



UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA DE MINAS

DESCUBRIMIENTO DE UNIDADES GEOMETALÚRGICAS POR MEDIO DE
ANÁLISIS DE CONGLOMERADOS GEOESTADÍSTICO

TESIS PARA OPTAR AL GRADO DE
DOCTOR EN INGENIERÍA DE MINAS

ROBERTO MIGUEL FUSTOS TORIBIO

PROFESOR GUÍA:
XAVIER EMERY

MIEMBROS DE LA COMISIÓN:
JAVIER RUÍZ DEL SOLAR SAN MARTÍN
JOSÉ SAAVEDRA ROSAS
RONNY VALLEJOS ARRIAGADA

SANTIAGO DE CHILE
2017

Resumen

El modelamiento geometalúrgico de depósitos minerales está basado en el análisis de variables regionalizadas cuantitativas y cualitativas de origen metalúrgico, geológico u otros relacionados. El objetivo de este modelamiento es identificar y delimitar dominios que particionen el depósito mineral, de tal forma que los datos pertenecientes a un mismo dominio posean características similares. Estos dominios son llamados Unidades Geometalúrgicas (U.G.). Actualmente, las herramientas disponibles para identificar estas unidades no incorporan toda la información disponible, en particular, éstas no consideran la naturaleza de las variables de estudio, las que están distribuidas en el espacio y presentan una estructura de correlación espacial. Se presentaron dos propuestas que buscaron resolver este problema incorporando la distribución espacial de los datos, basándose en el Análisis de Conglomerados Jerárquicos y en la teoría de Mezclas de Distribuciones. Se presentó un marco conceptual que resumió las técnicas utilizadas para la definición de Unidades Geometalúrgicas, mostrando las variantes de cada herramienta y la teoría básica de cada método. Las propuestas se basaron en casos sintéticos que reprodujeran la naturaleza de cada problema y se aplicaron a casos de estudio reales con información geoquímica, metalúrgica y geológica. En los casos de datos simulados (estudios de casos sintéticos), fue posible realizar un análisis de sensibilidad de las propuestas postulando escenarios con diferentes complejidades. En gran parte de las simulaciones las propuestas pudieron descubrir con precisión la distribución de las Unidades Geometalúrgicas. En los casos de estudio reales “Minera Escondida” y “Geoquímica del sector Colchane”, las Unidades Geometalúrgicas pudieron ser identificadas y validadas en base a descripciones geológicas de las regiones de interés. Se discutieron las ventajas de las propuestas por sobre los métodos y algoritmos tradicionales, así como diferentes oportunidades de mejoras futuras.

Abstract

The geometallurgical modeling of ore bodies is based on the analysis of quantitative and qualitative regionalized variables of metallurgical, geological or related origin. The objective of this modeling is to identify and delineate domains that partition the ore body, such that the data belonging to the same domain possess similar characteristics. These domains are called Geometallurgical Units (G.U.). Currently, the tools available to identify these units do not incorporate all available information, in particular, they omit the nature of the study variables that are distributed in space and exhibit a spatial correlation structure. Two proposals are presented that seek to solve this problem of the use of spatial information, one from the Hierarchical Clustering Analysis and the other from the Mixture Distribution theory. A theoretical framework is presented to summarize the techniques used in the definition of Geometallurgical Units, showing the variants of each tool and the basic theory of each method. The proposals are based on synthetic cases that reproduce the nature of each problem and are applied to real case studies with geochemical, metallurgical and geological information. In the case of simulated data (synthetic case studies), it is possible to make a sensitivity analysis of the proposals by postulating scenarios with different complexity. In most of the simulations the proposals are able to accurately discover the distribution of the Geometallurgical Units. In the real case studies “Minera Escondida” and “Geochemistry of the Colchane sector”, Geometallurgical Units could be identified and validated based on geological descriptions of the regions of interest. The advantages of the proposals over traditional methods and algorithms, as well as different opportunities for future improvements, are discussed.

A mi esposa, mi madre y familia, gracias.

Agradecimientos

En estas líneas me gustaría dar las gracias a todas las personas que estuvieron de alguna forma relacionadas con este trabajo.

Dar las gracias a mi esposa Trinidad, que me acompañó en este largo y extenso camino. Gracias por soportar las noches sin dormir y mis mañanas a la mañana siguiente.

Gracias a mi madre Asunción y a mi hermano Ivo que creyeron en mi capacidad y fueron la voz que siempre me impulsó a seguir adelante.

Darle las gracias a mi profesor guía sería poco. Gracias a él pude ingresar al programa de doctorado y desde el primer hasta el último día recibí su apoyo, orientación, supervisión y paciencia. Lo siento si muchas veces demoré por algún u otro motivo, pero fue todo con el ánimo de mejorar el trabajo de investigación.

Quisiera agradecer a los miembros de la comisión, los que velaron por la calidad de este trabajo al realizar las correcciones pertinentes y de esa forma me ayudaron a formar el criterio que todo buen investigador debe tener al recibir las críticas.

Quisiera agradecer también al Centro de Excelencia Internacional en Minería y Procesamiento de Minerales CSIRO-Chile, por el financiamiento recibido durante el desarrollo de esta tesis, al centro Advanced Mining Technology Center (AMTC) de la Universidad de Chile y al Departamento de Ingeniería de Minas de la Universidad de Chile, por su colaboración y disposición durante todo este tiempo. Por último, a la Comisión Nacional de Investigación Científica y Tecnológica (CONICYT), a través de los proyectos Fondecyt 1170101 y Conicyt PIA Anillo ACT1407 por su apoyo logístico.

A todos ellos, muchas gracias.

Tabla de Contenido

1. Introducción	1
1.1. Geometalurgia	1
1.2. Modelamiento Geometalúrgico	3
1.3. Planteamiento del problema	5
1.4. Hipótesis de trabajo	6
1.5. Objetivos	6
1.6. Alcances	6
2. Marco Teórico	7
2.1. Aspectos generales del análisis de conglomerados	7
2.1.1. Generalidades	7
2.1.2. Paso previo al método de agrupamiento	8
2.2. Conglomerados Jerárquicos	8
2.2.1. Medidas de proximidad	8
2.2.2. Algoritmos de agrupamiento	13
2.2.3. Criterios para testear la bondad de ajuste	20
2.3. Modelos de Mezclas de Distribuciones	21
2.3.1. Mezcla de Distribuciones Normales	21
2.3.2. Parámetros en la mezcla de densidades normales	22
2.3.3. Estimación Máximo Verosímil de parámetros en las componentes de la mezcla	23
2.4. Estadística Bayesiana	26
2.4.1. Introducción	26
2.4.2. Introducción a Estadística Bayesiana	26
2.4.3. Criterios de información para la selección, comparación y análisis de convergencia	31
2.5. Geoestadística	33
2.5.1. Conceptos y definiciones básicas	33
2.5.2. Análisis estructural	36
2.5.3. Geoestadística lineal clásica	37
2.6. Modelamiento Geometalúrgico usando Análisis de Conglomerados	39
2.6.1. Validación e interpretación de resultados	39
2.6.2. Límites de las técnicas de agrupamiento con datos regionalizados	40
2.6.3. Aplicaciones a modelamiento geológico o geometalúrgico	41
3. Conglomerados Geoestadísticos basados en Métodos Jerárquicos	42

3.1.	Introducción	42
3.2.	Metodología propuesta	42
3.3.	Caso de estudio sintético	45
3.3.1.	Metodología de simulación	45
3.3.2.	Resumen de resultados	51
3.4.	Caso de estudio real: Geoquímica sector Colchane, I Región de Tarapacá . .	56
3.4.1.	Descripción de la zona de estudio	56
3.4.2.	Descripción de la base de datos	57
3.4.3.	Conocimiento geológico del área	61
3.4.4.	Aplicación	62
3.4.5.	Resumen de resultados	64
3.5.	Conclusiones parciales	67
4.	Conglomerados basados en mezcla de distribuciones	69
4.1.	Introducción	69
4.2.	Metodología propuesta	70
4.3.	Caso de estudio sintético	78
4.3.1.	Metodología de simulación	78
4.3.2.	Resumen de resultados	79
4.4.	Caso de estudio real: Minera Escondida	86
4.4.1.	Descripción de la base de datos	86
4.4.2.	Aplicación	90
4.4.3.	Resumen de resultados	92
4.5.	Conclusiones parciales	98
5.	Discusiones generales	99
6.	Conclusiones	102
	Bibliografía	104
A.	Caso de estudio de geoquímica, sector Colchane, I Región de Tarapacá	113
A.1.	Resultados del caso de estudio aplicando Conglomerados Geoestadísticos con el método de Ward	113
A.2.	Resultados del caso de estudio aplicando Conglomerados Geoestadísticos con el método de la distancia máxima (Complete linkage)	115
A.3.	Resultados del caso de estudio aplicando Conglomerados Geoestadísticos con el método de la distancia mínima (Single linkage)	117
A.4.	Resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de Ward sin coordenadas espaciales	119
A.5.	Resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de la distancia máxima (Complete linkage) sin coordenadas espaciales	121
A.6.	Resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de la distancia mínima (Single linkage) sin coordenadas espaciales .	123
A.7.	Resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de Ward con coordenadas espaciales	125
A.8.	Resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de la distancia máxima (Complete linkage) con coordenadas espaciales	127

A.9. Resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de la distancia mínima (Single linkage) con coordenadas espaciales .	129
B. Caso de estudio yacimiento Escondida, II Región de Antofagasta	131
B.1. Resultados del caso de estudio aplicando Mezclas de Distribuciones Geoestadísticas para la ley de cobre (Cu) total y ley de fierro (Fe) total	133
B.2. Resultados del caso de estudio aplicando Mezclas de Distribuciones Gaussianas sin coordenadas espaciales para la ley de cobre (Cu) total y ley de fierro (Fe) total	135
B.3. Resultados del caso de estudio aplicando Mezclas de Distribuciones Gaussianas con coordenadas espaciales para la ley de cobre (Cu) total y ley de fierro (Fe) total	137

Índice de Tablas

3.1. Matriz de mesetas caso inicial.	46
3.2. Matriz de mesetas caso secundario.	48
3.3. Estadísticas descriptivas para variables en la base de datos.	57
4.1. Resumen de medias a posteriori para los casos de estudio en las 100 simulaciones	79
4.2. Resumen de pertenencia y porcentaje de asignación a posteriori para las 100 simulaciones	81
4.3. Resumen de log-verosimilitud para las 100 simulaciones	84
4.4. Estadísticas descriptivas para variables en la base de datos	87
4.5. Descriptivos de los parámetros μ para la Ley de Cu Total y Ley de Fe Total por unidad descubierta	91
4.6. Descriptivos de los parámetros μ para la Ley de Cu Total y Ley de Fe Total por unidad descubierta	92
4.7. Descriptivos para las verosimilitudes en el caso de 3 unidades descubiertas .	93
4.8. Descriptivos de los parámetros de correlación espacial en el caso de tres unidades descubiertas	94
4.9. Descriptivos para las verosimilitudes en el caso de cuatro unidades descubiertas	96
4.10. Descriptivos de los parámetros de correlación espacial en el caso de cuatro unidades descubiertas	96

Índice de Ilustraciones

1.1. Ejemplos de continuidad espacial.	1
1.2. Ejemplo de definición de Unidades Geometalúrgicas (U.G.M.).	2
1.3. Actividades de la Geometalurgia.	2
1.4. Análisis de Conglomerados o Clustering.	4
1.5. Mezcla de distribuciones.	5
2.1. Coeficiente de Correlación (Pearson = 0,3071 y Spearman = 0,7627).	10
2.2. Dendrograma	13
2.3. Procedimiento de la mínima distancia	14
2.4. Procedimiento de la máxima distancia	14
2.5. Procedimiento de la distancia promedio	15
2.6. Procedimiento de la distancia promedio ponderada	15
2.7. Procedimiento basado en el centroide ponderado	15
2.8. Procedimiento basado en el centroide no ponderado	16
2.9. Ilustración del Método de Ward	17
2.10. Algoritmo K-Medias	19
2.11. Histograma de las razones de la frente sobre la longitud del cuerpo de los 1000 cangrejos	22
3.1. Disimilitud entre valores de leyes de cobre para dos observaciones.	43
3.2. Disimilitud entre dos observaciones.	44
3.3. Simulación caso inicial.	47
3.4. Variables cosimuladas en cada una de las U.G.M. de forma independiente.	47
3.5. Simulación caso secundario.	48
3.6. Variables cosimuladas en cada una de las U.G.M. de forma independiente.	49
3.7. Resultados del clustering jerárquico usando la Distancia Geoestadística y el Método de Ward para tres simulaciones del caso inicial.	52
3.8. Resultado del clustering jerárquico con la Distancia Geoestadística (arriba, derecha) y con la Distancia Euclidean (abajo), en base a las 100 realizaciones utilizando la máxima frecuencia como clasificador.	53
3.9. Resultados del clustering jerárquico usando la Distancia Geoestadística y el Método de Ward para tres simulaciones del caso secundario.	54
3.10. Resultado del clustering jerárquico con la Distancia Geoestadística (arriba, derecha) y con la Distancia Euclidean (abajo), en base a las 100 realizaciones usando la máxima frecuencia como clasificador.	55
3.11. Zona de estudio, sector Colchane, I Región de Tarapacá.	56

3.12. Distribución espacial y frecuencial para la variable Cu (ppm).	58
3.13. Distribución espacial y frecuencial para la variable Mo (ppm).	58
3.14. Distribución espacial y frecuencial para la variable Ag (ppm).	58
3.15. Distribución espacial y frecuencial para la variable As (ppm).	59
3.16. Distribución espacial y frecuencial para la variable Pb (ppm).	59
3.17. Distribución espacial y frecuencial para la variable Au (ppb).	59
3.18. Distribución espacial y frecuencial para la variable RTP.	60
3.19. Distribución espacial y frecuencial para la variable Bouger.	60
3.20. Distribución espacial y frecuencial para la variable geológica Período.	61
3.21. Distribución espacial y frecuencial para la variable geológica Época.	61
3.22. Distribución espacial y frecuencial para la variable geológica Tipo de roca.	62
3.23. Cálculo y ajuste del modelo de semivariograma.	62
3.24. Distribución espacial para los clústers formados por el Método de Ward.	63
3.25. Asociación de conglomerados con el Período geológico.	64
3.26. Asociación de conglomerados con la Época geológica.	64
3.27. Asociación de conglomerados con el Tipo de roca.	64
3.28. Variables geoquímicas Cu (ppm), Mo (ppm), Ag (ppm) y As (ppm) por conglomerado formado.	65
3.29. Variables geoquímicas Pb (ppm) y Au (ppb), magnetométrica RTP y gravimétrica Bouger por conglomerado formado.	66
4.1. Proceso Gaussiano con media $\mu = \mathbf{1}$, varianza del proceso espacial $\sigma^2 = 1$, alcance $\phi = 40$ y covarianza exponencial.	71
4.2. Influencia del número de pares de muestras en la construcción del variograma experimental.	72
4.3. Paso preliminar utilizando algoritmo K-Medias en base a las coordenadas geográficas	74
4.4. Tipos de mallado para definir las particiones $\mathbf{Z} = (Z_1, \dots, Z_P)^\top$	75
4.5. Evolución de grilla irregular durante algoritmo iterativo	75
4.6. Casos de estudio con diferencias en la definición de sus fronteras	79
4.7. Distribución a posteriori de las medias poblacionales y gráfico de pertenencia máxima para los casos de estudio en una de las simulaciones	80
4.8. Distribución de medias a posteriori para los casos de estudio en las 100 simulaciones	81
4.9. Probabilidades de pertenencia a posteriori para los casos de estudio en una de las simulaciones	82
4.10. Resumen de Log-verosimilitudes para las 100 simulaciones en cada caso de estudio	83
4.11. Distribución de pertenencia a posteriori en los casos de estudio para una simulación	84
4.12. Distribución de los parámetros espaciales para los casos de estudio en una de las simulaciones (valores reales: $\sigma^2 = 0,1$ y $\phi = 50$)	85
4.13. Zona de estudio, Mina Escondida, II Región de Antofagasta	86
4.14. Mapa geológico del yacimiento Escondida (Richards y cols. (2001))	87
4.15. Mapa de alteraciones hidrotermales del yacimiento Escondida (Padilla y cols. (2001))	88
4.16. Distribución de las leyes de Cu Total y Fe Total	88

4.17. Distribución frecuencial, en planta y espacial para las variables Ley de Cu Total y Ley de Fe Total	89
4.18. Distribución en planta y espacial por mezcla para la Ley de Cu Total y la Ley de Fe Total para tres unidades descubiertas	90
4.19. Distribución en planta y espacial por mezcla para la Ley de Cu Total y la Ley de Fe Total para cuatro unidades descubiertas	91
4.20. Análisis de convergencia de los parámetros μ (medias) de cada mezcla para la Ley de Cu Total y la Ley de Fe Total	93
4.21. Distribuciones a posteriori de los parámetros μ (medias) de cada mezcla para la Ley de Cu Total y la Ley de Fe Total	94
4.22. Análisis de convergencia de los parámetros μ (medias) de cada mezcla para la Ley de Cu Total y la Ley de Fe Total	95
4.23. Distribuciones a posteriori de los parámetros μ (medias) de cada mezcla para la Ley de Cu Total y la Ley de Fe Total	95
4.24. Comparacion de resultados con conocimiento geológico de la zona de estudio	97
A.1. Distribución espacial para los clústers formados por el Método de Ward . . .	113
A.2. Distribución de las variables de estudio en función de los clústers formados por el Método de Ward	114
A.3. Distribución espacial para los clústers formados por el método de la distancia máxima (Complete linkage)	115
A.4. Distribución de las variables de estudio en función de los clústers formados por el método de la distancia máxima (Complete linkage)	116
A.5. Distribución espacial para los clústers formados por el método de la distancia mínima (Single linkage)	117
A.6. Distribución de las variables de estudio en función de los clústers formados por el método de la distancia mínima (Single linkage)	118
A.7. Distribución espacial para los clústers formados por el Método de Ward . . .	119
A.8. Distribución de las variables de estudio en función de los clústers formados por el Método de Ward	120
A.9. Distribución espacial para los clústers formados por el método de la distancia máxima (Complete linkage)	121
A.10. Distribución de las variables de estudio en función de los clústers formados por el método de la distancia máxima (Complete linkage)	122
A.11. Distribución espacial para los clústers formados por el método de la distancia mínima (Single linkage)	123
A.12. Distribución de las variables de estudio en función de los clústers formados por el método de la distancia mínima (Single linkage)	124
A.13. Distribución espacial para los clústers formados por el Método de Ward . . .	125
A.14. Distribución de las variables de estudio en función de los clústers formados por el Método de Ward	126
A.15. Distribución espacial para los clústers formados por el método de la distancia máxima (Complete linkage)	127
A.16. Distribución de las variables de estudio en función de los clústers formados por el método de la distancia máxima (Complete linkage)	128
A.17. Distribución espacial para los clústers formados por el método de la distancia mínima (Single linkage)	129

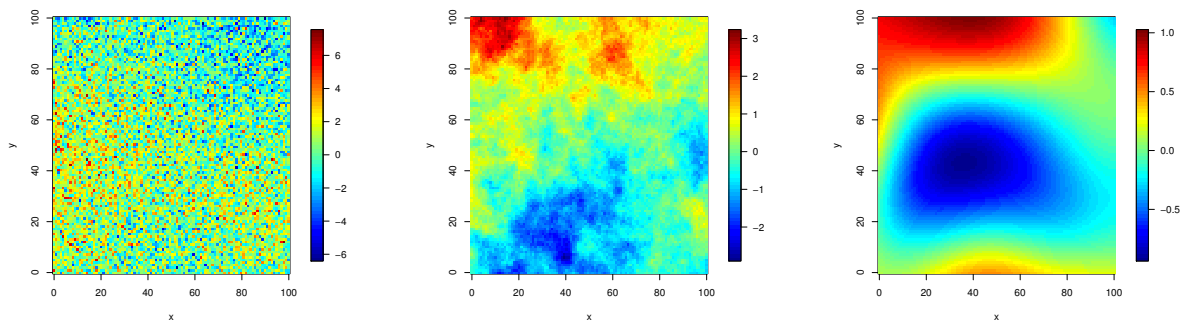
A.18. Distribución de las variables de estudio en función de los clústers formados por el método de la distancia mínima (Single linkage)	130
B.1. Comparación entre resultados de las tres mezclas encontradas para la ley de cobre (Cu) total	131
B.2. Comparación entre resultados de las tres mezclas encontradas para la ley de hierro (Fe) total	132
B.3. Distribución de las variables de estudio en función de las tres mezclas encontradas	133
B.4. Distribución de las variables de estudio en función de las cuatro mezclas encontradas	134
B.5. Distribución de las variables de estudio en función de las tres mezclas encontradas	135
B.6. Distribución de las variables de estudio en función de las cuatro mezclas encontradas	136
B.7. Distribución de las variables de estudio en función de las tres mezclas encontradas	137
B.8. Distribución de las variables de estudio en función de las cuatro mezclas encontradas	138

Capítulo 1

Introducción

1.1. Geometalurgia

En Evaluación de Yacimientos, uno de los aspectos cruciales es conocer a cabalidad la distribución espacial de variables regionalizadas (atributos geológicos, geoquímicos, metalúrgicos, etc.), así como su continuidad espacial. Este último concepto es utilizado en todas las fases del proceso de evaluación y permite obtener resultados que son confiables y lo más parecidos a la realidad bajo diferentes escenarios (Figura 1.1). El conocer cómo se distribuyen los atributos de interés en una región determinada no es el único aspecto crucial en las fases del proceso minero, pues existen factores determinantes que pueden condicionar la factibilidad técnica y económica de los proyectos. Muchas veces, estos factores están relacionados con características geológicas y metalúrgicas de los depósitos. El campo que estudia la incorporación de estos aspectos en los modelos de recursos se denomina Geometalurgia (Miller y cols. (2010)), la que se define como una disciplina que integra variables geológicas y metalúrgicas, creando modelos que relacionan propiedades intrínsecas de la roca con respuesta metalúrgica.



(a) Escenario con baja continuidad espacial (b) Escenario con continuidad espacial irregular (c) Escenario con alta continuidad espacial

Figura 1.1: Ejemplos de continuidad espacial.

