",lty=8,pch=20,col=3)
#-------------------------------------------------------------------------------#
windows()
plot(datos1[,3],coord[,2],xlab="Área Basal",ylab="Latitud",main="Latitud vs Área
 Basal",type="p",lty=8,pch=20,col=3)
#-------------------------------------------------------------------------------#
Abasal<-interp.new(datos1[,1],datos1[,2],datos1[,3])
#----------Interpolación de la altura Basal---------------------------------#
image(Abasal,main="Área Basal",xlab="Longitud",ylab="Latitud")
contour(Abasal,add=TRUE,nlev=15,filled=TRUE)
#--------------Distribución Espacial de la altura Basal----------------------#
#ángulos de visión.
windows()
persp(Abasal,col = "8",scale = TRUE, shade = 0.5, border = NA,box = TRUE,main="(a)",
theta = -150, phi = 30,ylab="Latitud",xlab="Longitud",zlab="Área Basal")
#-------------------------------------------------------------------------------#
windows()
persp(Abasal,col = "8",scale = TRUE, shade = 0.5, border = NA,box = TRUE,main="(b)",
theta = -10, phi = 30,ylab="Latitud",xlab="Longitud",zlab="Área Basal")
#-------------------------------------------------------------------------------#
windows()
persp(Abasal,col = "8",scale = TRUE, shade = 0.5, border = NA,box = TRUE,main="(c)",
theta = 30, phi = 30,ylab="Latitud",xlab="Longitud",zlab="Área Basal")
#-------------------------------------------------------------------------------#
#--------------Lectura de datos originales como geodata----------------------#
#-------------------------------------------------------------------------------#
DAbasal<-as.geodata(datos1,coords.col=1:2,data.col=3)
DAbasal
#-------------------------------------------------------------------------------#
basal.varg<-variog(DAbasal,estimator.type="classical")
basal.varg
plot(basal.varg,main="",xlab="h",pch=20,col=2,ylab=expression(gamma(h)))
#-------------------------------------------------------------------------------#
#------------Ajuste del Modelo de Semivariograma--------------------------------#
#----------model parameters estimated by WLS (weighted least squares)----------#
#-------------------------------------------------------------------------------#
basal.var.ajuste<-variofit(basal.varg,cov.model="exp", fix.nugget =FALSE,
ini.cov.pars=c(7,0.25),nugget=2)
#-------------------------------------------------------------------------------#
basal.var.ajuste1<-variofit(basal.varg,cov.model="spherical", fix.nugget =FALSE,
ini.cov.pars=c(7,0.25),nugget=2)
basal.var.ajuste1
#-------------------------------------------------------------------------------#
#  vg3 <- variog(geotrees3,trend=1st,max.dist= 9542.536/2)
#-------------------------------------------------------------------------------#
#tausq sigmasq     phi
#0.9762  3.4166  0.2500
#-------------------------------------------------------------------------------#
basal.varg<-variog(DAbasal,estimator.type="classical")
```

```
plot(basal.varg,main="Ajuste de semivariograma Exponencial",
ylab=expression(gamma(h)),xlab="h")
#-------------------------------------------------------------------------------#
sc1<-variofit(basal.varg,ini=c(2.6,0.25),weights="equal",cov.model="exponential")
sc2<-variofit(basal.varg,ini=c(2.6,0.25),weights="equal",cov.model="gaussian")
sc3<-variofit(basal.varg,ini=c(2.6,0.25),weights="equal",cov.model="spherical")
sc4<-variofit(basal.varg,ini=c(2.6,0.25),weights="equal",cov.model="matern")
#-------------------------------------------------------------------------------#
ols<-variofit(basal.varg,ini=c(3.4166,0.25),weights="equal",cov.model="matern"
,kappa=1)