Además, la Geometalurgia permite definir las zonas del depósito donde el mineral es factible técnicamente de ser procesado, como también, proyectar las primeras ecuaciones de recuperación y el consumo de los principales insumos del proceso (Rosales (2012)). Los estudios geometalúrgicos se definen en general sobre dominios o unidades geometalúrgicas (UGM), es decir, regiones del espacio con un comportamiento geológico y metalúrgico similar (Figura 1.2). La modelación geometalúrgica considera una serie de actividades (Figura 1.3) a diferentes escalas usando técnicas provenientes del área estadística multivariante y Data Mining, debido a su potencia y a la gran cantidad de información que pueden resumir.

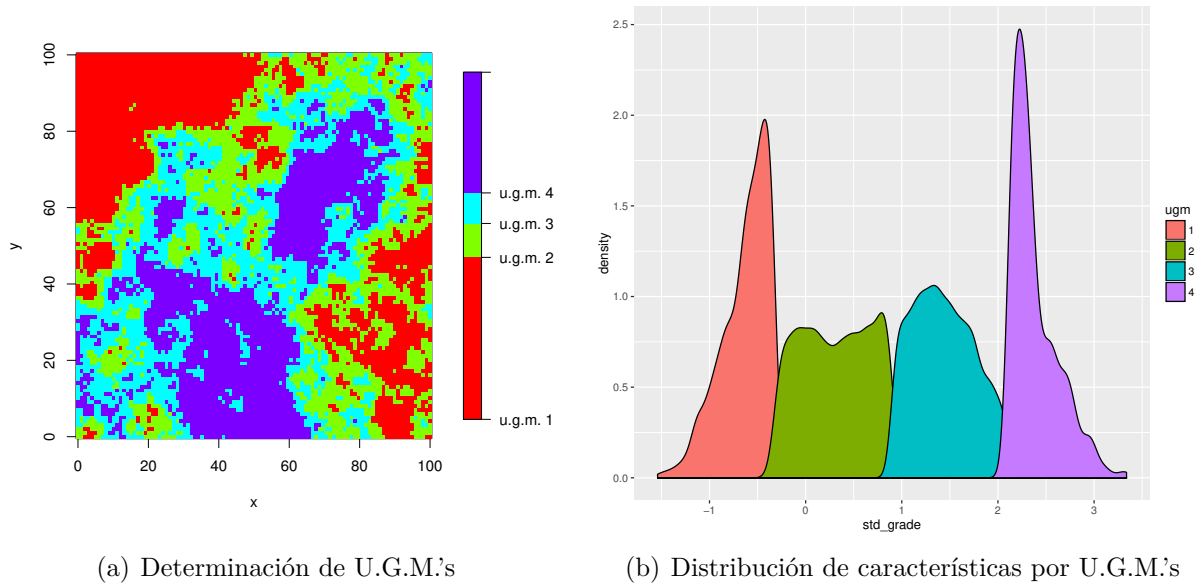
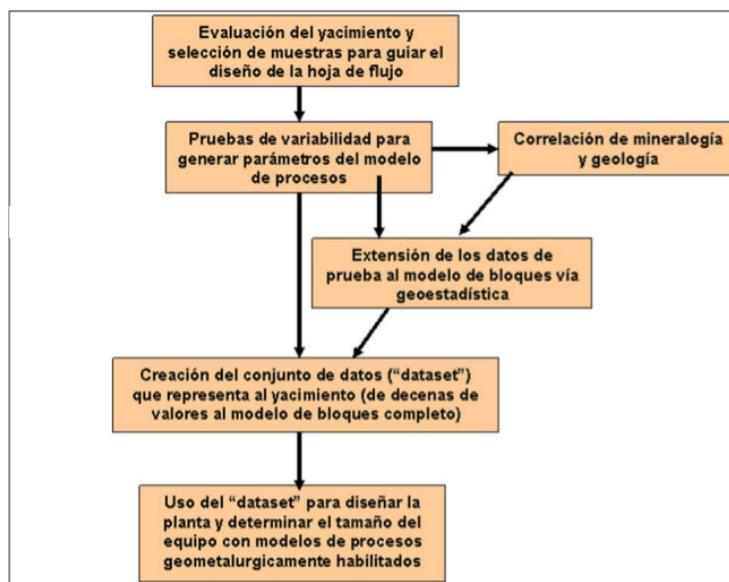


Figura 1.2: Ejemplo de definición de Unidades Geometalúrgicas (U.G.M.).



Fuente: SGS Minerals Services, 2007. <http://www.sgs.cl/>

Figura 1.3: Actividades de la Geometalurgia.

El objetivo es determinar relaciones entre la geología y la metalurgia, incorporando datos de interés como litología, alteración, mineralogía, ley del metal de interés, etc; a fin de retroalimentar a la exploración y realizar la estimación de reservas (Hallewell (2009)). Como producto se obtiene una agrupación de unidades geometalúrgicas con un comportamiento metalúrgico determinado, a las cuales se les asigna las mismas ecuaciones de recuperación y de consumo de insumos de proceso. De esta forma se realiza una estimación más exacta de los recursos y reservas. Los métodos de análisis de datos multivariados son frecuentemente utilizados en estos estudios, con bases de datos que contienen una gran cantidad de variables observadas (Dagnelie (1975)). Estos métodos permiten obtener diversas conclusiones e información que ayudan a los investigadores a descubrir relaciones entre individuos o variables. Muchas veces, la intención de utilizar estos métodos se centra en la capacidad de explicar una gran cantidad de información en función de la menor dimensionalidad posible (Jolliffe (1986)), así como poder relacionar variables observadas, individuos muestreados o combinaciones de ambos. Podemos agrupar los métodos multivariados de acuerdo a su orientación, es decir, si están dirigidos a las variables observadas o a los individuos de la muestra o población. Cuando el interés se centra en los individuos muestreados, surgen los métodos de clasificación y de agrupamiento en general. Esta última técnica, que corresponde a un tipo de aprendizaje no supervisado del área de Data Mining, produce grupos de datos de los cuales inicialmente se desconoce su pertenencia, a diferencia de los métodos discriminantes (Hardle y Simar (2007)), donde desde un principio se conocen los grupos de origen.

1.2. Modelamiento Geometalúrgico

Los métodos orientados a crear agrupaciones de individuos en función de sus variables observadas, se definen como técnicas de Agrupamiento, Análisis de Conglomerados o Clustering en inglés (Fortier y Solomon (1996)). Estas agrupaciones deben respetar ciertos criterios a la hora de formarse (Hartigan (1975)) y lo que se busca es tener el mayor grado de homogeneidad posible dentro de cada agrupación o conglomerado, con individuos de grupos diferentes lo más heterogéneos posible (Figura 1.4). El uso de estos métodos se ha extendido a diversas áreas (Bailey (1994)), académicas e industriales (Veyssieres y Plant (1998)), debido a su gran potencial y beneficios otorgados, que permiten obtener un mayor conocimiento de los datos cuando se trabaja con una alta dimensionalidad (Zhang y cols. (1996)). No obstante, existen problemas aún no resueltos cuando se trabaja con estos métodos. Por ejemplo, la carencia de una solución única en cuanto a pertenencia a los grupos formados, pues esta solución dependerá de varios factores involucrados en el proceso de formación de conglomerados (Fraley y Raftery (1998)). En vista de que el análisis de conglomerados es una técnica de tipo exploratorio, el resultado siempre será la formación de estas agrupaciones, independiente si existe una real relación entre los individuos. Además, la solución depende de las variables utilizadas en la construcción de los conglomerados, influyendo en el resultado de manera relevante ante adición u omisión de alguna de ellas (Strehl y cols. (2000)).

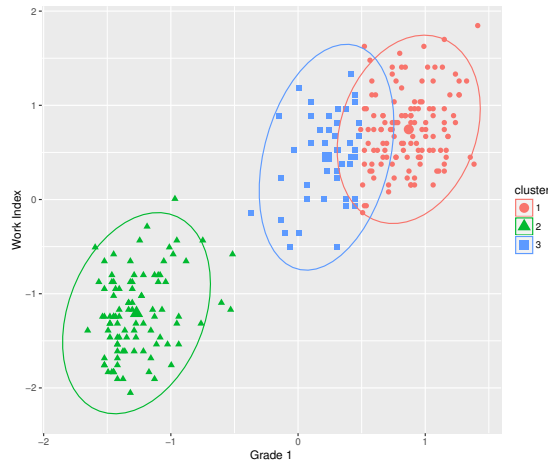


Figura 1.4: Análisis de Conglomerados o Clustering.

El objetivo principal del análisis de conglomerados es la formación de agrupaciones con individuos semejantes entre sí y muy disímiles con individuos pertenecientes a agrupaciones distintas. Existen diferentes enfoques para construir estos conglomerados; algunos se basan en medidas de similitud entre individuos y otros en maximizar la distribución de las variables de interés asociadas a los grupos.

La metodología básica para implementar el primer enfoque se basa en tres conceptos fundamentales: la medida de similitud (Strehl y Ghosh (2000)), la forma en cómo se crean las agrupaciones (Tryon (1970)) o conglomerados y el número de grupos (Tibshirani y cols. (2001)) a crear o descubrir. El primer concepto hace relación con los métodos existentes para definir el grado de similitud entre los individuos, los que dependerán fuertemente del tipo de objeto de estudio (Dice (1945)), al igual que de las variables medidas. El cómo se crean las agrupaciones (Bonner (1964)) está relacionado con el orden de pertenencia de los individuos hacia cada conglomerado, que se basa en la similitud obtenida por la etapa anterior. El último concepto es el más discutido, pues no existe una regla general que permita identificar el número óptimo de conglomerados a formar (Do-Jong y cols. (2001); Wang y Yu (2001)). Podemos colocarnos en los casos extremos, donde por un lado tenemos un solo conglomerado que incluye a todos los individuos y la heterogeneidad es alta; y en el caso contrario tendríamos tantos conglomerados como individuos, donde el grado de homogeneidad es completo. En estos métodos es crucial buscar el punto de equilibrio entre el número de conglomerados a formar, el grado de homogeneidad de cada uno de ellos y la variabilidad controlada por cada agrupación (Han (2001)). El segundo enfoque habla sobre lo que se conoce como mezcla de distribuciones (Figura 1.5) (Bilmes y cols. (1998); Shental y cols. (2004)), que se define como una combinación de densidades estadísticas provenientes de poblaciones heterogéneas, para las cuales asumimos una función de densidad respectiva. Los modelos de mezclas han resuelto problemas de distintas disciplinas. Por ejemplo, en el área de la salud al detectar diferencias en las concentraciones de compuestos en pacientes (Bahoura y Pelletier (2004)), a modo de controlar los parámetros para obtener un rango óptimo de los medicamentos, o en el área de la antropología para determinar diferencias entre especies basadas en mediciones físicas a modo de caracterizar nuevos hallazgos de manera eficiente y con el menor riesgo (Rajesh y Punithavalli (2014)).

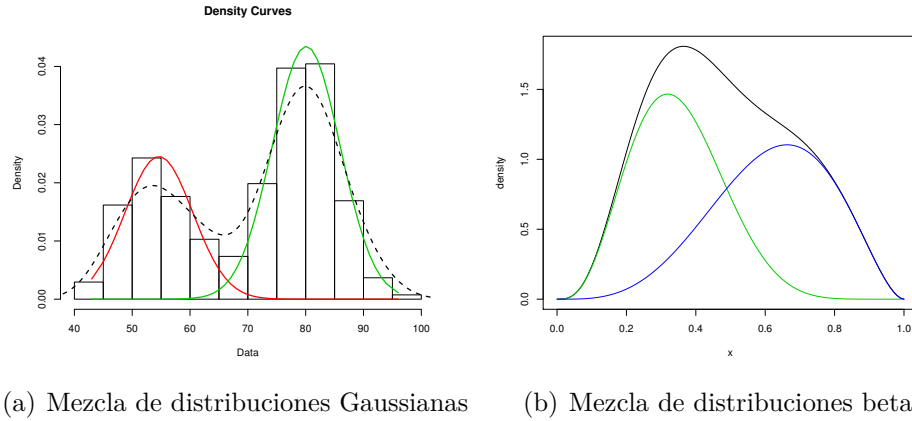


Figura 1.5: Mezcla de distribuciones.

La metodología para implementar el segundo enfoque se basa en definir en primer lugar la distribución de las variables con las que se va a trabajar, luego se escogen los parámetros que serán los que controlen las características de las distribuciones y por último se elige un algoritmo para poder encontrar los valores de los parámetros bajo el criterio de maximizar la función de verosimilitud de la muestra. Usualmente se utilizan distribuciones Gaussianas para describir el comportamiento de las mezclas, lo que conlleva a tener que ajustar los parámetros de media y varianza para cada densidad que compone la mezcla. Sin embargo, no existen restricciones para el uso de cualquier otra densidad, por lo que en casos donde no se cumplan las condiciones para ajustar un modelo Gaussiano se puede experimentar con otras distribuciones que sean mucho más flexibles (Figura 1.5(b)).

1.3. Planteamiento del problema

En los estudios de datos geometalúrgicos se cuenta con variables de diversos orígenes, las que pueden ser cuantitativas o cualitativas, por lo que se debe tener cuidado al aplicar diversas herramientas de análisis. Estos estudios en general tienen como objetivo final el generar unidades geometalúrgicas, que se definen como dominios con características geológicas y metalúrgicas similares. Las herramientas actualmente empleadas para construir estas unidades no consideran la correlación espacial existente entre los atributos cuantitativos de origen metalúrgico y tampoco incorporan atributos categóricos de origen geológico, por lo que una gran cantidad de información queda fuera de los estudios. Estas fuentes de información son la base de la Evaluación de Yacimientos, por lo que dejarlas fuera de los análisis puede generar problemas en la definición de estas unidades. Además, si no se toma en cuenta la escala de medición de las diversas variables disponibles es posible que se generen errores de ponderación de atributos cuando se definan similitudes entre datos.

1.4. Hipótesis de trabajo

Todas las variables comparten una característica en común: son variables medidas en el espacio. Esta característica hace que se les denomine Variables Regionalizadas, definición proveniente del área de la Geoestadística. Esta característica establece que en cada localidad del espacio existe una variable o proceso aleatorio, cuya realización corresponde a la muestra con la que disponemos en las bases de datos y además, cada variable del espacio depende de alguna forma de los procesos en localidades cercanas a ella. Por este motivo, a pesar de muchas veces no observar alguna relación entre variables, esto no quiere decir que el proceso completo se comporte de forma aleatoria, pues el tipo de dependencia puede estar reflejado en una cierta continuidad espacial. Esto y la falta de información exhaustiva que provoca incertidumbre en las estimaciones de recursos y reservas, sustentan la necesidad de considerar todas las propiedades pasadas por alto en un análisis estadístico multivariado tradicional. Asumiremos la existencia de unidades geometalúrgicas que presentan un grado de similitud entre los atributos pertenecientes a una misma unidad y además, comparten una estructura de correlación espacial que es independiente entre éstas.

1.5. Objetivos

Como objetivo general, la propuesta de tesis busca generar nuevos algoritmos basados en el análisis de conglomerados jerárquicos y las mezclas de distribuciones Gaussianas para el agrupamiento de datos regionalizados; y aplicar los resultados en el descubrimiento de unidades geometalúrgicas. Como objetivos específicos, en la primera propuesta se busca definir una medida de disimilitud que incorpore tanto los valores de los atributos medidos así como la correlación espacial existente. En la segunda propuesta se busca definir un algoritmo que maximice la verosimilitud de las unidades geometalúrgicas encontradas utilizando las estructuras de correlación espacial propias de cada unidad. En ambos casos se realizarán estudios en base a casos sintéticos que reproduzcan las interacciones entre las muestras simuladas y se realizarán comparaciones entre las propuestas y los resultados provenientes de metodologías tradicionales.

1.6. Alcances

El tema de investigación será desarrollado mostrando en una etapa inicial un panorama general de los métodos de análisis de conglomerados existentes y que son utilizados en estudios geometalúrgicos, mostrando los conceptos básicos que los sustentan y las diferentes opciones que generan soluciones, todas ellas igualmente aceptadas. El problema de investigación será abordado en base a las propuestas formuladas mostrando los resultados obtenidos en base a casos sintéticos y aplicaciones a datos reales. Los resultados serán aplicados al modelamiento geometalúrgico, definiendo dominios en el espacio con atributos similares y que compartan una estructura de correlación espacial. Esta información servirá para tener una mejor caracterización de los depósitos minerales y para reducir la variabilidad del proceso aguas abajo.

Capítulo 2

Marco Teórico

2.1. Aspectos generales del análisis de conglomerados

2.1.1. Generalidades

Los métodos de conglomerados tratan de identificar la forma en que distintos individuos a los cuales se les ha medido diversas características se relacionan entre sí y es posible agruparlos en base a sus similitudes, por lo que caen en la categoría de técnicas multivariantes que permiten comprender de mejor forma las posibles relaciones entre los individuos de una población ([Hair y cols. \(1998\)](#)). Al aplicar estos métodos se obtienen los siguientes resultados:

Caracterización global: Se obtiene una clasificación de individuos o variables relacionadas entre sí con un alto grado de similitud, la que puede ser contrastada con el conocimiento a priori de relaciones existentes.

Simplificación de la información: Debido a que los individuos pertenecientes a una misma categoría son relativamente homogéneos entre sí, es posible representarlos en base a sus características más relevantes.

Identificación de nuevas relaciones: Como los conglomerados están formados en base a las similitudes entre individuos, se pueden obtener nuevos conocimientos sobre relaciones existentes que no son visibles a partir de los datos en forma individual.

Los resultados obtenidos están fuertemente influenciados por elementos previos involucrados en el análisis de conglomerados. Estos elementos están insertos en la metodología que se ha de seguir en el proceso de agrupamiento (clustering).

2.1.2. Paso previo al método de agrupamiento

Una vez que se seleccionan las variables que se incluirán en el análisis de conglomerados, es necesario verificar la integridad de los datos. Ya se mencionó que la clasificación obtenida depende de varios factores, donde podemos destacar como punto de partida la integridad de los datos. Los datos de entrada, que servirán para obtener las medidas de similitud, deben ser representativos y confiables, pues cualquier valor aberrante puede alterar drásticamente los resultados y por ende la asociación de individuos en las etapas subsecuentes (Dixon (1953); Horn y cols. (2001)). Además, como se verá más adelante, las escalas de medición de algunas variables pueden condicionar la similitud entre los individuos, por lo que un paso previo consistiría en estandarizar todas las variables a una escala común siempre que sea posible para evitar posibles distorsiones o inconsistencias.

2.2. Conglomerados Jerárquicos

2.2.1. Medidas de proximidad

Las medidas de proximidad¹ buscan cuantificar el parentesco entre los individuos o variables que serán agrupadas (Anderberg (1973); Orlóci (2013)). En función del tipo de variable escogida para el estudio, se construye la medida de proximidad entre todos los pares de objetos, para luego agruparlos en conglomerados según distintos procedimientos vistos más adelante. Si la intención es agrupar variables, un caso frecuente para medir proximidad es el utilizar funciones de similitud, que miden la dependencia entre ellas. Los coeficientes de correlación, de Pearson en caso de normalidad de las variables o un alto número de observaciones (ley de los grandes números) y Spearman a falta de alguna distribución conocida, son los más utilizados en este aspecto.

En cambio si el objetivo es agrupar en función de la proximidad entre objetos, existen diversos métodos que originan lo que se conoce como medidas de distancia métrica, que están orientadas a distintos tipos de características medidas. También es posible destacar las medidas de distancia no métricas o de asociación, las que intentan comparar a los individuos en base a características cualitativas (nominales u ordinales).

2.2.1.1. Medidas de similitud entre variables

Partiremos dando la definición formal de lo que es una medida de similitud o las condiciones que debe cumplir una función para ser considerada como tal (Duda y cols. (2012)).

¹Cuando la medida de proximidad mida la cercanía entre variables hablaremos de una medida de similitud y cuando mida la cercanía entre individuos corresponderá a una medida de distancia

Una función $S : V \times V \rightarrow \mathbb{R}$ se dice de similitud si cumple las primeras dos propiedades:

$$\begin{aligned} \forall x, y \in V; \quad s(x, y) &\leq s_0 \\ \forall x, y \in V; \quad s(x, y) &= s(y, x) \\ \forall x, y \in V; \quad s(x, y) = 0 &\implies x = y \\ \forall x, y, z \in V; \quad |s(x, y) + s(y, z)|s(x, z) &\geq s(x, y)s(y, z) \end{aligned} \tag{2.2.1}$$

donde s_0 es un número real finito arbitrario. Además si la función cumple las últimas dos propiedades se dice que corresponde a una función de similitud métrica.

Medida del Coseno

Definiendo dos variables Z_i y Z_j , que son muestreadas sobre m individuos, y dados z_i y z_j los vectores cuyas k -ésimas componentes indican el valor de la variable correspondiente en el k -ésimo individuo, podemos definir la medida de similitud del Coseno como:

$$s(z_i, z_j) = \frac{z_i^T z_j}{\|z_i\| \|z_j\|} \tag{2.2.2}$$

expresión correspondiente al producto interno normalizado ([Singhal \(2001\)](#)) y siendo $\|\cdot\|$ la norma Euclídeana ([Williams \(2012\)](#)).

Coefficiente de Correlación de Pearson

El Coeficiente de Correlación de Pearson se define a través de la ecuación (2.2.3), donde \bar{z}_i y \bar{z}_j denotan los valores promedios para las variables Z_i y Z_j respectivamente ([Edwards \(1976\)](#)). La diferencia fundamental entre las dos medidas de similitud propuestas es que el Coseno se basa en datos originales, tomando en cuenta las desviaciones al origen, mientras que el coeficiente de correlación de Pearson utiliza las desviaciones de los datos centrados en torno a su media. Esta diferencia es crucial, pues la buena representatividad de la media de los datos conduce a la correcta utilización de cualquiera de las dos medidas propuestas.

$$s(z_i, z_j) = \frac{(z_i - \bar{z}_i)^T (z_j - \bar{z}_j)}{\|z_i - \bar{z}_i\| \|z_j - \bar{z}_j\|} \tag{2.2.3}$$

Coefficiente de correlación de Spearman

La forma más usual para medir la correlación entre variables es el utilizar el coeficiente de correlación de Pearson, pero para la correcta interpretación de éste, se suele asumir la normalidad de ambas variables, cosa que en raras ocasiones se cumple ([Siegel \(1972\)](#)). Para remediar esta situación, se dispone del coeficiente de correlación de Spearman, que corresponde a un test no-paramétrico, el que puede ser interpretado de la misma forma que el de Pearson y es

menos sensible ante valores extremos, tal y como se observa en la Figura 2.1. Para obtenerlo es necesario calcular los rangos para cada una de las variables y reemplazar las diferencias entre ellos en la ecuación (2.2.4), donde n corresponde al número de pares de casos y r_i son los rangos de la variable Z_i .

$$s(z_i, z_j) = 1 - \frac{6 \sum_{i=1}^n (r_i - r_j)^2}{n^3 - n} \quad (2.2.4)$$

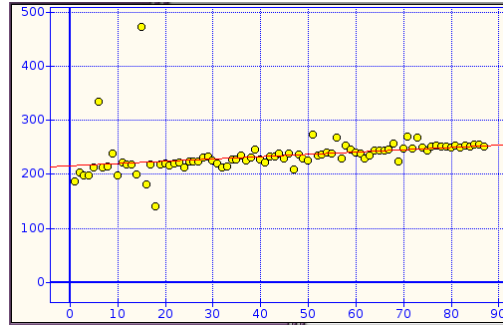


Figura 2.1: Coeficiente de Correlación (Pearson = 0,3071 y Spearman = 0,7627).

2.2.1.2. Medidas de distancia entre individuos

En presencia de atributos de tipo cuantitativo, ya sean continuos o discretos, se dispone de diversas medidas de distancia para cuantificar la proximidad entre los individuos de un conjunto. Para presentar estas medidas, en adelante denotaremos a los dos elementos por vectores de p atributos numéricos (p -dimensionales), de la forma $z_i = (z_{i1}, z_{i2}, \dots, z_{ip})$ y $z_j = (z_{j1}, z_{j2}, \dots, z_{jp})$. La medida de distancia entre ellos puede ser calculada a través de las siguientes métricas.

Distancia Euclideana

Es una de las distancias más utilizadas dada su sencillez. Tiene la ventaja de que la distancia entre dos elementos no está afectada por la adición de nuevos elementos al conjunto (Seber (2009)). Sin embargo, para utilizarla y obtener resultados fiables es necesario realizar un tratamiento de estandarización previo de los atributos que componen los vectores, pues las unidades en las que se miden los atributos son de gran importancia.

$$d(z_i, z_j) = \sqrt{(z_{i1} - z_{j1})^2 + (z_{i2} - z_{j2})^2 + \dots + (z_{ip} - z_{jp})^2} \quad (2.2.5)$$

Distancia Manhattan o City-Block

Es el promedio de la diferencia absoluta entre atributos (Deza y Deza (2009)). Casi siempre al emplear esta distancia obtenemos los mismos resultados que con la distancia Euclideana,

pero las diferencias mayores se disimulan porque no están elevadas al cuadrado:

$$d(z_i, z_j) = |z_{i1} - z_{j1}| + |z_{i2} - z_{j2}| + \dots + |z_{ip} - z_{jp}| \quad (2.2.6)$$

Distancia de Chebychev

La distancia entre los individuos se obtiene como la maxima diferencia absoluta de los atributos medidos sobre los individuos ([Cantrell \(2000\)](#)).

$$d(z_i, z_j) = \max_p \{|z_{i1} - z_{j1}| + |z_{i2} - z_{j2}| + \dots + |z_{ip} - z_{jp}|\} \quad (2.2.7)$$

Distancia de Minkowski

$$d(z_i, z_j) = \{|z_{i1} - z_{j1}|^g + |z_{i2} - z_{j2}|^g + \dots + |z_{ip} - z_{jp}|^g\}^{1/g} \quad (2.2.8)$$

con $g \in \mathbb{N}$. Es posible asignar un peso a cada atributo según su grado de importancia, por lo que es posible obtener una nueva distancia ponderada de la forma:

$$d(z_i, z_j) = \{w_1|z_{i1} - z_{j1}|^g + w_2|z_{i2} - z_{j2}|^g + \dots + w_p|z_{ip} - z_{jp}|^g\}^{1/g} \quad (2.2.9)$$

Esta distancia es la que reúne los casos anteriores, para valores de $g = 2$ se obtiene la distancia Euclideana, para valores de $g = 1$, se obtiene la distancia de Manhattan y para valores de $g = \infty$ se obtiene una aproximación de la distancia de Chebychev ([Kruskal \(1964\)](#)).

Distancia de Mahalanobis

$$d(z_i, z_j) = \{(z_i - z_j)\Sigma^{-1}(z_i - z_j)^T\} \quad (2.2.10)$$

Se diferencia de la distancia Euclideana en que tiene en cuenta la correlación entre los atributos involucrados. Aquí, Σ corresponde a la matriz de varianzas-covarianzas entre los atributos ([Mahalanobis \(1936\)](#); [Mardia y Kent \(1979\)](#)).

2.2.1.3. Medidas de distancia entre individuos con atributos discretos binarios

Para atributos de tipo binario, por ejemplo, presencia (1) o ausencia (0) de alguna característica, se definen medidas de distancia utilizando tablas de contingencia. Al igual que en el caso de las distancias para atributos numéricos, la distancia para atributos binarios debe cumplir con la propiedad de simetría, por lo que diremos que se tiene simetría si ambos de los estados del atributo son igualmente probables. En tal caso, usando los coeficientes provenientes de

una tabla de contingencia, obtendremos una medida de distancia entre dos vectores de la forma

$$d(z_i, z_j) = \frac{r + s}{q + r + s + t} \quad (2.2.11)$$

donde q es el número de atributos clasificados como 1's para ambos vectores, t es el número de atributos clasificados como 0's para ambos vectores; y s y r corresponden al número de atributos clasificados de forma distinta.

Un atributo binario no es simétrico si sus estados (1 ó 0) correspondientes no son igualmente probables. En este caso, el denominador de la expresión anterior no toma en consideración los atributos nulos para ambos vectores. Este nuevo coeficiente es conocido como Coeficiente de Jaccard ([Jaccard \(1900\)](#); [Jaccard \(1901\)](#); [Jaccard \(1908\)](#)):

$$d(z_i, z_j) = \frac{r + s}{q + r + s} \quad (2.2.12)$$

2.2.1.4. Medidas de distancia entre individuos con atributos cualitativos

Medidas de distancia para atributos nominales

Cuando los atributos son de tipo nominal, existen dos procedimientos ampliamente utilizados

1. Concordancia simple

$$d(z_i, z_j) = \frac{p - m}{p} \quad (2.2.13)$$

donde p es el número total de atributos y m es el número de concordancias.

2. Crear un atributo binario para cada nivel de los atributos nominales y calcular su distancia como ya se planteó previamente en el caso de datos binarios.

Medidas de distancia para atributos de tipo mixto

En el caso donde los objetos contienen atributos de distinto origen, podemos utilizar una medida de distancia para estos casos ([Anderberg \(1973\)](#)). Esta medida $d(z_i, z_j)$, que contiene p atributos de diferentes tipos se define como

$$d(z_i, z_j) = \frac{\sum_{n=1}^p \delta_{ij}^{(n)} d_{ij}^{(n)}}{\sum_{n=1}^p \delta_{ij}^{(n)}} \quad (2.2.14)$$

donde la indicadora $\delta_{ij}^{(n)} = 0$ si alguno de los valores no se encuentra. La contribución que hace el atributo n a la distancia entre dos vectores $d_{ij}^{(n)}$ es obtenida de acuerdo a las siguientes posibilidades:

- Si el atributo es de tipo binario o categórico, $d_{ij}^{(n)} = 0$ si $z_{in} = z_{jn}$, en otro caso, $d_{ij}^{(n)} = 1$.

- Si el atributo es de tipo continuo, $d_{ij} = \frac{|z_{in} - z_{jn}|}{\max_h(z_{hn}) - \min_h(z_{hn})}$, donde h toma todo el rango de valores del atributo n .

2.2.2. Algoritmos de agrupamiento

El aspecto crucial en esta etapa es la selección del algoritmo de agrupamiento, que permitirá, en base a la matriz de similitudes obtenida en la etapa anterior, la formación secuencial de los conglomerados. Este paso dista mucho de ser sencillo, pues existe una alta gama de algoritmos para este propósito, cada uno con ventajas específicas y campos de validez ([Estivill-Castro y Yang \(2000\)](#)). Sin embargo, todos los métodos existentes están enfocados en maximizar las diferencias entre conglomerados, y minimizarla dentro de los conglomerados. En cuanto a los algoritmos, podemos clasificarlos en dos grandes agrupaciones, los jerárquicos y los no jerárquicos.

2.2.2.1. Algoritmos Jerárquicos

Los algoritmos jerárquicos consisten en la construcción de una estructura en forma de árbol que recibe el nombre de Dendrograma (Figura 2.2), en el cual se puede seguir de forma gráfica el procedimiento de unión de los conglomerados. La formación de los conglomerados involucra fases sucesivas de agrupaciones o disociaciones, de tal forma que se minimice o maximice alguna medida de proximidad. Los algoritmos jerárquicos pueden subdividirse en aglomerativos y disociativos, donde cada una de estas categorías presentan una gran cantidad de posibilidades. La manera de formar los conglomerados, ya sea por algoritmos aglomerativos o disociativos es relativamente sencilla y siguen un orden general. En el caso de los aglomerativos, se inicia con tantos conglomerados como individuos haya en la muestra. Luego, se selecciona una medida de proximidad, agrupando primero los conglomerados con mayor proximidad entre ellos. De esta forma se prosigue con el algoritmo hasta formar un solo grupo que contenga a todos los individuos, hasta cuando se forme un número a priori de conglomerados y se tengan argumentos basados en la homogeneidad de los conglomerados para no proseguir con el algoritmo ([Jain y cols. \(1999\)](#)). En el caso de los disociativos, se parte con un conglomerado que contiene a todos los individuos de la muestra y en base a medidas de proximidad se va desagrupando hasta cumplir con algún criterio de parada para el algoritmo de desagrupamiento elegido.

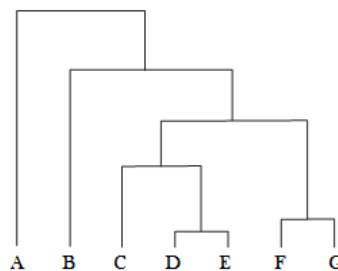


Figura 2.2: Dendrograma

Algoritmos Jerárquicos Aglomerativos

Para estos algoritmos existen variantes a las que denominamos procedimientos y que llevan a resultados distintos, por lo que no existe una elección única que lleve a los resultados correctos. De esta forma, la elección de cada uno de los procedimientos se debe basar en la información disponible y el conocimiento del investigador.

- Procedimiento de la mínima distancia o similitud máxima

Este procedimiento también conocido como “amalgamiento simple” o “simple linkage”, basa la construcción en que la distancia o similitud entre conglomerados corresponde a la mínima distancia (o máxima similitud) entre sus componentes, tal y como se aprecia en la Figura 2.3 (Sneath y cols. (1973)).

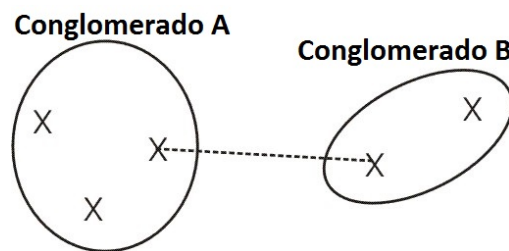


Figura 2.3: Procedimiento de la mínima distancia

- Procedimiento de la máxima distancia o similitud mínima

Conocido como “amalgamiento completo” o “complete linkage”, considera que la distancia entre dos conglomerados corresponde a la máxima distancia o mínima similitud entre cualquier par de sus componentes, tal y como se aprecia en la Figura 2.4 (King (1967)).

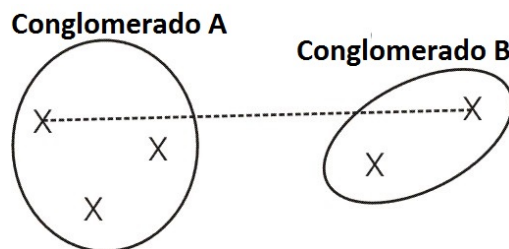


Figura 2.4: Procedimiento de la máxima distancia

- Procedimiento de la distancia o similitud promedio

En esta estrategia la distancia o similitud entre los conglomerados se obtiene como la media aritmética entre la distancia o similitud de las componentes de dichos conglomerados (Figura 2.5). Cabe destacar que no se tiene en cuenta el tamaño de ninguno de los conglomerados involucrados, por lo que entrega el mismo grado de importancia a los conglomerados, sin importar que sean de dimensiones muy distintas (Ward Jr (1963); Murtagh (1983)).

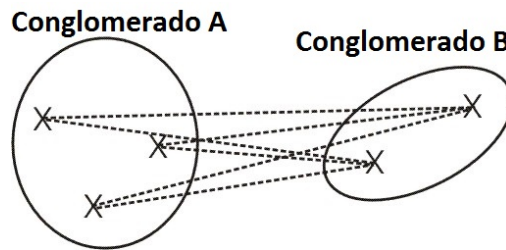


Figura 2.5: Procedimiento de la distancia promedio

- Procedimiento de la distancia o similitud promedio ponderada

Este método consiste en el mismo cálculo de las distancias o similitudes promedio, pero a diferencia del método anterior, pondera estas distancias por el número de componentes de cada nuevo conglomerado a formar (Figura 2.6).

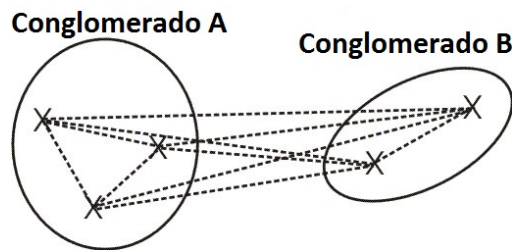


Figura 2.6: Procedimiento de la distancia promedio ponderada

- Procedimiento basado en el centroide

La distancia o similitud entre dos conglomerados viene dada por la distancia o similitud entre sus centroides, que corresponde al vector de medias de las variables que son medidas sobre las componentes de los conglomerados. Para este método, se dispone de dos opciones:

- Centroide ponderado: en el que los tamaños de los conglomerados son considerados en las etapas de construcción. Esto se ejemplifica con la Figura 2.7, donde el centroide del clúster con mayor número de elementos (Clúster A) tiene un peso mayor que el que tiene un número menor de elementos (Clúster B).

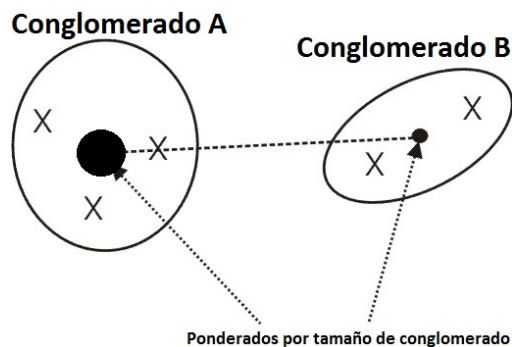


Figura 2.7: Procedimiento basado en el centroide ponderado

- Centroides no ponderado: los tamaños de los conglomerados no son considerados. En la Figura 2.8 los dos centroides poseen el mismo peso, independiente de la cantidad de elementos que conforman los clústers.

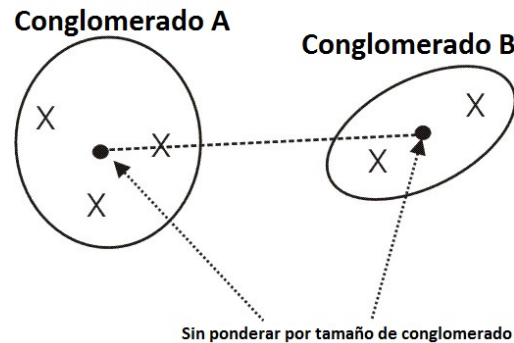


Figura 2.8: Procedimiento basado en el centroide no ponderado

- Procedimiento de Ward

Considera que la distancia o similitud entre dos conglomerados es la suma de las desviaciones al cuadrado entre los individuos y las medias de sus conglomerados respectivos (Wolfe (1967)). En cada etapa de construcción de conglomerados se busca minimizar esta cantidad, también llamada “Suma de Cuadrados”, para todas las particiones obtenidas mediante la combinación de dos conglomerados en un paso previo; es decir, se unirán dos conglomerados mientras ellos posean el menor incremento en el valor de la Suma de Cuadrados dentro de cada conglomerado y dentro de cada componente al centroide del conglomerado. El procedimiento de Ward (definido e implementado recursivamente a través del algoritmo de Lance-Williams (Lance y Williams (1967)) es uno de los más utilizados en la práctica; posee casi todas las ventajas del procedimiento *average linkage* (Gordon (1999)) y suele ser más discriminativo en la determinación de los niveles de agrupación. Incluso una investigación llevada a cabo por Kuiper (Kuiper y Fisher (1975)), probó que este procedimiento era capaz de acertar mejor con la clasificación óptima que otros métodos (distancia mínima, máxima, promedio y basado en el centroide). Cada uno de estos procedimientos presentan ventajas y desventajas dependiendo del tipo de estudio que se esté realizando (Guha y cols. (1998)). Por ejemplo, en el caso del procedimiento de la mínima distancia se presenta el problema denominado “efecto cadena”, donde unas pocas componentes pertenecientes a dos conglomerados muy diferentes, son unidas en etapas de aglomeración, causando un aumento en la heterogeneidad del conglomerado final. El procedimiento de la máxima distancia, tiene la ventaja de producir conglomerados compactos y jerarquías generalmente mucho más útiles que el procedimiento de la mínima distancia, no obstante, este último resulta ser un procedimiento mucho más versátil. Además, se tiene que el procedimiento de la máxima distancia es mucho menos sensible a “datos aberrantes”² que el procedimiento de la mínima distancia. En el caso del procedimiento de Ward, se obtienen conglomerados mucho más compactos y de tamaños similares (Figura 2.9).

²Se define un dato aberrante como a un valor anómalo, extremo o alejado de la distribución esperada de una muestra

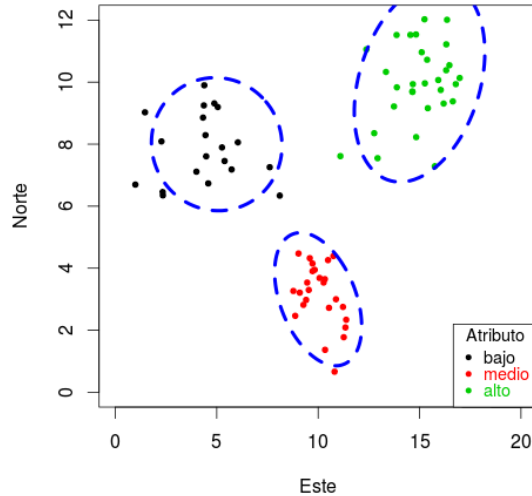


Figura 2.9: Ilustración del Método de Ward

Algoritmos Jerárquicos Disociativos

En el caso de los algoritmos disociativos, a diferencia de los aglomerativos, se parte con un conglomerado que contiene a todos los individuos o componentes del estudio y a partir de ahí, se comienzan a formar conglomerados en etapas sucesivas a través de particiones hasta conseguir tantos conglomerados como componentes existan. Estos algoritmos son poco utilizados dada la escasa documentación existente, pero en cuanto a la estrategia a seguir, se mantiene la misma idea que en el caso de los aglomerativos. Se busca maximizar las distancias o minimizar las similitudes entre las componentes del conglomerado inicial. Estos algoritmos pueden clasificarse en dos ramas, una, los Monotéticos, donde se dividen los conglomerados sobre la base de un solo atributo o variable; y los Politéticos, cuyas divisiones se basan en los valores tomados por todas las variables.

La pregunta que surge al utilizar estos algoritmos es ¿cuándo un conglomerado debe dejar de dividirse para dar paso a otro? Respuestas a esta pregunta son documentadas por P. MacNaughton-Smith ([Macnaughton-Smith y cols. \(1964\)](#)).

Coficiente de correlación cofenético

En vista que los métodos jerárquicos imponen una estructura definida en los componentes, es necesario verificar si tal estructura es aceptable o no para representar las relaciones originales entre los datos. Para verificar si se mantienen las relaciones originales con esta nueva estructura se utiliza la matriz de correlación cofenética, la cual es simplemente la correlación entre los $\frac{n(n-1)}{2}$ elementos de la parte superior de la matriz de distancias observada y los correspondientes en la llamada matriz cofenética, C , cuyos elementos, c_{ij} , se definen como aquellos que determinan la distancia entre los elementos i y j cuando éstos se unen en el mismo conglomerado. Así, si tras el empleo de varios procedimientos de conglomerados dis-

tintos, éstos conducen a soluciones parecidas, surge la pregunta de qué método elegiremos como definitivo. La respuesta la da el coeficiente cofenético, ya que aquel método que tenga un coeficiente cofenético más elevado será aquel que presente una menor distorsión en las relaciones originales existentes entre los elementos en estudio.

2.2.2.2. Algoritmos de particionamiento

El concepto fundamental de los algoritmos de particionamiento es que se especifica a priori el número de conglomerados a formar. El algoritmo parte con una solución inicial, la que es actualizada en base a algún criterio óptimo, el que intercambia los elementos³ pertenecientes a los grupos iniciales, para obtener nuevos conglomerados hasta lograr una solución final. La diferencia de todos los algoritmos de particionamiento radica en la forma de cómo se define lo que es una mejor partición y la forma de cómo se obtiene. Como el proceso parte con un número conocido a priori de conglomerados, es necesario contar con una partición inicial del conjunto de elementos o en su defecto con un número inicial de centroides en torno a los cuales se construirán los conglomerados iniciales.

Elección de centroides iniciales

Estos centroides corresponden a los núcleos de los conglomerados iniciales y existen distintos procedimientos, todos ellos subjetivos, para seleccionar tales puntos del total de elementos.