wls<-variofit(basal.varg,ini=c(3.4166,0.25),weights="cressie",kappa=1)
#-------------------------------------------------------------------------------#
ml <- likfit(DAbasal,ini=c(3.4166,0.25),trend="1st",lik.method = "ML",
cov.model = "exponential")
rml <- likfit(DAbasal,ini=c(3.4166,0.25),trend="1st",lik.method = "RML",
cov.model = "exponential")
#-------------------------------------------------------------------------------#
plot(basal.varg)
lines(ols)
lines(wls,lwd=2)
lines(ml,lty=2)
lines(rml,lty=2,lwd=2)
legend(5000,3,c("OLS","WLS","ML","REML"),lty=c(1,1,2,2),lwd=c(1,2,1,2))
#-------------------------------------------------------------------------------#
#------------Ajuste del Modelo de Semivariograma-------------------------------#
#----------model parameters estimated by REML, ML-----------------------------#
#-------------------------------------------------------------------------------#
geo.data<- as.geodata(datos1,coords.col=1:2, data.col=3)
rml1<-likfit(geo.data,geo.data$coords,data=geo.data$data,ini=c(3.4166,0.2500),
lik.method="REML",cov.model = "exponential")

rml2<-likfit(geo.data,geo.data$coords,data=geo.data$data,ini=c(3.4166,0.2500),
lik.method="REML",cov.model = "spherical")
#-------------------------------------------------------------------------------#
#---------------------------Abre hoja txt---------------------------------------#
#-------------------------------------------------------------------------------#
datos2<-read.table('D:/DOC JUAN CARLOS/Año_2010/Tesis Magister 2010/Datos para
tesis/Datos_arboles1.txt',col.names=c('x','y','altura'))

coord<-read.table('D:/DOC JUAN CARLOS/Año_2010/Tesis Magister 2010/Datos para
tesis/coordenadas.txt',col.names=c('x','y'))

altura<-read.table('D:/DOC JUAN CARLOS/Año_2010/Tesis Magister 2010/Datos
para tesis/altura.txt',,col.names=c('altura'))
#-------------------------------------------------------------------------------#
describe(altura[,1])
#-------------------------------------------------------------------------------#
coordinates(datos2) <- ~ x + y;
str(datos2);
summary(datos2);
```

```
windows()
spplot(datos2, main="Altura")
#-------------------------------------------------------------------------#
windows()
Z<-cbind(coord[,1],coord[,2],altura[,1])

scatterplot3d(Z,pch=20,angle=50,xlab="",ylab="",zlab="",main="Altura",type="h",
highlight.3d=TRUE)
#-------------------------------------------------------------------------#
windows()
hist(altura[,1], probability=TRUE,xlab="",ylab="",main="Altura",sub="",col=3)
lines(density(altura[,1]),col=4)
rug(altura[,1])
#-------------------------------------------------------------------------#
windows()
hist(log(altura[,1]), probability=TRUE,xlab="",ylab="",main="",sub="",col=3)
lines(density(log(altura[,1])),col=4)
rug(log(altura[,1]))
#-------------------------------------------------------------------------#
windows()
boxplot(altura[,1],xlab="",ylab="",main="Altura",col="Lightblue")
#-------------------------------------------------------------------------#
windows()
qqnorm(altura[,1],xlab="",ylab="",main="")
qqline(altura[,1], col = 2)
#-------------------------------------------------------------------------#
geo.data<- as.geodata(datos2,coords.col=1:2, data.col=3)
plot(geo.data)
#-------------------------------------------------------------------------#
windows()
plot(coord[,1],coord[,2],xlab="",ylab="",main="",type="p",lty=8,pch=20,col=3)
#-------------------------------------------------------------------------#
windows()
plot(coord[,1],altura[,1],xlab="Latitud",ylab="Altura",main="Latitud vs Altura"
,type="p",lty=8,pch=20,col=3)
#-------------------------------------------------------------------#
windows()
plot(altura[,1],coord[,2],xlab="Longitud",ylab="Altura",main="Longitud vs Altura"
,type="p",lty=8,pch=20,col=3)
#-------------------------------------------------------------------------#
Altura<-interp.new(coord[,1],coord[,2],altura[,1])
mg<-merge(Altura$x,Altura$y)
#----------Interpolación de la altura Basal----------------------------------#
image(Altura,main="Altura",xlab="Longitud",ylab="Latitud")
contour(Altura,add=TRUE,nlev=15,filled=TRUE)
#-------------Distribución Espacial de la altura Basal-----------------------#
#ángulos de visión.
windows()
persp(Altura,col = "7",scale = TRUE, shade = 0.5, border = NA,box = TRUE,
main="(c)",theta = -150, phi = 30,ylab="Latitud",xlab="Longitud",zlab="Altura")
windows()
persp(Altura,col = "7",scale = TRUE, shade = 0.5, border = NA,box = TRUE,
```