- Elección de los primeros k elementos de la base de datos. Este es el más simple de los métodos para elección de centroides iniciales y tiene validez sólo en el caso en que el orden de los elementos no esté relacionado con alguna estructura inherente a los datos (MacQueen y cols. (1967)).
- Enumerar los casos desde 1 hasta r y escoger los k números de forma aleatoria (McRae (1971)), también denominado como inicialización aleatoria.
- Tomar una partición de casos en k grupos mutuamente excluyentes y usar los centroides como puntos iniciales (Forgy (1965)).
- Tomar el vector de media de los datos como el primer punto inicial y luego tomar cualquier punto que esté a una distancia por lo menos d de éste. Seguir con los puntos siguientes usando el mismo criterio hasta completar los k puntos requeridos (Ball y Hall (1967)).

Elección de particiones iniciales

El interés se centra en encontrar una partición del conjunto de datos inicial más que determinar un conjunto de centroides iniciales. Algunos procedimientos corresponden a:

³Los elementos o individuos corresponden a vectores de dimensión dada por el número de atributos medidos

- Si se dispone de un conjunto de centroides iniciales, entonces sobre cada uno de ellos se construyen las particiones iniciales incluyendo las componentes más próximas a cada partición. Los centroides se mantienen fijos durante el proceso, de modo que la partición resultante es independiente de la secuencia de introducción a cada partición (Forgy (1965)).
- Dado un conjunto de centroides iniciales, éstos son tomados como conglomerados iniciales de tamaño uno. Luego, se asigna cada elemento a alguno de los conglomerados actualizando el centroide y prosiguiendo con las fases sucesivas. Este procedimiento hace referencia al método de conglomerados jerárquicos y asegura que la partición formada es independiente del orden en el que los elementos fueron asignados.
- Emplear cualquier otro método de conglomerados jerárquicos para construir la partición. Se destaca el método de Ward por sus propiedades antes descritas (Wolfe (1967)).

Algoritmo K-Medias

Se denomina K-Medias al procedimiento que busca asignar cada elemento o individuo al conglomerado con el centroide más cercano (MacQueen y cols. (1967)). La clave de este procedimiento es que el centroide es actualizado luego de cada asignación. Se presenta el pseudo código para implementar este algoritmo en la Figura 2.10. El algoritmo inicia con el conjunto de centroides de cada conglomerado escogidos al azar o por alguna designación específica. En cada iteración, los elementos o individuos son asignados al centroide más cercano de acuerdo a la distancia Euclideana. Luego, el centroide es recalculado. El centro de cada conglomerado se obtiene como la media de todos los elementos o individuos que lo componen. Es necesario definir condiciones de convergencia para establecer un alto en el algoritmo. Por ejemplo, se puede detener el algoritmo si la varianza intra-grupal no cambia significativamente en etapas sucesivas o definir un número máximo de iteraciones. Es posible obtener mucha más información acerca de la convergencia del algoritmo K-Medias en el trabajo de Shokri Selim y Mohamed Ismail (Selim y Ismail (1984)). Los métodos propuestos son sencillos y bastante populares dada la flexibilidad y los diferentes resultados que entregan (Dhillon y Modha (2001)). No obstante, generan muchas dudas en cuanto a las etapas de selección de configuración inicial, en cuanto a los criterios de convergencia utilizando uno u otro método, el orden en el que se ingresan los individuos a cada conglomerado, cuál método converge y asegura una convergencia en menos cantidad de pasos y si el número de individuos o atributos afecta al número de iteraciones necesarias para obtener convergencia.

Entrada: Individuos en base de datos y número de conglomerados a formar

Salida: Conglomerados

1. Inicializar los k conglomerados
2. **Mientras que** no se cumplan las condiciones **Hacer**
 - a. Asignar los individuos al conglomerado más cercano
 - b. Actualizar el centroide del conglomerado
3. Finalizar algoritmo

Figura 2.10: Algoritmo K-Medias

2.2.3. Criterios para testear la bondad de ajuste

De la misma forma en que un análisis predictivo puede ser evaluado en términos de la precisión de las predicciones, es de interés testear la bondad del ajuste realizado por los métodos de conglomerados. Sin embargo, el hecho de verificar si un ajuste es bueno o no, es un problema actualmente sin solución debido a la inexistencia de un criterio universal que diga cuán bueno fue el ajuste (Baker y Hubert (1976)). Como mencionamos al principio, los métodos de conglomerados corresponden a herramientas de aprendizaje no supervisado y descubrimiento de nuevas categorías para el conjunto de elementos, por lo que al desconocer las categorías reales no es posible cuantificar la exactitud de la solución obtenida.

En términos prácticos, la bondad de ajuste de la solución obtenida queda a criterio de los investigadores, la que pasa por el concepto de asociar conocimientos previos y en base a ellos validar la solución obtenida. No obstante, diversos criterios han sido desarrollados para evaluar este concepto, los que pueden ser divididos en dos categorías: Validez Interna y Externa.

2.2.3.1. Criterios de Validez Interna

Los criterios de validez interna están relacionados con el concepto de compacidad de los conglomerados formados, midiendo esta característica con alguna función específica. Usualmente, estas funciones incorporan medidas de homogeneidad de los elementos dentro de un mismo conglomerado y medidas de separabilidad entre elementos pertenecientes a conglomerados diferentes. En general, se utiliza la información contenida en el conjunto de datos para la construcción de estas funciones, dejando de lado cualquier tipo de conocimiento extra con la que muchas veces se dispone. Las sumas de cuadrados del error (SSE) corresponden al criterio de validez interna más utilizado para cuantificar la bondad de ajuste en los métodos de conglomerados. Se destacan por su sencillez y porque utilizan los conceptos de homogeneidad y separabilidad antes mencionados. Corresponden a un valor numérico, el que es obtenido como sigue

$$SSE = \sum_{k=1}^K \sum_{z_i \in C_k} \|z_i - \mu_k\|^2 \quad (2.2.15)$$

donde C_k denota al conjunto que reúne los elementos del conglomerado k -ésimo, μ_k corresponde al vector que contiene los valores medios de los atributos de los elementos pertenecientes a C_k . Las componentes de μ_k son obtenidas de la forma

$$\mu_{k,j} = \frac{1}{N_k} \sum_{z_i \in C_k} z_{i,j} \quad (2.2.16)$$

donde $N_k = |C_k|$ corresponde al número de elementos pertenecientes al conglomerado C_k . Los métodos de conglomerados usualmente tratan de minimizar esta función para lograr el concepto de compacidad, la que a través de álgebra simple, puede ser reescrita de la siguiente forma

$$SSE = \frac{1}{2} \sum_{k=1}^K N_k \bar{S}_k \quad \bar{S}_k = \frac{1}{N_k^2} \sum_{z_i, z_j \in C_k} \|z_i - z_j\|^2 \quad (2.2.17)$$

razón por la cual, el criterio de Sumas de Cuadrados del Error también es conocido como Criterio de Mínima Varianza. Es ampliamente utilizado en muchos estudios, pero es necesario mencionar ciertas restricciones al aplicarlo, pues, es efectivo sólo cuando existe una clara separación de un conglomerado a otro en el espacio característico. En casos excepcionales, donde no se cuente con esta propiedad, no se aconseja su uso.

2.2.3.2. Criterios de Validez Externa

Se puede realizar una validación comparando los resultados obtenidos con un criterio externo (por ejemplo, clasificaciones obtenidas por evaluaciones independientes o analizando en los grupos obtenidos, el comportamiento de variables no utilizadas en el proceso de clasificación) o realizando un análisis de conglomerados con una muestra diferente de la realizada. Entre ellos destacamos la Medida Basada en la Mútua Información ([Strehl y Ghosh \(2000\)](#)), el Índice de Rand ([Rand \(1971\)](#); [Hubert y Arabie \(1985\)](#)) y los criterios relativos como el Índice de Calinski-Harabasz ([Caliński y Harabasz \(1974\)](#)).

2.3. Modelos de Mezclas de Distribuciones

2.3.1. Mezcla de Distribuciones Normales

2.3.1.1. Reseña histórica

Las mezclas de distribuciones fueron estudiadas por primera vez en 1894 por Karl Pearson, analizando un conjunto de 1000 cangrejos (la razón del ancho de frente sobre la longitud de cuerpo), cuyos datos fueron proporcionados a principios de 1890 por el profesor W. R. Weldon (1860-1906). Observando el histograma (Figura 2.11) para ver su distribución se pudo observar que existía sesgo a la derecha. Weldon ([Weldon \(1889\)](#)) sugirió que la razón para esta asimetría podría ser que la muestra tenía representantes de dos tipos de cangrejo, pero cuando los datos fueron recolectados no habían sido diferenciados como tal. Esto hizo que Pearson propusiera que la distribución de las medidas podrían ser modeladas por la suma de los productos entre la proporción del tipo de cangrejo y su distribución normal, con los dos pesos dados a la proporción de cangrejos de cada tipo. Esto es conocido hoy en día como el concepto de mezcla finita de distribuciones, la que puede expresarse de la siguiente manera:

$$f(x) = p N(\mu_1, \sigma_1^2) + (1 - p) N(\mu_2, \sigma_2^2) \quad (2.3.1)$$

donde p es la proporción del tipo de cangrejo para la cual la razón (frente/longitud del cuerpo) tiene media μ_1 y varianza σ_1^2 , y $(1 - p)$ es la proporción del tipo de cangrejo para que los valores correspondientes sean μ_2 y σ_2^2 . El estudio para saber la manera en la que se ajustarían los datos empezó con Pearson ([Pearson \(1894\)](#)), utilizando el método de momentos para estimar los cinco parámetros de dos componentes de una mezcla normal univariada.

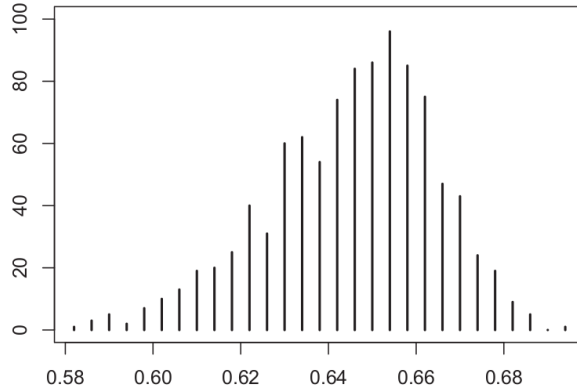


Figura 2.11: Histograma de las razones de la frente sobre la longitud del cuerpo de los 1000 cangrejos

2.3.1.2. Función de densidad para una mezcla

Si definimos a X como a una variable aleatoria definida sobre una población que contienen K grupos homogéneos, su distribución puede ser escrita a través de su densidad de la siguiente forma:

$$G(\mathbf{X}) = G_1(\mathbf{X}) + \dots + G_K(\mathbf{X}) = \sum_{k=1}^K G_k(\mathbf{X}) \quad (2.3.2)$$

donde $G_k(\mathbf{X}) = \pi_k f_k(\mathbf{X})$, $0 \leq \pi_k \leq 1$ y $\sum_{k=1}^K \pi_k = 1$ para $k \in \{1, \dots, K\}$. Diremos que \mathbf{X} tiene una mezcla finita de distribuciones y que $G(\cdot)$ es la función de densidad de la mezcla o mezcla de densidades de toda la población. Además $G_1(\cdot), \dots, G_K(\cdot)$ son llamadas componentes de la mezcla, los parámetros π_1, \dots, π_K son llamados pesos de las componentes de la mezcla o proporciones de las componentes de la mezcla y $f_1(\cdot), \dots, f_K(\cdot)$ densidades de las componentes de la mezcla. Como ejemplo, para una mezcla de densidades normales univariadas se tendrá

$$G(\mathbf{x}) = \pi_1 f_1(\mathbf{x}|\mu_1, \sigma_1^2) + \dots + \pi_K f_K(\mathbf{x}|\mu_K, \sigma_K^2) \quad (2.3.3)$$

En el caso multivariante, en vez de trabajar con los parámetros μ_k y σ_k^2 , asumiremos una distribución normal multivariante, la que está controlada por el vector de medias $\mu = (\mu_1, \dots, \mu_K)$ y la matriz de varianzas-covarianzas $\Sigma = (\Sigma_1, \dots, \Sigma_K)$.

2.3.2. Parámetros en la mezcla de densidades normales

Los parámetros de la mezcla de densidades normales p-variantes corresponden a (μ, Σ) , donde llamaremos μ al vector de medias de la mezcla y Σ a la matriz de varianzas-covarianzas de la distribución mezclada; estos se obtienen fácilmente conocidas los vectores de medias μ_k y las matrices de varianzas Σ_k , con $k = 1, \dots, K$, de la densidades componentes de la mezcla.

2.3.2.1. Vector de medias

Vamos a definir $\mu(K)$ como el vector de medias p -dimensional de la distribución mezclada, donde una componente de este vector $\mu(K)_j$, con $j = 1, \dots, p$, es la media ponderada de la mezcla en la j -ésima variable; esta se obtiene sumando, para los K -grupos, los productos entre las proporciones de mezcla π_k (con $k = 1, \dots, K$) y las j -ésimas componentes del vector de medias de cada grupo μ_{jk} , es decir:

$$\mu(K) = \begin{pmatrix} \mu(K)_1 \\ \vdots \\ \mu(K)_p \end{pmatrix} = \sum_{k=1}^K \pi_k \begin{pmatrix} \mu_{1k} \\ \vdots \\ \mu_{pk} \end{pmatrix} = \sum_{k=1}^K \begin{pmatrix} \pi_k \mu_{1k} \\ \vdots \\ \pi_k \mu_{pk} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^K \pi_k \mu_{1k} \\ \vdots \\ \sum_{k=1}^K \pi_k \mu_{pk} \end{pmatrix} \quad (2.3.4)$$

2.3.2.2. Matriz de varianzas-covarianzas

La matriz de varianzas-covarianzas de la distribución mezclada con densidades componentes normales viene dada por

$$V(K) = \sum_{k=1}^K \pi_k V_k + \sum_{k=1}^K \pi_k (\mu_k - \mu(K))(\mu_k - \mu(K))' \quad (2.3.5)$$

Luego la variabilidad total, que es la matriz de varianzas-covarianzas $V(K)$, se descompone en una variabilidad explicada, $\sum_{k=1}^K \pi_k (\mu_k - \mu(K))(\mu_k - \mu(K))'$ que tiene en cuenta las diferencias entre medias de las densidades de las componentes μ_k y el vector de medias $\mu(K)$, y una variabilidad no explicada, $\sum_{k=1}^K \pi_k V_k$, que es la variabilidad con respecto a las componentes.

2.3.3. Estimación Máximo Verosímil de parámetros en las componentes de la mezcla

Como se puede apreciar en la sección anterior, los parámetros de una mezcla de densidades normales quedan totalmente determinados si conocemos los parámetros de cada una de las componentes que forman la mezcla; esto es para nuestro caso conocer la terna (π_k, μ_k, V_k) , que llamaremos θ_k , ya que está formada por los parámetros del grupo k para $k = 1, \dots, K$. Dificilmente en la práctica contamos con esta información, por lo que se hace necesario estimar los valores del peso π_k , el vector de medias μ_k y la matriz de varianzas-covarianzas V_k , para $k = 1, \dots, K$. Estimaremos estos parámetros por el método de máxima verosimilitud (MV).

2.3.3.1. Verosimilitud, soporte y puntaje para mezclas

Sea x_1, \dots, x_n una muestra de datos independientes que pueden estratificarse en K grupos, de manera que existe n_1 observaciones del grupo 1, n_2 observaciones del grupo 2, ..., n_K del

grupo K . El vector $x_i = (x_{i1}, \dots, x_{ip})'$ representa un individuo particular para $i = 1, \dots, n$. Definimos la función de verosimilitud para la mezcla, $l(\theta)$, como:

$$l(\theta) = \prod_{i=1}^n G(x_i|\theta) = \prod_{i=1}^n \left(\sum_{k=1}^K \pi_k f_k(x_i) \right) \quad (2.3.6)$$

donde $\theta = (\theta_1, \dots, \theta_K)$ es el vector de parámetros para los K grupos, $G(x_i|\theta)$ es valor de la mezcla de densidades para el i -ésimo individuo dado el vector de parámetros θ , π_k es el peso de la mezcla para el k -ésimo grupo y $f_k(x_i)$ es el valor de la densidad en el k -ésimo grupo para el i -ésimo individuo.

Sea $l(\theta)$ la función de verosimilitud de una mezcla de K densidades, con $\theta = (\theta_1, \dots, \theta_K)'$. La función soporte para la mezcla se define como:

$$L(\theta) = \sum_{i=1}^n \ln(G(x_i|\theta)) = \sum_{i=1}^n \ln \left(\sum_{k=1}^K \pi_k f_k(x_i) \right). \quad (2.3.7)$$

mientras que la función de puntaje para la mezcla se define como:

$$Z(\theta) = \sum_{i=1}^n \frac{\frac{\partial}{\partial \theta} [G(x_i|\theta)]}{G(x_i|\theta)} = \sum_{i=1}^n \frac{\frac{\partial}{\partial \theta} \left[\sum_{k=1}^K \pi_k f_k(x_i) \right]}{\sum_{k=1}^K \pi_k f_k(x_i)} \quad (2.3.8)$$

2.3.3.2. Ecuaciones de máxima verosimilitud para mezclas de densidades normales

En nuestro caso, asumimos cada $f_k(\mathbf{X})$ como una densidad normal p -variante con vector de medias μ_k y matriz de varianzas \mathbf{V}_k ; de esta forma $\theta = (\pi_1, \dots, \pi_K; \mu_1, \dots, \mu_K; \mathbf{V}_1, \dots, \mathbf{V}_K)$. Luego nuestra tarea ahora es maximizar la función soporte dada en (2.3.7), que equivale a encontrar solución al sistema homogéneo que se produce en la función de puntaje para mezclas dada en (2.3.8). Para evitar inconvenientes en la maximización supondremos que el orden de las K distribuciones estará determinado por $\pi_1 \geq \pi_2 \geq \dots \geq \pi_K$, pues el orden $1, \dots, K$ dado anteriormente es arbitrario. También supondremos que como mínimo hay p -observaciones en cada distribución y trataremos de encontrar un máximo local que proporcione un estimador consistente de los parámetros. Luego la función de puntaje para mezclas nos proporciona el siguiente sistema homogéneo para el k -ésimo grupo:

$$\frac{\partial L(\theta)}{\partial \pi_k} = 0 \quad (2.3.9)$$

$$\frac{\partial L(\theta)}{\partial \mu_k} = 0 \quad (2.3.10)$$

$$\frac{\partial L(\theta)}{\partial \mathbf{V}_k} = 0 \quad (2.3.11)$$

Las soluciones a la ecuación (2.3.9) queda de la forma:

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^n \pi_{ik} \quad (2.3.12)$$

donde se define a π_{ik} como la probabilidad a posteriori de que la observación i halla sido generada por la población k , esto es

$$\pi_{ik} = \frac{\pi_k f_k(\mathbf{x}_i)}{\sum_{k=1}^K \pi_k f_k(\mathbf{x}_i)}. \quad (2.3.13)$$

Para la ecuación (2.3.10) tenemos:

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{\sum_{i=1}^n \pi_{ik}} \sum_{i=1}^n \pi_{ik} \mathbf{x}_i \quad (2.3.14)$$

y para (2.3.11)

$$\widehat{\mathbf{V}}_k = \frac{1}{\sum_{i=1}^n \pi_{ik}} \sum_{i=1}^n \pi_{ik} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k) (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)'. \quad (2.3.15)$$

Para resolver las ecuaciones (2.3.12), (2.3.14) y (2.3.15) y así obtener los estimadores se necesitan las probabilidades π_{ik} dadas con (2.3.13), pero para esto se necesitan los parámetros del modelo. Entonces tenemos el conflicto de estimar parámetros los cuales necesitamos conocer previamente para estimarlos. Para resolver el problema de estimación de parámetros se utiliza el algoritmo EM ([Dempster y cols. \(1977\)](#)) que consiste en una estimación y maximización iterativa de la verosimilitud de la muestra, como se puede ver en el Algoritmo 1.

Paso inicial

Hacer $k = 1$ y $r = 999$, definir una *tolerancia* y un conjunto de parametros iniciales $\hat{\boldsymbol{\theta}}^{(0)}$

Mientras $r > \textit{tolerancia}$ **hacer**

Paso de estimación

Calcular $\hat{\pi}_{ig}^{(k)} = \frac{\hat{\pi}_g^{(k-1)} f_g(\mathbf{x}_i)}{\sum_{g=1}^G \hat{\pi}_g^{(k-1)} f_g(\mathbf{x}_i)}$ donde $f_g(\mathbf{x}_i) = N_p(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_g^{(k-1)}, \widehat{\mathbf{V}}_g^{(k-1)})$

Paso de maximización

Hallar $\hat{\boldsymbol{\theta}}^{(k)} = (\hat{\pi}_1^{(k)}, \dots, \hat{\pi}_G^{(k)}, \hat{\boldsymbol{\mu}}_1^{(k)}, \dots, \hat{\boldsymbol{\mu}}_G^{(k)}, \widehat{\mathbf{V}}_1^{(k)}, \dots, \widehat{\mathbf{V}}_G^{(k)})$, donde

$$\hat{\pi}_g^{(k)} = \frac{1}{n} \sum_{i=1}^n \hat{\pi}_{ig}^{(k)}$$

$$\hat{\boldsymbol{\mu}}_g^{(k)} = \frac{1}{\sum_{i=1}^n \hat{\pi}_{ig}^{(k)}} \sum_{i=1}^n \hat{\pi}_{ig}^{(k)} (\mathbf{x}_i)$$

$$\widehat{\mathbf{V}}_g^{(k)} = \frac{1}{\sum_{i=1}^n \hat{\pi}_{ig}^{(k)}} \sum_{i=1}^n \hat{\pi}_{ig}^{(k)} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g^{(k)}) (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g^{(k)})'$$

$$r = \|\hat{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}}^{(k-1)}\|$$

Hacer $k = k + 1$

Finalizar

Algoritmo 1: Algoritmo EM para mezclas de densidades normales

2.4. Estadística Bayesiana

2.4.1. Introducción

Muchos problemas estadísticos involucran conjuntos de observaciones que pueden considerarse relacionadas entre sí debido a la estructura del problema. Por ejemplo, en un estudio sobre la efectividad de un tratamiento renal, si los pacientes del hospital i -ésimo tienen una probabilidad de supervivencia $P(X_i = x) = \theta_i$, entonces parece razonable suponer que tales probabilidades (las cuales representan una muestra de hospitales) están relacionadas entre dos o más hospitales. Una manera natural de modelar esta situación consiste en usar una distribución de probabilidad inicial, bajo la cual, las probabilidades de supervivencia de cada hospital pueden verse como una muestra de una distribución poblacional común, que a su vez depende de un hiperparámetro desconocido ϕ . Una característica importante de este enfoque es que los datos observados pueden utilizarse para estimar ciertos aspectos de la distribución poblacional (funciones de ϕ) a pesar de que sus valores no son observables. Esta estructura jerárquica resulta muy útil al analizar modelos con muchos parámetros, como los que se usan para analizar datos que provienen de muestreos estratificados. Por ejemplo, en un estudio para pronosticar la proporción de votantes que favorecerán a un candidato determinado en las próximas elecciones, pueden considerarse factores a nivel nacional, regional, estatal y distrital. Es importante destacar que en la práctica los modelos no jerárquicos no son apropiados para analizar datos con estructura jerárquica. En general, si el modelo tiene pocos parámetros entonces no puede ajustar adecuadamente conjuntos de datos relativamente grandes. Por el contrario, si el modelo tiene muchos parámetros entonces tenderá a “sobreajustar” a los datos, en el sentido de que ajustará bien a los datos observados pero no necesariamente producirá buenas predicciones. En contraste, los modelos jerárquicos pueden tener suficientes parámetros para ajustar bien los datos evitando el problema de sobreajuste al modelar la estructura de dependencia entre los parámetros a través de la distribución poblacional.

2.4.2. Introducción a Estadística Bayesiana

Dentro de los modelos lineales jerárquicos, colocamos especial atención a aquellos que incorporen el principio de incertidumbre dentro de la distribución poblacional para las probabilidades θ_i y además, incorporen la información aportada por los datos observados x_i . Una de las maneras de plantear dicho esquema es mediante una metodología bayesiana, que consta de tres pasos fundamentales:

1. Especificar un modelo de probabilidad que incluya algún tipo de conocimiento previo (a priori) sobre los parámetros del modelo dado.
2. Actualizar el conocimiento sobre los parámetros desconocidos condicionando este modelo de probabilidad a los datos observados.
3. Evaluar el ajuste del modelo a los datos y la sensibilidad de las conclusiones a cambios en los supuestos del modelo.

La diferencia fundamental entre la estadística clásica (frecuentista) y la bayesiana es el concepto de probabilidad. Para la estadística clásica es un concepto *objetivo*, que se encuentra en la naturaleza, mientras que para la estadística bayesiana se encuentra en el *observador*, siendo así un concepto subjetivo. De este modo, en estadística clásica sólo se toma como fuente de información las muestras obtenidas suponiendo, para los desarrollos matemáticos, que se puede tomar tamaños límite de las mismas. En el caso bayesiano, sin embargo, además de la muestra también juega un papel fundamental la información previa o la historia que se posee relativa a los fenómenos que se tratan de modelar. El concepto básico en estadística bayesiana es el de probabilidad condicional: Para dos sucesos A y B ,

$$P(A|B) = \frac{P(A \cup B)}{P(B)}. \quad (2.4.1)$$

Se puede aplicar la definición también a variables discretas o continuas. Desde el punto de vista bayesiano, casi todas las probabilidades son condicionales porque siempre existe algún conocimiento previo o historia de los sucesos. Un concepto importante es el expresado por la ley de la probabilidad total: Para un suceso A y una partición B_1, \dots, B_k ,

$$P(A) = \sum_{i=1}^k P(A|B_i)P(B_i). \quad (2.4.2)$$

Se puede aplicar el teorema a variables discretas:

$$P(x) = \sum_y P(x|Y = y)P(Y = y) \quad (2.4.3)$$

o a variables continuas:

$$f(x) = \int f(x|y)f(y)dy. \quad (2.4.4)$$

2.4.2.1. Intercambiabilidad

Consideremos una serie de experimentos $\{E_i : i = 1, \dots, k\}$, donde E_i produce observaciones x_i cuya distribución depende del parámetro θ_i . De esta manera, cada E_i tiene asociada una función de verosimilitud $p(x_i|\theta_i)$. Cabe mencionar que los métodos descritos en esta sección son también aplicables a datos observacionales que tengan una estructura jerárquica. Si no contamos con información que nos permita distinguir cualquiera de los parámetros θ_i de los otros (aparte de las observaciones x_i), y si no es razonable establecer algún orden o agrupamiento de los parámetros, entonces debemos suponer alguna forma de simetría entre éstos. Dicha simetría debe verse reflejada en la distribución inicial de los parámetros θ_i y se representa probabilísticamente a través del concepto de intercambiabilidad. De esta manera,

$$p(\theta_1, \dots, \theta_k) = \int \prod_{i=1}^k p(\theta_i|\phi)dQ_0(\phi). \quad (2.4.5)$$

Por otro lado, a nivel de las observaciones se tiene

$$\begin{aligned} p(x_1, \dots, x_k) &= \int \prod_{i=1}^k p(x_i|\theta_i)dQ(\theta_1, \dots, \theta_k) \\ &= \int \prod_{i=1}^k \prod_{j=1}^{n_i} p(x_{ij}|\theta_i)dQ(\theta_1, \dots, \theta_k) \end{aligned}$$

donde $dQ_0(\phi) = p(\phi)d\phi$ y $dQ(\theta_1, \dots, \theta_k) = p(\theta_1, \dots, \theta_k)d\theta_1 \dots d\theta_k$. La forma de estas distribuciones como mezclas de distribuciones de observaciones independientes e idénticamente distribuidas, es generalmente todo lo que se necesita en la práctica para capturar la idea de intercambiabilidad. A manera de resumen, podemos decir que un modelo jerárquico tiene la siguiente estructura:

Nivel I. Observaciones

$$\begin{aligned} p(x|\theta) &= p(x_1, \dots, x_k | \theta_1, \dots, \theta_k) \\ &= \prod_{i=1}^k p(x_i | \theta_i). \end{aligned}$$

Nivel II. Parámetros

$$\begin{aligned} p(\theta|\phi) &= p(\theta_1, \dots, \theta_k | \phi) \\ &= \prod_{i=1}^k p(\theta_i | \phi). \end{aligned}$$

Nivel III. Hiperparámetros

$$p(\phi).$$

Este modelo puede interpretarse de la siguiente manera:

- Las observaciones x_1, \dots, x_k provienen de experimentos distintos pero de alguna forma relacionados entre sí (por ejemplo, experimentos realizados en k centros de investigación involucrados en el mismo estudio sobre una afección renal).
- Los parámetros $\theta_1, \dots, \theta_k$ se suponen intercambiables, en vista de la posible relación entre las k muestras observadas (por ejemplo, θ_i podría representar la probabilidad de supervivencia en cada uno de los centros de investigación).
- El parámetro ϕ describe alguna característica relevante de la población (por ejemplo, ϕ podría representar la probabilidad de supervivencia “global” para pacientes que sufren de la afección renal bajo estudio).

Es claro que estas ideas pueden generalizarse para construir modelos jerárquicos con más de tres niveles.

2.4.2.2. Covariables

En ocasiones se cuenta con información adicional en forma de covariables $z = (z_1, \dots, z_r)$, de manera que el conjunto de observaciones está dado por $(x_1, z_1), \dots, (x_k, z_k)$. La forma usual

de modelar la intercambiabilidad de los parámetros en presencia de covariables es a través del concepto de independencia condicional,

$$p(\theta_1, \dots, \theta_k | z_1, \dots, z_k) = \int \prod_{i=1}^k p(\theta_i | \phi, z_{i1}, \dots, z_{ir}) p(\phi | z_{i1}, \dots, z_{ir}) d(\phi). \quad (2.4.6)$$

En esencia, esto supone cierta forma de intercambiabilidad parcial.

2.4.2.3. Análisis

El problema básico al analizar un modelo jerárquico bayesiano es que la distribución inicial de los parámetros no está completamente especificada, sino que depende de un hiperparámetro que a su vez tiene una distribución inicial propia. En otras palabras, el problema consiste en hacer inferencias sobre las características individuales $\theta_1, \dots, \theta_k$, así como sobre la característica poblacional ϕ . La distribución inicial apropiada es entonces

$$p(\theta, \phi) = p(\theta | \phi) p(\phi). \quad (2.4.7)$$

La distribución final correspondiente es

$$\begin{aligned} p(\theta, \phi | x) &\propto p(\theta, \phi) p(x | \theta, \phi) \\ &= p(\theta | \phi) p(\phi) p(x | \theta) \end{aligned}$$

donde la última igualdad se debe a que la distribución de las observaciones sólo depende de θ ; el hiperparámetro ϕ afecta a x sólo a través de θ . Dicho de otra forma, θ y ϕ son condicionalmente independientes dado x . La distribución final puede reescribirse como

$$p(\theta, \phi | x) = p(\theta | \phi, x) p(\phi | x) \quad (2.4.8)$$

donde

$$p(\theta | \phi, x) = \frac{p(\theta | \phi) p(x | \theta)}{p(x | \phi)} \quad (2.4.9)$$

y

$$p(\phi | x) \propto p(x | \phi) p(\phi) \quad (2.4.10)$$

con

$$p(x | \phi) = \int p(x | \theta) p(\theta | \phi) d\theta \quad (2.4.11)$$

Por otra parte, la distribución marginal final de los parámetros está dada por

$$p(\theta | x) \propto \int p(x | \theta) p(\theta | \phi) p(\phi) d\phi. \quad (2.4.12)$$

No siempre es posible calcular estas distribuciones de manera analítica. Afortunadamente, existen técnicas de simulación que permiten calcular las integrales donde el número de parámetros es relativamente grande (Gelfand y Smith (1990)). Específicamente, los métodos de Monte Carlo vía cadenas de Markov (Gilks y cols. (1996)) han demostrado ser muy útiles en el análisis de los modelos jerárquicos.

2.4.2.4. Predicción

Un modelo jerárquico bayesiano está caracterizado por los parámetros $\theta_1, \dots, \theta_k$ y el hiperparámetro ϕ . En general, hay dos distribuciones predictivas que podrían ser de interés: (i) la distribución de una observación futura X_j^* correspondiente a uno de los parámetros θ_j existentes; y (ii) la distribución de una observación X^{**} correspondiente a una futura θ^* que proviene de la misma población que dio lugar a los parámetros θ_i . Como en el caso de la distribución a posteriori, en general estas distribuciones predictivas no pueden encontrarse analíticamente pero pueden ser aproximadas a través de métodos de simulación.

2.4.2.5. Métodos para el cálculo

Los problemas de inferencia paramétrica bayesiana se centran en la obtención de una determinada cantidad de posteriors de interés, que serán funciones de los parámetros del problema. Así, si lo que se desea es la inferencia sobre $\theta = (\theta_1, \theta_2, \dots, \theta_k) \in \Theta$, o sobre cualquier transformación del parámetro, $g(\theta)$, desde el punto de vista bayesiano, la formulación ha de contemplar la información muestral, dada por los datos $x = (x_1, x_2, \dots, x_n)$ con función de verosimilitud $f(x|\theta)$, y la elección de la distribución a priori sobre el parámetro, $\pi(\theta)$, que recoge la opinión subjetiva del experto. La combinación de ambos ingredientes conduce a la obtención, vía Teorema de Bayes, de la distribución a posteriori,

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int_{\Theta} \pi(\theta)f(x|\theta)d\theta} \propto \pi(\theta)f(x|\theta) \quad (2.4.13)$$

bajo la que se pretende calcular la cantidad a posteriori de interés, que puede expresarse como

$$E[g(\theta)|x] = \int_{\Theta} g(\theta)\pi(\theta|x)d\theta. \quad (2.4.14)$$

El caso es que tanto la obtención de la densidad a posteriori, como el cálculo de la cantidad a posteriori, contienen integrales no factibles analíticamente, que se solían resolver mediante diferentes métodos de integración numérica hasta que se empezaron a utilizar los métodos de simulación Monte Carlo de Cadenas de Markov, MCMC ([Gilks y cols. \(1995\)](#)), que permiten simular una cadena que converja a la distribución a posteriori, de manera que cantidades muestrales obtenidas a partir de las realizaciones de dicha cadena converjan a la cantidad a posteriori de interés ([Møller y cols. \(2006\)](#)).

2.4.2.6. Métodos MCMC

Los métodos MCMC permiten la resolución de problemas que no son analíticamente tratables y precisan distintas aproximaciones numéricas para las integrales implicadas, que no tienen en cuenta la naturaleza aleatoria del problema ni el carácter probabilístico de las funciones implicadas. Estos métodos permiten muestrear la distribución a posteriori, aunque ésta sea desconocida, gracias a la construcción de una cadena de Markov cuya distribución estacionaria sea, precisamente, $\pi(\theta|x)$.

Una cadena de Markov es una sucesión de variables aleatorias $\{X_t\}$ tal que X_{t+1} sólo depende

del estado actual X_t , mediante la probabilidad de transición de la cadena, $p(X_{t+1}|X_t)$. Bajo ciertas condiciones, la probabilidad de transición de la cadena no depende del instante t y converge a una distribución estacionaria, ϕ , de manera que si $X_t \rightarrow X$ cuando $t \rightarrow \infty$ con distribución ϕ entonces

$$\bar{g}_N = \frac{1}{N} \sum_{t=1}^N g(X_t) \rightarrow E_\phi[g(X)] \quad (2.4.15)$$

cuando $N \rightarrow \infty$. Se trata, por tanto, de simular una cadena de Markov sobre Θ , $\{\theta^{(t)}\}$, cuya distribución estacionaria sea $\pi(\theta|x)$, lo que permitirá que

$$E[g(\theta)|x] \simeq \frac{1}{N-m} \sum_{t=m+1}^N g(\theta^{(t)}) = \bar{g}_{N-m}. \quad (2.4.16)$$

donde m indica el número de realizaciones de la cadena que se desechan para evitar correlación (muestra “burn in”), y N debe ser suficientemente grande como para garantizar la convergencia (Gilks y cols. (1996); Robert y Casella (1999)). Para diseñar la cadena, $\{\theta^{(t)}\}$, se muestrearán mediante simulación la distribución apropiada, $\pi(\theta|x)$. Cuando las distribuciones utilizadas en la simulación son las condicionadas a posteriori de cada parámetro con respecto al resto (porque tengan forma estándar, a diferencia de la distribución a posteriori) se habla de Muestreo de Gibbs, mientras que el Algoritmo de Metropolis-Hastings (que incluye el muestreo de Gibbs como caso particular) utiliza distribuciones más generales para la simulación de la cadena (Y. Chen y cols. (2014); Tierney (1994)).

2.4.3. Criterios de información para la selección, comparación y análisis de convergencia

Existen diversos criterios para determinar la bondad de ajuste del modelo elegido durante el proceso de modelado vía simulaciones por cadenas de Markov. Para comparar modelos, el criterio más usado es el de la desviación o devianza. La devianza se calcula a partir del logaritmo de la función de verosimilitud o verosimilitud restringida. También se utilizan otros, como el Criterio de Información de Akaike (AIC) (Bozdogan (1987)), el AIC Corregido (AICC) (Hurvich y Tsai (1991)) y el Bayesiano (BIC) (S. Chen y Gopalakrishnan (1998)), así como diversas versiones que surgen a partir de estos criterios. En mayor o menor medida, todos penalizan el logaritmo de la función de verosimilitud por el número de parámetros.

2.4.3.1. Criterio de Información de Akaike (AIC)

El Criterio de Información de Akaike (AIC) es un método para la comparación y selección de modelos. Este método permite determinar con qué eficiencia los modelos se ajustan a una base de datos a través de un valor numérico de referencia. Este enfoque se puede aplicar a modelos jerárquicos como no jerárquicos, y no se sustenta en cuanto a valores de probabilidad o en el concepto de significación estadística. Se define como:

$$\text{AIC} = -2\{\ln(\theta) - k\} \quad (2.4.17)$$

donde k corresponde al número de parámetros libres en el modelo. Los puntajes de AIC se demuestran a menudo como puntuaciones ΔAIC o como diferencias entre el mejor modelo (el que posea un valor de AIC menor) y cada modelo. El modelo que mejor explica los datos con el mínimo número de parámetros es el que presenta más bajo valor de AIC (Molinero (2003); Balzarini y cols. (2004)).

Otra variante del criterio de información de Akaike, es el AICc (el criterio AIC corregido), que tiene en cuenta el tamaño muestral. Esencialmente, penaliza la complejidad del modelo con los conjuntos de datos pequeños. Se define como:

$$\text{AICc} = -2\{\ln(\theta)\} + 2k \frac{n}{n - k - 1} \quad (2.4.18)$$

donde n corresponde al tamaño muestral. A medida que n aumenta, AICc converge a AIC, por lo que no hay nada de malo en utilizar siempre AICc independiente del tamaño de la muestra (Burnham y Anderson (2002)). La lógica que sigue este método no es la de las pruebas de hipótesis. Por tanto, no se debe plantear una hipótesis nula o calcular una probabilidad, y no es necesario decidir acerca de la tendencia del valor $-p$ para determinar su significación estadística. El valor de AIC puede ser negativo o positivo, en dependencia de las unidades en que se expresen los datos, y no se puede interpretar como un valor individual.

2.4.3.2. Criterio de información bayesiano de Schwarz (BIC)

La estadística bayesiana surge del famoso teorema de Bayes, que en esencia permite, en caso de conocer la probabilidad de que ocurra un suceso, modificar su valor cuando se dispone de nueva información (Molinero (2002)). El Criterio de Información de Schwarz o Bayesiano, se denomina de esa forma por basarse en argumentos de la llamada estadística bayesiana. Los métodos bayesianos constituyen una alternativa a la estadística tradicional, que se basa en el contraste de hipótesis. Estos métodos se diferencian en que incorporan información externa al estudio. Con esta información y con los datos observados se estima una distribución de probabilidad para la magnitud efecto que se está investigando (Batanero y Batanero (2008)). La fórmula para el Criterio de Información Bayesiano (BIC) es similar al criterio de Akaike, así como su interpretación:

$$\text{BIC} = -2\{\ln(\theta)\} + 2k \cdot \ln(n) \quad (2.4.19)$$

donde k es el número de parámetros libres del modelo y n es el tamaño muestral. El criterio para elegir el mejor modelo es el mismo que el de Akaike: el que tenga el menor valor de BIC (Calegario y cols. (2005); Carrero (2008)).

2.4.3.3. Análisis de convergencia

Aunque teóricamente la convergencia del algoritmo está garantizada, en la práctica se necesitan usar algunos criterios de decisión. Existen varias posibilidades, pero entre ellas, usaremos el criterio de Gelman (Gelman y Rubin (1992)). El objetivo del análisis bayesiano es obtener estimaciones de $E(\theta|x)$, por lo que usualmente se corren cadenas con un cierto número de

iteraciones. Presentaremos el siguiente ejemplo, para ilustrar el criterio de Gelman, implementado en la librería *coda* del software R (Plummer y cols. (2006)). Corremos J cadenas durante R iteraciones. Sea

$$\bar{\theta}_{.j} = \frac{1}{R} \sum_{i=1}^R \theta^{(ij)} \quad (2.4.20)$$

donde $\theta^{(ij)}$ es el valor i -ésimo generado por la cadena j -ésima. Sea

$$\bar{\theta}_{..} = \frac{1}{J} \sum_{j=1}^J \bar{\theta}_{.j} \quad (2.4.21)$$

Se calculan las varianzas (B) entre y (W) dentro de cada cadena.

$$B = \frac{R}{J-1} \sum_{j=1}^J (\bar{\theta}_{.j} - \bar{\theta}_{..})^2 \quad (2.4.22)$$

$$W = \frac{1}{J} \sum_{j=1}^J s_j^2 \quad (2.4.23)$$

donde s_j^2 es la varianza estimada de la cadena j -ésima. Con esto, podemos obtener una estimación de la varianza a posteriori $V(\theta|x)$, dada por

$$\widehat{V(\theta|x)} = \frac{J-1}{J} W + \frac{1}{R} B \quad (2.4.24)$$

que sobrestima la verdadera varianza si los puntos iniciales son muy dispersos pero es insesgado suponiendo estacionariedad (cuando $J \rightarrow \infty$). No obstante, para J finita, la varianza estimada dentro de cada secuencia, W , debe subestimar $V(\theta|x)$ porque las cadenas no han tenido tiempo para muestrear toda la distribución estacionaria, pero cuando $J \rightarrow \infty$, W también es insesgada. Calculando la razón

$$\frac{\widehat{V(\theta|x)}}{W}, \quad (2.4.25)$$

si es mucho más grande que 1, proporciona evidencia de falta de convergencia. Para problemas con múltiples simulaciones, se tiene un tiempo de cálculo superior, por lo que es conveniente asignar un número apropiado, tanto de simulaciones como de cadenas.

2.5. Geoestadística

2.5.1. Conceptos y definiciones básicas

2.5.1.1. Variables regionalizadas

Definiremos como variable regionalizada $z(x)$, donde x representa a un punto genérico del espacio, a una variable numérica distribuida en el espacio, de forma que presenta una estructura de correlación espacial. Una definición equivalente, más rigurosa matemáticamente, consiste en decir que una variable regionalizada es una realización de un proceso estocástico Z_D con el conjunto de índices contenido en un espacio euclidiano d -dimensional,

$Z_D \equiv \{Z(x) : x \in D \subset \mathbb{R}^d\}$. En general x es un punto en el espacio bi- o tridimensional ($d = 2$ o 3), es decir $x = (u_1, u_2)$ o $x = (u_1, u_2, u_3)$. En términos prácticos $z(x)$ puede verse como una medición de una variable aleatoria en un punto x de una región de estudio D . Una muestra de Z_D está dada por

$$\mathbf{z}_n = \{z(x_i), i = 1, \dots, n, x_i \in D\} \quad (2.5.1)$$

2.5.1.2. Función aleatoria

Si a cada punto x que pertenece a un dominio en el espacio le hacemos corresponder una variable aleatoria $Z(x)$, entonces el conjunto de variables aleatorias espacialmente distribuidas $\{Z(x) : x \in D\}$, corresponderá a un proceso estocástico o función aleatoria. Al tomar una realización de una función aleatoria $Z(x)$, se obtendrá una función espacial $\{z(x) : x \in D\}$, la cual constituye una variable regionalizada.