```
main="(d)",theta = -10, phi = 30,ylab="Latitud",xlab="Longitud",zlab="Altura")
windows()
persp(Altura,col = "7",scale = TRUE, shade = 0.5, border = NA,box = TRUE,
main="(c)",theta = 30, phi = 30,ylab="Latitud",xlab="Longitud",
zlab="Cantidad Extraída")
#-------------------------------------------------------------------------#
#--------------Lectura de datos originales como geodata-------------------#
#-------------------------------------------------------------------------#
DAltura<-as.geodata(datos2,coords.col=1:2,data.col=3)
altura.varg<-variog(DAltura,estimator.type="classical")
altura.varg
plot(altura.varg,main="",xlab="h",pch=20,col=2,ylab=expression(gamma(h)))
#-------------------------------------------------------------------------#
sc1<- variofit(altura.varg,ini=c(0.9,0.1),weights="equal",cov.model = "exponential")
sc2<- variofit(altura.varg,ini=c(0.9,0.1),weights="equal",cov.model = "gaussian")
sc3<- variofit(altura.varg,ini=c(0.9,0.1),weights="equal",cov.model = "spherical")
sc4<- variofit(altura.varg,ini=c(0.9,0.1),weights="equal",cov.model = "matern")
#-------------------------------------------------------------------#
ols <- variofit(altura.varg,ini=c(0.9,0.1),weights="equal",cov.model = "matern")
wls <- variofit(altura.varg,ini=c(0.9,0.1),weights="cressie")
ml <- likfit(DAltura,ini=c(0.9,0.1),lik.method = "ML",cov.model = "exponential")
rml <- likfit(DAltura,ini=c(3.4166,0.25),lik.method = "RML",cov.model = "exponential")
#------------Ajuste del Modelo de Semivariograma-----------------------------#
#----------model parameters estimated by WLS (weighted least squares)----------#
altura.var.ajuste<-variofit(altura.varg, fix.nugget =FALSE,ini.cov.pars=c(0.8,0.25))
altura.var.ajuste
#-------------------------------------------------------------------#
altura.var.ajuste1<-variofit(basal.varg,cov.model="spherical", fix.nugget
=FALSE,ini.cov.pars=c(7,0.25),nugget=2)
altura.var.ajuste1
#-------------------------------------------------------------------#
#tausq sigmasq     phi
#0.9762  3.4166  0.2500
#-------------------------------------------------------------------#
basal.varg<-variog(DAltura,estimator.type="classical")
plot(basal.varg,main="Ajuste de semivariograma Exponencial",
ylab=expression(gamma(h)),xlab="h")
lines.variomodel(cov.model="exp",cov.pars=c(3.4166,0.2500),
nug=0.9762,max.dist=9000, col=2)
#-------------------------------------------------------------------#
ols <- variofit(altura.varg,ini=c(0.1,0.2),weights="equal",
cov.model = "matern",kappa=1)
wls <- variofit(altura.varg,ini=c(3.4166,0.25),weights="cressie",kappa=1)
#-------------------------------------------------------------------#
ml <- likfit(DAbasal,ini=c(3.4166,0.25),trend="1st",lik.method = "ML",
cov.model = "exponential")
rml <- likfit(DAbasal,ini=c(3.4166,0.25),trend="1st",lik.method = "RML",
cov.model = "exponential")
#-------------------------------------------------------------------#
#------------Ajuste del Modelo de Semivariograma-----------------------------#
#----------model parameters estimated by REML, ML-----------------------------#
geo.data<- as.geodata(datos2,coords.col=1:2, data.col=3)
```

```
rml1<-likfit(geo.data,geo.data$coords,data=geo.data$data,ini=c(0.1,0.2),
lik.method="REML",cov.model = "exponential")
rml2<-likfit(geo.data,geo.data$coords,data=geo.data$data,ini=c(0.1,0.2),
lik.method="REML",cov.model = "spherical")
#-----------------------------------------------------------------#
```