2.5.1.3. Distribución espacial de una función aleatoria

Sea Z una función aleatoria definida en $D \subset \mathbb{R}^d$, entonces el vector aleatorio $\{Z(x_1), Z(x_2), \dots, Z(x_n)\}$ se caracteriza por su función de distribución finito-dimensional:

$$F_{Z(x_1), Z(x_2), \dots, Z(x_n)}(z_1, z_2, \dots, z_n) = P(Z(x_1) \leq z_1, Z(x_2) \leq z_2, \dots, Z(x_n) \leq z_n). \quad (2.5.2)$$

El conjunto de todas las distribuciones para todo valor de n y para cualquier $(z_1, \dots, z_n) \in \mathbb{R}^n$ constituye la distribución espacial de la función aleatoria. Esta distribución en la práctica es imposible de determinar a partir de un número finito de observaciones y sólo se puede esperar inferir sus primeros momentos.

2.5.1.4. Momentos de una función aleatoria Z

- El momento de primer orden de Z es la esperanza matemática definida como

$$\mu(x) = E[Z(x)], \quad x \in D. \quad (2.5.3)$$

- Los momentos de segundo orden considerados en geoestadística son:
 - La varianza de $Z(x)$

$$\sigma^2 = Var[Z(x)] = E[\{Z(x) - \mu(x)\}^2], \quad x \in D. \quad (2.5.4)$$

- La covarianza entre dos variables aleatorias $Z(x_i)$ y $Z(x_j)$ definida como

$$C(x_i, x_j) = E[\{Z(x_i) - \mu(x_i)\}\{Z(x_j) - \mu(x_j)\}], \quad x_i, x_j \in D. \quad (2.5.5)$$

Esta función también es conocida como función de autocovarianza.

– El semivariograma $\gamma(x_i, x_j)$ que se define como

$$\gamma(x_i, x_j) = \frac{1}{2} \text{Var} [Z(x_i) - Z(x_j)] \quad (2.5.6)$$

ó

$$\gamma(x_i, x_j) = \frac{1}{2} E[\{Z(x_i) - Z(x_j)\}^2], \quad x_i, x_j \in D, \quad (2.5.7)$$

si la función aleatoria Z es estacionaria en la media. También es conocido como función de semivarianza. Se debe notar que tanto la varianza como el variograma son siempre positivos o nulos, mientras que la covarianza puede tomar valores negativos.

2.5.1.5. Funciones aleatorias estacionarias

Se dice que una función aleatoria Z es estrictamente estacionaria si su distribución espacial es invariante a cualquier traslación respecto a un vector h o, lo que es equivalente, la función de distribución del vector aleatorio $Z(x_1), Z(x_2), \dots, Z(x_n)$ es idéntica a la del vector $Z(x_1 + h), Z(x_2 + h), \dots, Z(x_n + h)$ para cualquier h . Puesto que, como se planteó anteriormente, usualmente se trabaja sólo con los momentos, con lo cual resulta práctico limitar la hipótesis de estacionariedad a estos primeros momentos. Se dice que una función aleatoria es estacionaria de segundo orden o débil si se cumple que:

1. su valor esperado existe y no depende de x

$$E[Z(x)] = \mu; \quad \forall x \in D, \quad (2.5.8)$$

2. para cualquier par de variables aleatorias $Z(x)$ y $Z(x + h)$, su covarianza existe y sólo depende del vector de separación h

$$C(h) \equiv C(x + h, x) = E[Z(x + h)Z(x)] - \mu^2, \text{ tal que } x, x + h \in D. \quad (2.5.9)$$

La estacionaridad de la covarianza implica que la varianza existe, es finita y no depende de x , es decir

$$\sigma^2 = C(0) = \text{Var}[Z(x)]. \quad (2.5.10)$$

Asimismo bajo esta hipótesis el variograma también es estacionario y se cumple que

$$\gamma(h) \equiv \gamma(x + h, x) = \frac{1}{2} E[\{Z(x + h) - Z(x)\}^2]. \quad (2.5.11)$$

Además existe una relación directa entre el variograma y la función de covarianza

$$\gamma(h) = C(0) - C(h). \quad (2.5.12)$$

En este caso resulta suficiente usar una de las dos funciones (covarianza o variograma) para caracterizar la dependencia espacial.

Asumiendo la existencia de los primeros dos momentos de $Z(x)$ para todo $x \in D$, el primer momento $E\{Z(x)\} = \mu(x)$ es usualmente llamado tendencia, mientras que la existencia del

segundo momento permite la definición de estacionaridad (débil).

Además de los dos tipos de estacionaridad previamente mencionados (estricta y débil), existe un tercer tipo de estacionaridad, llamada estacionariedad intrínseca. Aquí, no es necesario el supuesto de $E\{Z(x)\} \equiv \mu$, mientras que $E\{(Z(x+h) - Z(x))^2\}$ depende sólo de h . Si este es el caso, escribiremos $2\gamma(h) = E\{(Z(x+h) - Z(x))^2\}$. En lo siguiente, usaremos $|\cdot|$ para denotar la norma euclidiana de un vector.

2.5.2. Análisis estructural

Después de realizar el análisis descriptivo de los datos, lo que sigue es determinar la dependencia espacial entre los datos medidos de una variable. Este análisis se llama análisis estructural, y para realizarlo se utiliza la información muestral y tres funciones: variograma, covariograma y correlograma. El análisis estructural consiste en estimar y modelar una función que refleje la correlación espacial de la variable regionalizada a partir de la adopción razonada de la hipótesis más adecuada acerca de su variabilidad. Esto quiere decir, que en dependencia de las características de estacionaridad del fenómeno, se modelará la función de covarianza, o correlograma.

2.5.2.1. Variograma

El variograma es una herramienta que permite analizar el comportamiento espacial de una variable sobre un área definida. En otras palabras un variograma mide la variación de las observaciones con la dirección y la distancia y describe la relación espacial entre los valores observados. En la sección anterior se definió la estacionaridad intrínseca en la cual se mencionó que se asumía que la varianza de los incrementos de la función aleatoria asociada a la variable regionalizada era finita.

2.5.2.2. Covarianza y Correlograma

La función de covarianza muestral entre pares de observaciones que se encuentran a una distancia h se calcula, empleando la fórmula clásica de la covarianza, por:

$$\begin{aligned} \widehat{C}(h) &= \widehat{Cov}(Z(x+h), Z(x)) = \frac{\sum_{i=1}^n (z(x_i+h) - \widehat{\mu})(z(x_i) - \widehat{\mu})}{n} \\ &= \frac{\sum_{i=1}^n (z(x_i+h))(z(x_i))}{n} - \widehat{\mu}^2 \end{aligned}$$

donde $\widehat{\mu}$ representa el valor promedio estimado en todo punto de la región de estudio y n es el número de parejas de puntos con datos que se encuentran a una distancia h . Asumiendo que el fenómeno es estacionario y estimando la varianza de la variable regionalizada a través

de la varianza muestral, se tiene que el correlograma muestral está dado por:

$$\hat{\rho}(h) = \frac{\widehat{Cov}(Z(x+h), Z(x))}{\sqrt{\widehat{Var}(Z(x+h)) \times \widehat{Var}(Z(x))}} = \frac{\widehat{C}(h)}{\widehat{C}(0)}.$$

Bajo el supuesto de estacionaridad estricta o débil, cualquiera de las tres funciones de dependencia espacial mencionadas, es decir variograma, covarianza o correlograma, puede ser usada en la determinación de la relación espacial entre datos.

2.5.2.3. Aspectos generales del modelo de variograma

Si el variograma $\gamma(h)$ depende solamente de la longitud del vector h , estaremos en presencia de un proceso isótropo, lo que refleja que las correlaciones por pares entre la variable de interés en dos localidades diferentes depende sólo de sus distancias. De lo contrario, se denomina anisótropo. Los procesos isótropos son muy usados en el análisis de datos espaciales por su interpretación y disponibilidad de funciones paramétricas para el variograma, que puede ser escrito como $\gamma(|h|)$. El modelo *Matèrn* ha surgido como un modelo flexible, el que está dado por

$$\gamma(|h|) = \begin{cases} \tau^2 + \sigma^2 - m(\sigma^2, \zeta, \nu, |h|) & \text{si } |h| > 0 \\ 0 & \text{en otro caso} \end{cases}$$

donde

$$m(\sigma^2, \zeta, \nu, |h|) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (2\zeta\sqrt{\nu}d)^\nu K_\nu(2\zeta\sqrt{\nu}d)$$

es la función *Matèrn*. Aquí, ζ es una medida de decaimiento de la correlación con la distancia que separa a las observaciones, ν es un parámetro de suavizamiento, $\Gamma(\cdot)$ es la función Gamma convencional y $K_\nu(\cdot)$ es la función de Bessel modificada de segundo tipo de orden ν . Numerosos otros modelos de variograma se utilizan en la práctica, tales como el efecto pepita, el modelo esférico o el modelo exponencial ([Chilès y Delfiner \(2012\)](#)), por citar los más usados. Teniendo en cuenta una variedad de opciones para la función variograma, una pregunta natural, es cómo decidir el que mejor se ajusta a unos datos observados o si los datos pueden ser distinguidos por él.

2.5.3. Geoestadística lineal clásica

Con el desarrollo de la tecnología ha ido surgiendo un creciente interés por el análisis estadístico de datos georeferenciados. Este interés seguirá en aumento, puesto que los desarrollos tecnológicos capaces de generar bases de datos espacialmente referenciadas son cada vez más y de mejor calidad. El interés de las autoridades gubernamentales y el sector productivo por este tipo de información crece día a día y el elenco de campos del saber que tratan con información georeferenciada es cada vez mayor. Como señalaba Tobler ([Moellering y Tobler \(1972\)](#)), “todas las cosas están relacionadas con todas, pero las cosas que están cercanas lo están más que aquellas que están alejadas”.

Esta aseveración, conocida como la primera ley de la geografía, pone una piedra en el camino a la estadística clásica, al considerar la dependencia espacial de las observaciones e invita a

desarrollar nuevos procedimientos que tengan en cuenta tal circunstancia.

Dado que el valor de una observación en una localización puede informar sobre el valor en otra localización de interés cercana a ella, resultará crucial el análisis de la estructura de la dependencia espacial entre las observaciones disponibles y el desarrollo de procedimientos que permitan predecir el valor de la característica objeto de estudio en una localización no observada aprovechando la estructura de la dependencia espacial que presenta dicha característica. Entre los procedimientos que ocupan un lugar privilegiado es el kriging ([Diggle y cols. \(2010\)](#)).

La geoestadística interpreta cada valor $z(x_i)$ de una variable regionalizada como una realización particular de una variable aleatoria $Z(x_i)$, y el conjunto de éstas para $x_i \in D$, constituye una función aleatoria ($Z = \{Z(x_i), \forall x_i \in D\}$). El problema de caracterizar la variabilidad espacial se reduce entonces a caracterizar la correlación entre las variables aleatorias que integran la función aleatoria.

2.5.3.1. Ventajas de la geoestadística lineal clásica

Como en todo procedimiento el resumen y análisis de la información, en la práctica están sujetos a sus ventajas y desventajas. En relación al kriging, tenemos:

1. El kriging obtiene como resultados los mejores predictores lineales no sesgados y hoy en día, el término está referido a toda una familia de técnicas de predicción.
2. Si los sitios de observación son todos distintos, se demuestra que el sistema de kriging es regular, es decir entrega una solución única.
3. El kriging es un interpolador exacto, es decir, que la predicción en un sitio muestreado vuelve a dar el valor medido y la varianza de kriging (varianza del error de predicción) en ese punto es nula.
4. Sesgo condicional: El error promedio puede no tener esperanza nula cuando se considera sólo los sitios donde la predicción es alta (o baja). Sin embargo, en general el sesgo condicional es pequeño si se usa suficientes datos ([Rivoirard \(1987\)](#)).

2.5.3.2. Desventajas de la geoestadística lineal clásica

1. La predicción tiende a ser poco confiable si el ajuste de los parámetros del variograma resulta ser incorrecta. En general, no se considera información previa sobre los parámetros que modelan la dependencia espacial, como en el caso bayesiano.
2. Es poco confiable para predecir la ocurrencia de valores extremos, pues al tener una tendencia hacia la media global, reduce la variabilidad y produce predicciones suavizadas.
3. Puede suceder que sitios de observación muy cercanos causen inestabilidad en la inversión de la matriz del sistema de ecuaciones lineales a resolver, la cual presenta entonces dos líneas casi iguales. Una solución a este problema consiste en agregar al modelo de variograma un “efecto pepita”; otra solución es agrupar al momento de estimar los sitios muy próximos y considerar un sitio promedio.

4. El mapa de las predicciones obtenidas por kriging es siempre más suave que el mapa de los valores puntuales reales, es decir, que presenta menos fluctuaciones. La búsqueda de una predicción precisa se acompaña inevitablemente de este efecto de suavizamiento, pues no se puede inventar los detalles que no aparecen en las observaciones. Si se desea reproducir la variabilidad de la variable regionalizada, es necesario recurrir a otros procedimientos como el de simulación.

2.5.3.3. Modelo espacial con respuesta Gaussiana

En muchos estudios con datos espaciales, junto con los resultados, también se observan variables de exploración, midiendo por ejemplo las características de cada individuo en lugares propuestos. En el marco del proceso espacial, denotaremos a las q covariables del proceso por $\{R(x), x \in D \subset \mathbb{R}^d\}$. El modelo espacial lineal de $Z(x)$ dado $R(x)$ puede ser escrito como

$$Z(x) = R(x)' \beta + U(x) \quad (2.5.13)$$

donde $R(x)$ es una matriz de diseño que contiene a las q covariables del proceso espacial, β es un vector de coeficientes de regresión en algún subconjunto abierto de \mathbb{R}^q y $U(x)$ es una función aleatoria Gaussiana con media cero (débilmente estacionario) y función de covarianza espacial dada por

$$\text{cov}\{U(x), U(x')\} = C(x, x'; \theta) \quad (2.5.14)$$

donde θ es un vector de parámetros de $k \times 1$ en un subconjunto abierto de \mathbb{R}^k . El parámetro β caracteriza la parte determinista de los datos espaciales y se llama a veces el parámetro de tendencia, mientras que θ caracteriza la variabilidad espacial subyacente a través de la función de covarianza espacial $C(x, x'; \theta)$. Una práctica común es asumir $U(x)$ como estacionario de modo que la función de covarianza $C(\cdot)$ dependa de x y x' sólo a través de su separación $h = x - x'$. En este caso, podemos escribir $C(x, x'; \theta)$ como $C(x - x'; \theta) = C(h; \theta)$, el que puede ser fácilmente especificado a partir del variograma γ , por ejemplo por el modelo Matèrn descrito anteriormente. Estamos en posición de realizar inferencia basada en el modelo anterior por métodos de máxima verosimilitud. Dado un conjunto finito de ubicaciones x_1, \dots, x_n , entonces la matriz $[C(h; \theta)]$ de $n \times n$, que denotaremos por Σ_n es definida positiva. Denotaremos por $Z_n = (Z(x_1), \dots, Z(x_n))$ a los datos de la variable de interés y por $R_n = (R(x_1), \dots, R(x_n))'$ a los vectores que componen la matriz de diseño, con cada $R(x_i)$ de $(q+1) \times 1$. Dado $\Theta = (\beta', \theta)'$ denota el vector de $(q + k + 1)$ parámetros. Entonces la log-verosimilitud es

$$L_n(\Theta) = -(n/2) \log(2\pi) - (1/2) \log(|\Sigma_n|) - (1/2) (Z_n - R_n \beta)' \Sigma_n^{-1} (Z_n - R_n \beta). \quad (2.5.15)$$

2.6. Modelamiento Geometalúrgico usando Análisis de Conglomerados

2.6.1. Validación e interpretación de resultados

La validación corresponde a la última etapa del análisis de conglomerados, no así la menos importante.

En esta etapa se analizan los resultados obtenidos y se obtienen las conclusiones definitivas del estudio. Los métodos para validar un análisis de conglomerados difieren en el caso jerárquico y no jerárquico. En el primero de ellos, se busca cuantificar el grado de similitud entre los objetos de la solución final y encontrar el número óptimo de conglomerados a formar con tal de representar de buena forma la estructura natural de los datos. Con respecto a los métodos no jerárquicos, las preguntas anteriores pierden sentido, pues el objetivo principal es estudiar la homogeneidad de los grupos encontrados. Es necesario estar abierto a la posibilidad que no todos los conglomerados formados presenten una estructura definida. Existe una gran variedad de técnicas estadísticas que sirven para validar los resultados obtenidos, entre ellas, destacamos el Análisis de Varianza (ANOVA) y el Análisis de Varianza Multivariado (MANOVA), que sirven para identificar los grupos que son significativamente distintos y en base a qué variables lo son. Es posible realizar Análisis Discriminantes ([Hardle y Simar \(2007\)](#)) para poder determinar las fronteras naturales entre los conglomerados formados, además de Análisis Factoriales ([Bailey \(1994\)](#)) o de Componentes Principales ([Seber \(2009\)](#)) para representar gráficamente los grupos obtenidos y observar las diferencias existentes entre ellos. Determinar las características que definen a cada conglomerado obtenido es un paso crucial para su correcta interpretación, ya que en base a esto podremos evaluar la correspondencia de los resultados obtenidos con el conocimiento experimental con el que disponemos.

2.6.2. Límites de las técnicas de agrupamiento con datos regionalizados

En vista que se está trabajando con variables medidas en el espacio, es necesario contar con métodos que tomen en cuenta la continuidad espacial de los datos. Una alternativa muy utilizada es incorporar las coordenadas espaciales como variables al proceso de conglomerados, de esta forma, actúan como valores que condicionan las etapas de los algoritmos involucrados y se consigue cierta continuidad en las unidades creadas. No obstante, podrían existir problemas con esta metodología, pues si no se estandarizan las coordenadas antes de iniciar los algoritmos de conglomerados, éstas podrían tener un peso mucho mayor que el de las variables de interés reales, haciendo que la dependencia espacial pase a segundo plano y sea reemplazada por la compacidad de las unidades a formar. Por este motivo, es necesario buscar el equilibrio entre continuidad espacial y dependencia espacial, poniendo énfasis en este último, pues corresponde a la característica que define a los procesos de interés en el área de estudio y además, la correcta caracterización de esta dependencia condiciona los resultados en todas las etapas de la evaluación de depósitos minerales. Un ejemplo de esto es posible verlo en el artículo “Domaining by clustering multivariate geostatistical data” ([Romary y cols. \(2012\)](#)). En el documento se muestra la metodología tradicional del análisis de conglomerados y se entrega una alternativa que, además de respetar las relaciones entre variables, incorpora el concepto de continuidad espacial al momento de agrupar conglomerados ([Hervada-Sala y Jarauta-Bragulat \(2004\)](#); [Oliver y Webster \(1989\)](#)).

2.6.3. Aplicaciones a modelamiento geológico o geometalúrgico

En evaluación de yacimientos existe la necesidad de reducir la incertidumbre en el proceso de evaluación de recursos y reservas, por lo que el definir dominios con cierta homogeneidad en relación a variables geológicas o geometalúrgicas suele ser un aporte para este propósito (Rosales (2012)). Al tener dominios homogéneos, se espera que las observaciones pertenecientes a él se comporten de manera similar, por lo que se reduce la variabilidad en las evaluaciones de recursos y se utilizan las relaciones inherentes entre variables. No se debe olvidar el contexto en el que se trabaja, pues las variables medidas son regionalizadas, por lo que el uso de técnicas geoestadísticas se hace imprescindible (Deutsch (2013); Shaw y cols. (2013)). Además de esto, la existencia de bases de datos de gran volumen hace necesario el uso de técnicas óptimas, con las que se pueda obtener resultados en corto tiempo y de manera sencilla. Técnicas del área de Data Mining son muy utilizadas en estas condiciones (Allard y Guillot (2000); Ambroise y cols. (1997)) por la capacidad que poseen para identificar relaciones entre variables y descubrir nuevo conocimiento a partir de ellas (Zhu (2007)).

La ventaja radica en la existencia de técnicas de conocimiento no supervisado, las que permiten obtener estas relaciones sin la necesidad de intervención por parte del usuario, disminuyendo cualquier tipo de sesgo. Se utilizan estas técnicas para ayudar al diseño de mejores métodos de recuperación de recursos, así como para reducir al mínimo el impacto ambiental de la explotación de recursos al tener un mayor control sobre los procesos que conlleva la operación (Hoal y cols. (2013)).

Capítulo 3

Conglomerados Geoestadísticos basados en Métodos Jerárquicos

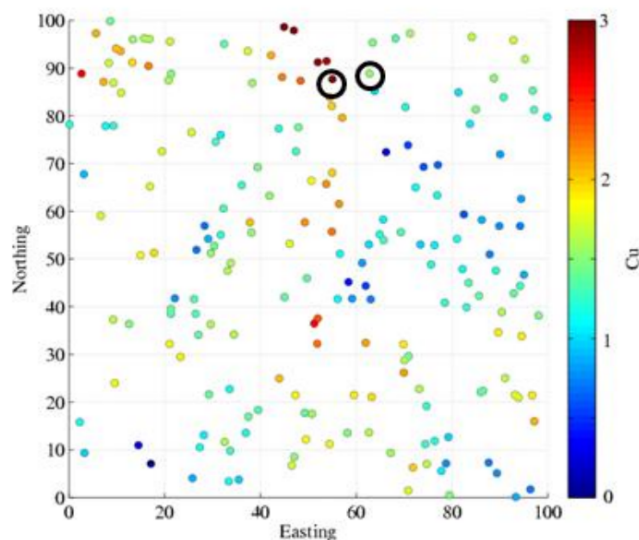
3.1. Introducción

En el estudio de variables geometalúrgicas se cuenta con variables de diversos orígenes, las que pueden ser cuantitativas o cualitativas, por lo que se debe tener precaución al establecer algún tipo de relación entre ellas. No obstante, todas las variables comparten una característica en común: son variables medidas en el espacio (variables regionalizadas). Esta característica establece que en cada localidad del espacio existe una variable o proceso aleatorio, cuya realización corresponde a la muestra con la que disponemos en las bases de datos y, además, cada variable del espacio depende de alguna forma de los procesos en localidades cercanas a ella. Por este motivo, a pesar no observar alguna relación entre variables, esto no quiere decir que el proceso completo se comporte de forma puramente aleatoria, pues el tipo de dependencia puede estar reflejado en una cierta continuidad espacial. Esto y la falta de información exhaustiva que provoca incertidumbre en las evaluaciones de recursos y reservas, sustentan la necesidad de considerar las propiedades pasadas por alto en un análisis estadístico multivariado tradicional.

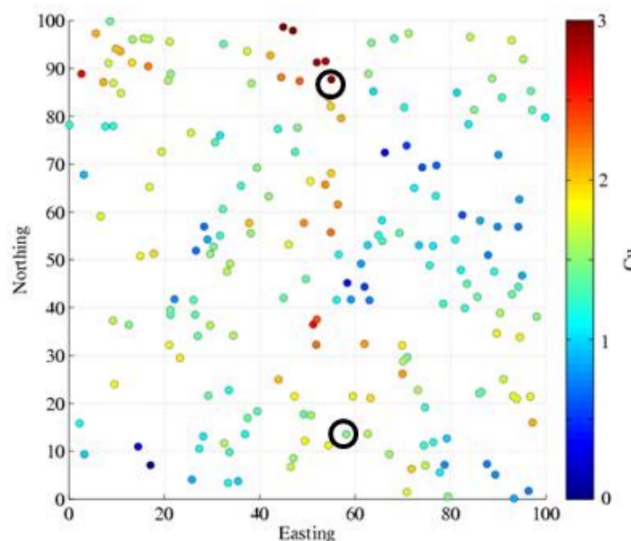
3.2. Metodología propuesta

Una vez que las unidades geológicas del depósito se encuentran definidas, se procede a identificar las Unidades Geometalúrgicas (UGM), es decir, regiones del espacio con un comportamiento geológico y metalúrgico similar. Para realizar este proceso, se cuenta con diferentes procedimientos que involucran una serie de análisis, utilizando técnicas de análisis de datos multivariantes y de Data Mining. En específico, para formar regiones cuyos individuos sean lo más semejantes entre sí, se utiliza el análisis de conglomerados en cualquiera de sus versiones (jerárquico-no jerárquico). Para este propósito, es necesario definir una distancia entre tales individuos, la que tradicionalmente se limita a la medida Euclideana o de Mahalanobis.

El problema con usar estas medidas radica en que pasan por alto el concepto de variables regionalizadas y la dependencia espacial existente entre los individuos. Para ejemplificar este problema, proponemos considerar el caso planteado en la Figura 3.1 de disimilitud entre leyes de cobre medidas en distintas localidades que justifica la necesidad de la propuesta. Es cuestionable el que ambas disimilitudes sean iguales, dado que el primer par de puntos son geográficamente mucho más cercanos que el segundo par de puntos. Debemos medir la disimilitud entre pares de puntos utilizando alguna función que dependa de la distancia geográfica entre las observaciones, tal y como se muestra en la Figura 3.2. Para este propósito, usaremos la función variograma, que cuantifica la desviación esperada entre observaciones en función de la distancia de separación espacial (Isaaks y Srivastava (1989)).



(a) $d_1 = rojo(3,0\%Cu) - verde(1,5\%Cu) = 1,5\%Cu$



(b) $d_2 = rojo(3,0\%Cu) - verde(1,5\%Cu) = 1,5\%Cu$

Figura 3.1: Disimilitud entre valores de leyes de cobre para dos observaciones.

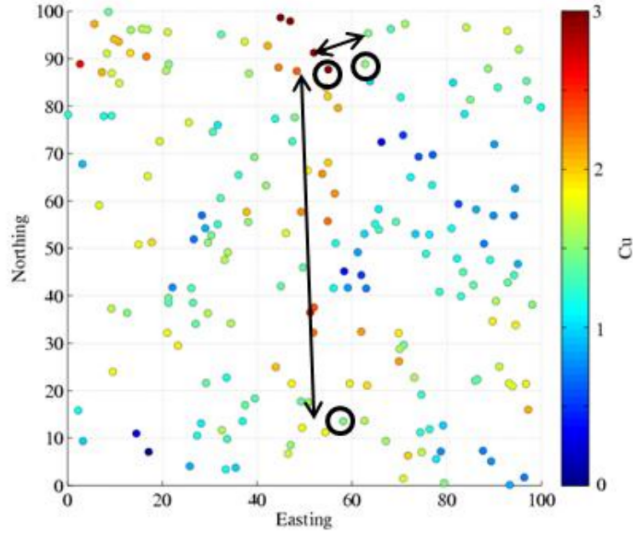


Figura 3.2: Disimilitud entre dos observaciones.

Así, podemos definir la distancia cuadrática entre dos observaciones, z_1 y z_2 , ubicadas en los puntos x_1 y x_2 respectivamente, como:

$$d_{12}^2 = \frac{(z_1 - z_2)^2}{\gamma(x_1 - x_2)} \quad (3.2.1)$$

donde γ representa el variograma asociado a la variable considerada (vista como realización de una función aleatoria Z):

$$\gamma(h) = \frac{1}{2} E \{ [Z(x+h) - Z(x)]^2 \} \quad (3.2.2)$$

En un caso más general, consideraremos dos vectores de p variables de interés, en dos localidades del espacio ($z_1 = z(x_1)$ y $z_2 = z(x_2)$; $x_1, x_2 \in D$) Los métodos de conglomerados tradicionales suelen utilizar las siguientes medidas para cuantificar la proximidad entre individuos

- Distancia Euclideana:

$$d(z_1, z_2)^2 = [z(x_1) - z(x_2)]^T [z(x_1) - z(x_2)] \quad (3.2.3)$$

- Distancia de Mahalanobis:

$$d(z_1, z_2)^2 = [z(x_1) - z(x_2)]^T \Sigma^{-1} [z(x_1) - z(x_2)] \quad (3.2.4)$$

donde Σ corresponde a la matriz de varianzas-covarianzas de las variables que componen el vector medido en las localidades.

Para tener en cuenta la correlación espacial de las observaciones, lo que haremos será proponer una nueva medida que incluya la dependencia espacial entre las localidades. Suponiendo que tenemos la estructura de correlación de las variables de estudio modelada ([Cressie \(1985\)](#)),

reemplazaremos la matriz de varianzas-covarianzas Σ de la distancia de Mahalanobis por la matriz variograma $\Gamma(h)$ (Wackernagel (2003)), que contiene los variogramas directos y cruzados para todos los pares de datos ubicados a una distancia h de separación espacial (Euclídeana), obteniendo una nueva medida de distancia espacial que llamaremos

- Distancia de Mahalanobis geoestadística:

$$d(z_1, z_2)^2 = [z(x_1) - z(x_2)]^T \Gamma(x_1 - x_2)^{-1} [z(x_1) - z(x_2)] \quad (3.2.5)$$

$$\text{con } \Gamma(x_1 - x_2) = \frac{1}{2} E \left\{ [Z(x_1) - Z(x_2)]^T [Z(x_1) - Z(x_2)] \right\}.$$

Una distancia similar ha sido propuesta por Vallejos y cols. (2015) en el contexto de clasificación de imágenes obtenidas por percepción remota. De cierto modo, la disimilitud entre las dos observaciones $z(x_1)$ y $z(x_2)$ se mide relativamente a la disimilitud esperada para el par de puntos (x_1, x_2) , pudiendo así detectar observaciones anormalmente diferentes en localidades cercanas, las cuales podrían pertenecer a conglomerados diferentes. Como aplicación, se espera obtener una agrupación de unidades geometalúrgicas por los métodos de conglomerados tradicionales con un comportamiento metalúrgico determinado, a las cuales se les asigna las mismas ecuaciones de recuperación y de consumo de insumos de proceso. De esta forma se asegura una estimación más precisa de los recursos disponibles y un mayor control sobre las relaciones existentes entre las variables medidas.

3.3. Caso de estudio sintético

3.3.1. Metodología de simulación

Debido a la falta de algún conjunto de datos apropiados para el estudio (para los cuales existiese un conocimiento perfecto y sin incertidumbre de las unidades geometalúrgicas), se utilizaron datos simulados con características especiales para poner a prueba la metodología propuesta. Por una parte, el conjunto de datos debía poseer un número determinado de variables geometalúrgicas que presentaran una estructura de correlación espacial entre ellas, pues el interés era utilizar esta información para la búsqueda de unidades geometalúrgicas que, en la práctica, poseen esta característica que se refleja en los análisis multivariados realizados en numerosos estudios. Por otro lado, el conjunto de datos simulados debía representar las diferencias existentes entre localidades pertenecientes a unidades geometalúrgicas distintas. Esto debía estar marcado por cambios abruptos en la continuidad espacial más que por cambios en medidas estadísticas globales, pues es usual que unidades geometalúrgicas distintas, presenten estadísticas globales semejantes, pero medidas de continuidad espacial completamente diferentes.

Para cumplir con el primer requerimiento, tuvimos que hacer uso de procedimientos de cosimulación Gaussiana (no condicional), descritos en numerosos textos (Chilès y Delfiner (2012); Lantuéjoul (2013)). Este proceso hace uso de una estructura de correlación espacial fija para realizar simulaciones de funciones aleatorias Gaussianas correlacionadas entre ellas. Con esto, se garantizó la existencia de relaciones entre las variables simuladas y que éstas mismas

presentaran una dependencia espacial. Además de esto, se tuvo que garantizar que existiesen diferencias para unidades geometalúrgicas distintas, por lo que se optó por la siguiente metodología:

- Simular coordenadas de manera aleatoria
- Particionar el conjunto de coordenadas en un número fijo de unidades geometalúrgicas (partición geográfica, de modo que cada unidad geometalúrgica sea espacialmente continua/conexa)
- Realizar el proceso de cosimulación Gaussiana en cada unidad geometalúrgica de forma independiente.

Como resultado de este proceso, se obtuvo una base de datos sintéticos, con información de unidades geometalúrgicas por un lado y, por otro lado, de variables Gaussianas (emulando variables geometalúrgicas) inter-dependientes dentro de cada unidad geometalúrgica, pero independientes entre una unidad y otra.

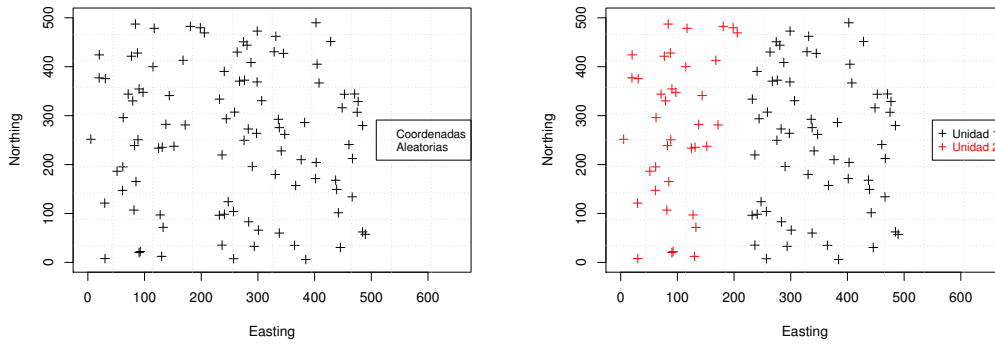
Una posibilidad fue la de asignar estructuras de correlación diferentes para cada unidad geometalúrgica, aunque por simplicidad y como etapa inicial, se optó por la misma estructura para todas las unidades, con el mismo rango o alcance de dependencia espacial y sin presencia de direcciones preferenciales, es decir, en el caso isótropo.

3.3.1.1. Caso inicial

Para el estudio, se co-simularon 6 variables regionalizadas en un dominio bidimensional de tamaño 500×500 , las que para una demostración previa, presentaron la estructura de correlación espacial presentada en la Tabla 3.1, con rangos o alcances idénticos iguales a 100, con una estructura esférica de dependencia espacial y sin presencia de efecto pepita. Luego, se procedió a simular 100 coordenadas aleatorias en el plano $[0, lim] \times [0, lim]$, con $lim = 500$, las que fueron particionadas para crear 2 grupos de forma aleatoria que hicieron de unidades geometalúrgicas distintas, tal y como se aprecia en la Figura 3.3. En cada una de ellas, se realizó el procedimiento de cosimulación Gaussiana no condicional, obteniéndose 100 realizaciones para cada una de las 6 variables en una misma localidad, las que servirán para cuantificar la efectividad de los métodos de análisis de conglomerados tradicionales y propuestos.

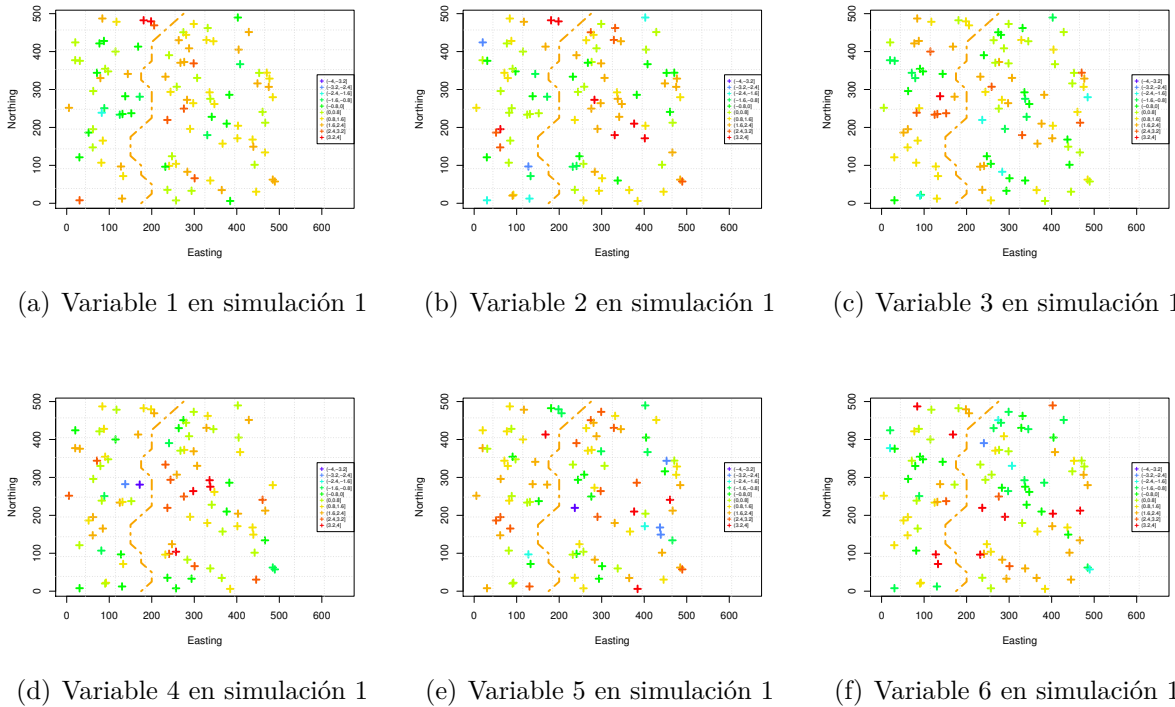
	var1	var2	var3	var4	var5	var6
var1	0.71	-0.23	0.14	0.11	-0.40	-0.31
var2	-0.23	1.06	-0.21	-0.29	-0.58	-0.32
var3	0.14	-0.21	1.95	0.05	-0.27	-0.05
var4	0.11	-0.29	0.05	0.86	0.28	-0.54
var5	-0.40	-0.58	-0.27	0.28	1.14	-0.08
var6	-0.31	-0.32	-0.05	-0.54	-0.08	1.79

Tabla 3.1: Matriz de mesetas caso inicial.



(a) Coordenadas simuladas de forma aleatoria (b) Clasificación de las mismas en dos U.G.M.

Figura 3.3: Simulación caso inicial.



(a) Variable 1 en simulación 1 (b) Variable 2 en simulación 1 (c) Variable 3 en simulación 1
 (d) Variable 4 en simulación 1 (e) Variable 5 en simulación 1 (f) Variable 6 en simulación 1

Figura 3.4: Variables cosimuladas en cada una de las U.G.M. de forma independiente.

Es posible apreciar de manera clara en la Figura (3.4) el cambio que existe en la continuidad espacial de las variables cuando se pasa de una unidad metalúrgica a otra para cada realización (presencia de discontinuidades al cruzar la frontera entre ambas unidades), razón de porqué se simuló de forma independiente.

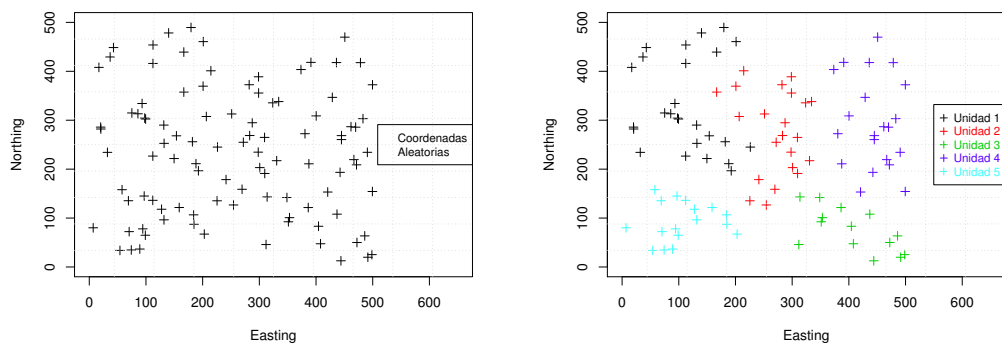
Cada realización fue construida a partir de la matriz de mesetas, rangos y efecto pepita definidos inicialmente, por lo que si se reconstruyen los variogramas directos y cruzados experimentales en cada unidad geometalúrgica, éstos debieran ser semejantes a las matrices que los originaron en forma aproximada.

3.3.1.2. Caso secundario

Ahora vamos a presentar un segundo caso con 5 unidades geometalúrgicas, donde la complejidad aumenta al considerar todas las interacciones entre estas unidades. Al igual que para el caso inicial, debemos especificar una matriz de mesetas para iniciar el proceso de simulación, la que se muestra en la Tabla 3.2. También cosimularemos variables isótropas, sin efecto pepita, con variogramas (directos y cruzados) esféricos y con los mismos rangos o alcances de manera independiente en cada unidad, obteniendo 100 realizaciones para hacer un resumen de ellas y poder cuantificar la efectividad del método de análisis de conglomerados utilizado. De la misma forma, particionamos el conjunto de coordenadas aleatorias en el plano $[0, lim] \times [0, lim]$, con $lim = 500$, en 5 unidades geometalúrgicas distintas, como se aprecia en la Figura (3.5). En cada uno de los grupos formados procedemos a hacer cosimulación no condicional Gaussiana de forma independiente, utilizando la matriz de mesetas antes definida. En la Figura 3.6 se muestra la distribución espacial de las variables para una de las simulaciones.

	var1	var2	var3	var4	var5	var6
var1	1.62	-0.16	-0.28	0.28	0.04	-0.22
var2	-0.16	1.79	0.15	0.07	0.59	-0.17
var3	-0.28	0.15	2.19	-0.24	0.28	-0.13
var4	0.28	0.07	-0.24	1.49	-0.47	0.03
var5	0.04	0.59	0.28	-0.47	2.05	0.02
var6	-0.22	-0.17	-0.13	0.03	0.02	1.64

Tabla 3.2: Matriz de mesetas caso secundario.



(a) Coordenadas simuladas de forma aleatoria (b) Clasificación de las mismas en dos U.G.M.

Figura 3.5: Simulación caso secundario.

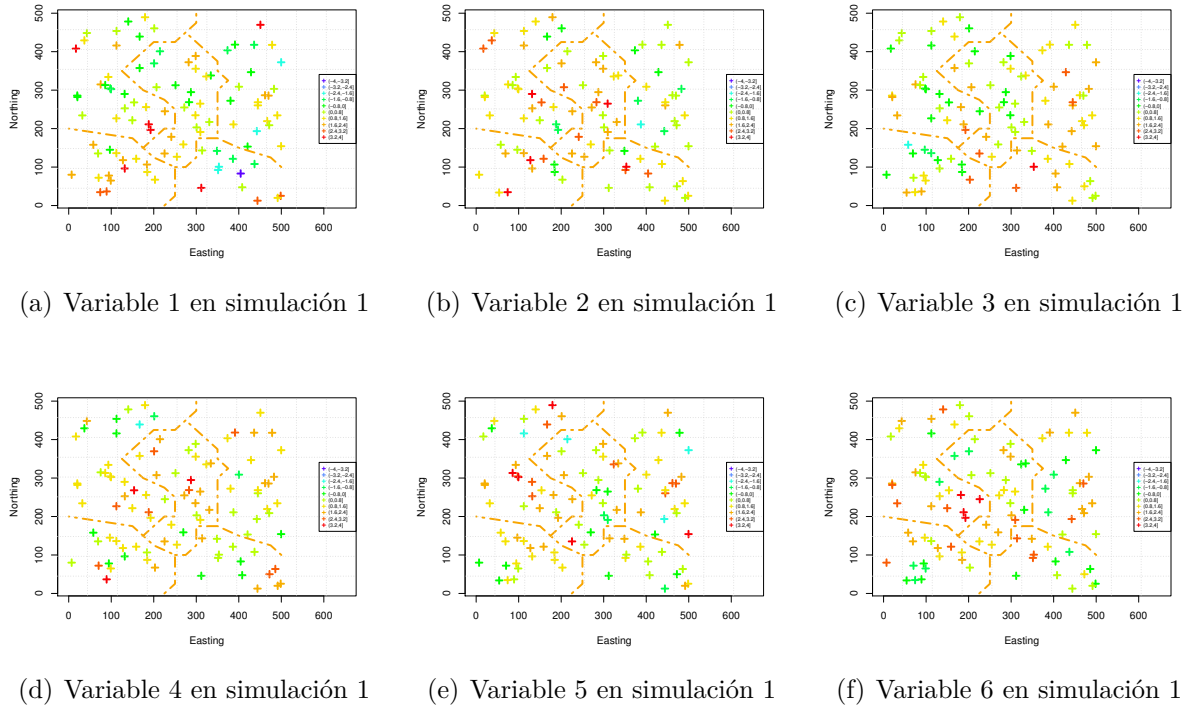


Figura 3.6: Variables cosimuladas en cada una de las U.G.M. de forma independiente.

3.3.1.3. Uso de conocimiento geológico

Además de leyes de distintos metales y otras variables geometalúrgicas, muchas veces disponemos de información proveniente del campo geológico, como definición de dominios, tipos de rocas, alteraciones u otros. La incorporación de esta información permitiría reducir la variabilidad de los datos y obtener estimaciones mucho más confiables y que representen la realidad de forma más fidedigna. Este tipo de información muchas veces es difícil de tratar, pues corresponde generalmente a variables categóricas nominales u ordinales, que no reciben el mismo tratamiento que una variable cuantitativa y los conceptos de dependencia espacial no son fáciles de asimilar. En el problema del descubrimiento de unidades geometalúrgicas a través de técnicas de clustering, tratamos de formar grupos con características semejantes, por lo que necesitamos alguna medida de proximidad en presencia de variables categóricas o mezclas de tipos de variables. Para este objetivo, existen las llamadas matrices de disimilitud, usadas en diversas áreas y que permiten el uso de un gran número de combinaciones de variables. En ecología, por ejemplo, existen diferentes medidas que permiten la construcción de estas matrices (Lance y Williams (1967); Legendre y Legendre (2012)). Cuando se disponen de datos de tipo mixto, es decir, variables cuantitativas y categóricas, una medida muy utilizada es la medida de Gower (Mountford (1962)), que está definida como $d_{ij}^2 = 1 - s_{ij}$, donde

$$s_{ij} = \frac{\sum_{h=1}^{p_1} (1 - |z_{ih} - z_{jh}|/G_h) + a + \alpha}{p_1 + (p_2 - d) + p_3}$$

es el coeficiente de similitud de Gower, donde z_{ih} corresponde al valor de la h -ésima variable cuantitativa continua para el i -ésimo individuo, p_1 es el número de variables cuantitativas

continuas, p_2 es el número de variables binarias, p_3 es el número de variables cualitativas (no binarias), a es el número de coincidencias (1,1) en las variables binarias, d es el número de coincidencias (0,0) en las variables binarias, α es el número de coincidencias en las variables cualitativas (no binarias) y G_h es el rango (o recorrido) de la h -ésima variable cuantitativa z_h .

Como no se contaba con datos reales para el estudio, utilizamos un supuesto que estaba relacionado con los alcances prácticos de las unidades geometalúrgicas. Se discutió previamente que los dominios geológicos aportaban información al problema, pues en base a ellos se puede definir una distancia geográfica (Euclideana) máxima de pertenencia a un dominio o a otro, es decir, a partir de cierta distancia para observaciones distintas, éstas pertenecen a dominios distintos. Esta restricción se tradujo como un acondicionamiento de la matriz de distancias Euclidianas entre localidades. Si la distancia euclideana $d_{ij} \geq d_{max}$, entonces las observaciones i y j pertenecen a unidades geometalúrgicas distintas, por lo que al emplear técnicas de clustering jerárquico para formar los grupos, éstas observaciones deben quedar en unidades diferentes de alguna manera. Esto actúa como una restricción sobre el proceso de formación de unidades geometalúrgicas y permite aportar una fuente de variabilidad al tener opciones de cómo elegir adecuadamente el valor del parámetro d_{max} .

Existe una segunda opción que de la misma forma impone restricciones sobre el proceso de clustering jerárquico, que es el de utilizar matrices de similitud para incrementar la distancia entre observaciones, siempre y cuando la diferencia en base a las variables categóricas sea considerable.

3.3.1.4. Medida de proximidad

Se definieron una serie de medidas que cuantificaban la proximidad entre observaciones, entre ellas, contamos con medidas de distancia para atributos numéricos. Dados dos vectores con p atributos numéricos (p -dimensionales), $z_i = (z_{i1}, z_{i2}, \dots, z_{ip})$ y $z_j = (z_{j1}, z_{j2}, \dots, z_{jp})$, la distancia entre ellos será calculada a través de las siguientes métricas en el estudio:

- Distancia Euclideana entre los atributos z_i y z_j

$$d(z_i, z_j) = \sqrt{\{(z_i - z_j)(z_i - z_j)^T\}} \quad (3.3.1)$$

En el estudio se propone una nueva métrica, basada en la distancia de Mahalanobis, pero que considera la dependencia espacial de las observaciones para descubrir los alcances de las unidades geometalúrgicas.

- Distancia geoestadística

$$d(z_i, z_j) = \sqrt{\{(z_i - z_j)\Gamma^{-1}(x_i - x_j)(z_i - z_j)^T\}} \quad (3.3.2)$$

donde z_i corresponde al vector de variables regionalizadas observadas en la localidad x_i , $h = x_i - x_j$ corresponde a la distancia euclideana utilizando como información las coordenadas de las localidades i y j , y $\Gamma(h)$ corresponde a la matriz que contiene los variogramas directos y cruzados para una distancia h .

3.3.1.5. Algoritmo de formación de unidades geometalúrgicas

La formación de los grupos o unidades geometalúrgicas se realiza mediante el proceso de clustering jerárquico (Everitt (1974)) utilizando el algoritmo de Ward (Ward Jr (1963)) para realizar el agrupamiento.

El Método de Ward (definido e implementado recursivamente a través del algoritmo de Lance-Williams (Lance y Williams (1967)) es uno de los más utilizados en la práctica y el utilizado en este estudio; posee casi todas las ventajas del método *average linkage* (Gordon (1999)) y suele ser más discriminativo en la determinación de los niveles de agrupación.

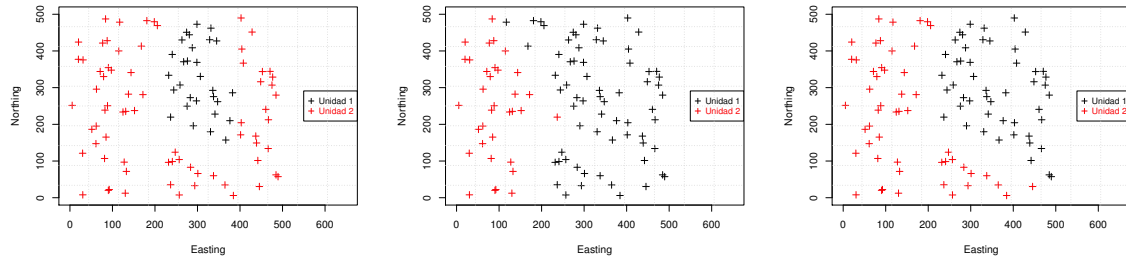
3.3.2. Resumen de resultados

En las simulaciones pertenecientes tanto al al conjunto de datos con dos así como con cinco unidades geometalúrgicas (caso inicial y secundario respectivamente), se calculó la “Distancia geoestadística” para cada par de localidades. Luego, mediante el método de clustering jerárquico aglomerativo, tomando como criterio de agrupación el principio de Ward o de varianza mínima, se formaron los grupos, suponiendo que el número de unidades geometalúrgicas era conocido. Esto último se sustenta en el hecho de que nuestro interés es verificar si la distancia geoestadística propuesta es capaz de descubrir los límites entre las unidades geometalúrgicas existentes utilizando la dependencia espacial de las variables y sus interrelaciones. Como conocimiento extra supondremos que existe una distancia geográfica máxima a partir de la cual dos observaciones deberían pertenecer a unidades geometalúrgicas diferentes. Esto condiciona la formación de los grupos y está reflejado en un condicionamiento de la distancia geoestadística utilizada, de la forma:

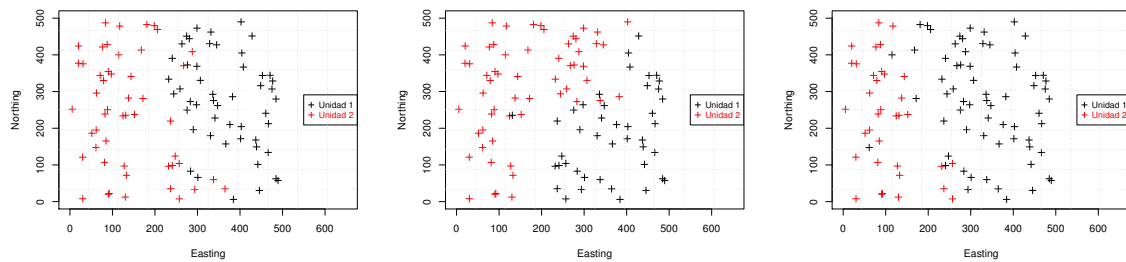
$$d_g(z_i, z_j) = \begin{cases} d_g(z_i, z_j) & \text{si } d_1(x_i, x_j) < d_{max} \\ d_g(z_i, z_j) \times \alpha & \text{si } d_1(x_i, x_j) \geq d_{max} \end{cases} \quad (3.3.3)$$

donde $d_g(z_i, z_j)$ corresponde a la distancia geoestadística entre las observaciones z_i y z_j utilizando las realizaciones de las variables regionalizadas para su cálculo, $d_1(x_i, x_j)$ es la distancia euclideana entre las localidades x_i y x_j , d_{max} es la distancia a partir de la cual dos observaciones deberían pertenecer a dos unidades geometalúrgicas distintas y α es un valor positivo grande. La pregunta que surge de forma natural es cómo determinar el valor d_{max} . Se presentan resultados obtenidos para diferentes valores de d_{max} en el caso inicial y secundario respectivamente. El modo de elegir el valor d_{max} de interés será discutido en las conclusiones.

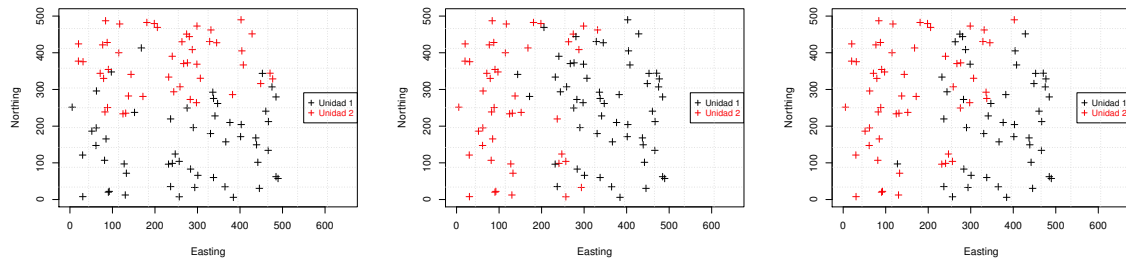
3.3.2.1. Resultados caso inicial



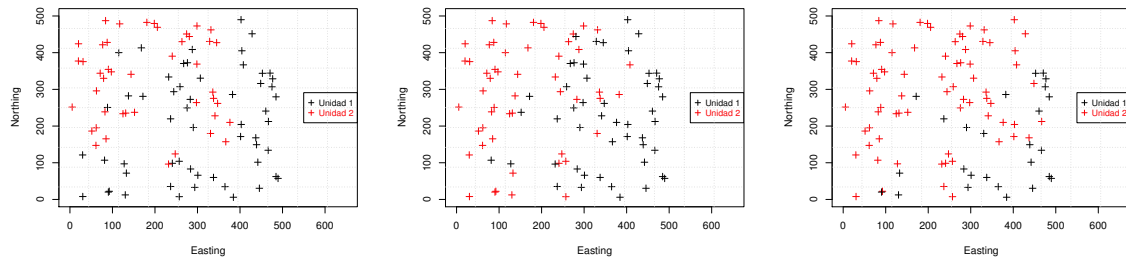
(a) $d_{max} = 100$



(b) $d_{max} = 200$



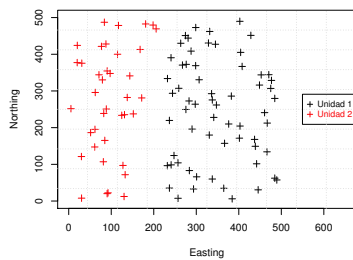
(c) $d_{max} = 300$



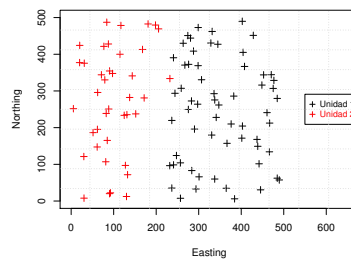
(d) $d_{max} = 400$

Figura 3.7: Resultados del clustering jerárquico usando la Distancia Geoestadística y el Método de Ward para tres simulaciones del caso inicial.

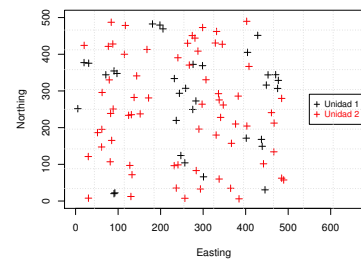
Podemos observar los clusters formados para cada una de las realizaciones del caso simple en la Figura (3.7), los que están ordenados en función del valor d_{max} , para $d_{max} = \{100, 200, 300, 400\}$. Se destaca que, a medida que el valor de d_{max} aumenta, se pierde conectividad en la formación de clusters, a partir de $d_{max} \geq 300$. Esta información nos permite hacer una estimación del valor d_{max} que permita la formación de clusters más conexos, por lo que la figura siguiente, muestra el resumen de los resultados para $d_{max} = 200$. La Figura (3.8) muestra el resumen de las 100 realizaciones obtenidas y los alcances de las unidades geometalúrgicas encontradas por los métodos de Clustering Geoestadístico y Clustering Multivariado tradicional; en el último caso, usando distancia Euclideana con y sin uso de coordenadas espaciales como variables auxiliares. Para determinar la pertenencia a una u otra unidad, se utilizó la máxima frecuencia obtenida del conjunto de realizaciones. Es posible observar que la clasificación obtenida por el Clustering Geoestadístico identificó la frontera entre las unidades geometalúrgicas reales, no obstante existen unas pocas observaciones mal clasificadas, las que pudieran corregirse aumentando el número de simulaciones buscando la convergencia del método. En el caso de los métodos de clustering tradicionales, sin usar las coordenadas podemos apreciar que no se identificaron las unidades geometalúrgicas reales y la clasificación carece del concepto de continuidad aparentando en mayor parte un campo aleatorio. Usando las coordenadas se observa cierta continuidad espacial en las unidades creadas, sin embargo, el error de clasificación es alto y no se aprecia la frontera que divide las unidades geometalúrgicas reales.



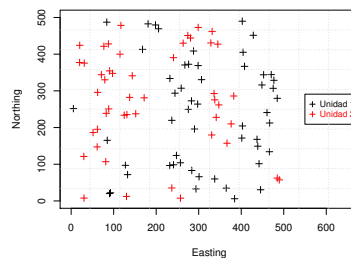
(a) U.G. Reales



(b) U.G. Clust. Geoestadístico



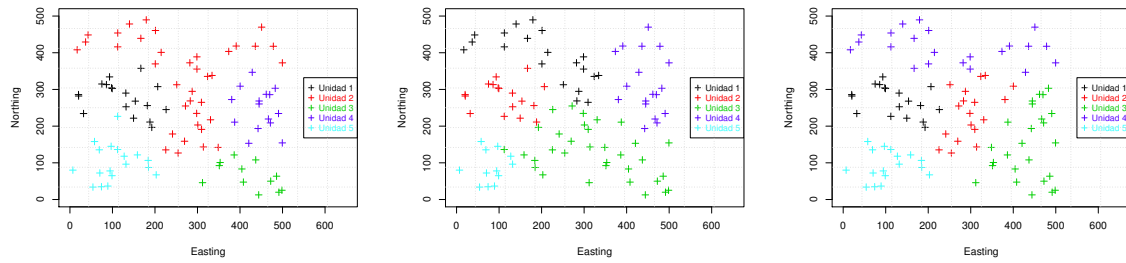
(c) U.G. Clust. sin coordenadas



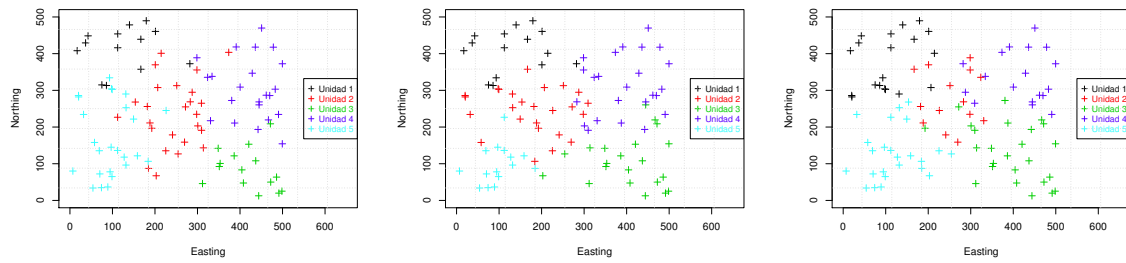
(d) U.G. Clust. con coordenadas

Figura 3.8: Resultado del clustering jerárquico con la Distancia Geoestadística (arriba, derecha) y con la Distancia Euclideana (abajo), en base a las 100 realizaciones utilizando la máxima frecuencia como clasificador.

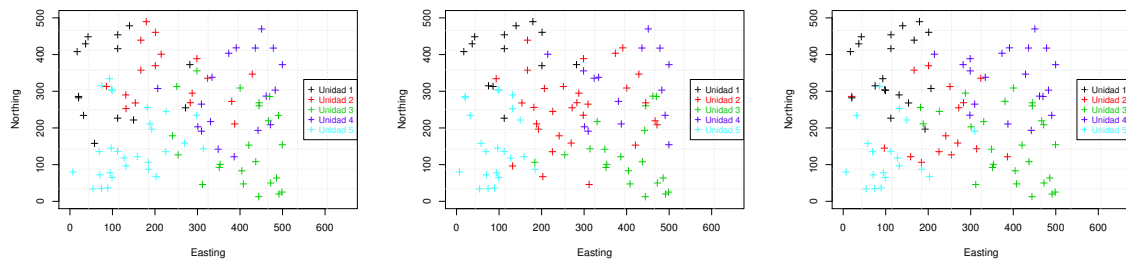
3.3.2.2. Resultados caso secundario



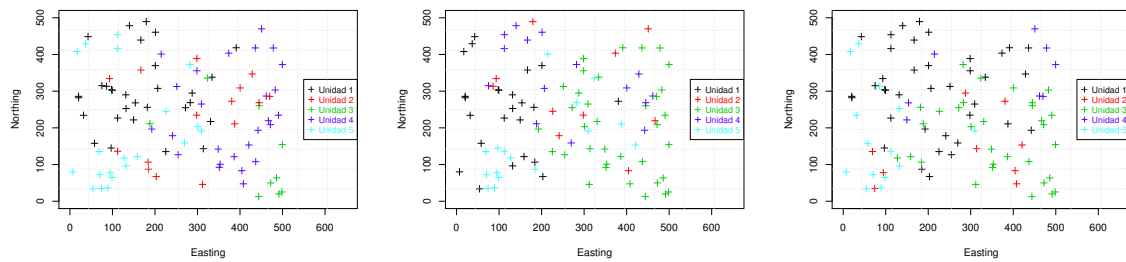
(a) $d_{max} = 100$



(b) $d_{max} = 200$

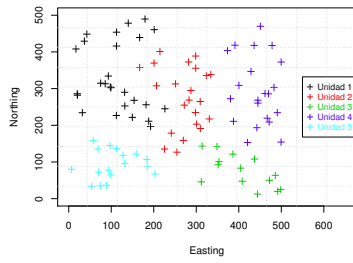


(c) $d_{max} = 300$

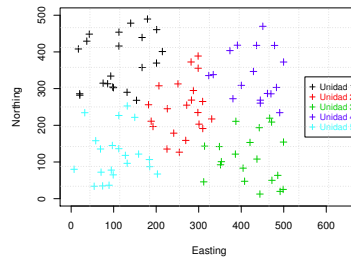


(d) $d_{max} = 400$

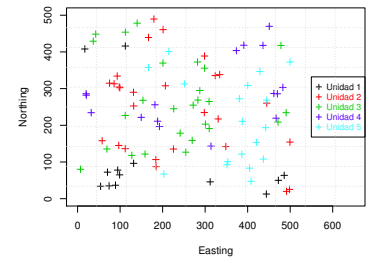
Figura 3.9: Resultados del clustering jerárquico usando la Distancia Geoestadística y el Método de Ward para tres simulaciones del caso secundario.



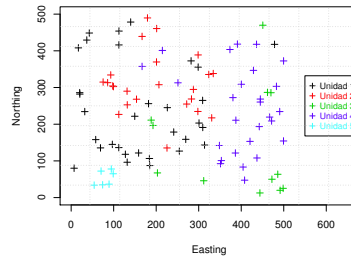
(a) U.G. Reales



(b) U.G. Clust. Geoestadístico



(c) U.G. Clust. sin coordenadas



(d) U.G. Clust. con coordenadas

Figura 3.10: Resultado del clustering jerárquico con la Distancia Geoestadística (arriba, derecha) y con la Distancia Euclideana (abajo), en base a las 100 realizaciones usando la máxima frecuencia como clasificador.

Podemos observar los clusters formados para cada una de las realizaciones del caso secundario en la Figura (3.9). Se destaca que a medida que el valor de d_{max} aumenta, se pierde conectividad en la formación de clusters, al igual que en el caso inicial. Esta información nos permite hacer una estimación del valor d_{max} que permita la formación de clusters más conexos, por lo que las figuras siguientes, muestran el resumen de los resultados para $d_{max} = 200$. La Figura (3.10) muestra el resumen de las 100 realizaciones obtenidas en el caso secundario y los alcances de las unidades geometalúrgicas encontradas por los métodos de Clustering Geoestadístico y Clustering Multivariado tradicional; en el último caso, usando distancia Euclideana con y sin uso de coordenadas espaciales como variables auxiliares. Para determinar la pertenencia a una u otra unidad, se utilizó la máxima frecuencia obtenida del conjunto de realizaciones. Con respecto a la clasificación, podemos observar que el método de Clustering Geoestadístico fue el que tuvo la mayor cantidad de observaciones bien clasificadas, pudiendo identificar la distribución espacial de las unidades geometalúrgicas reales, no obstante, tuvo problemas al identificar las fronteras entre tales unidades. En el caso de los métodos tradicionales de clustering, ya sea con y sin las coordenadas, se repiten las conclusiones obtenidas en el caso inicial; utilizando las coordenadas como variables auxiliares se obtiene mayor continuidad en las unidades descubiertas en comparación a no utilizarlas, sin embargo, las unidades descubiertas no concuerdan con las unidades geometalúrgicas reales y aparentan un campo de clasificación aleatorio que no identifica las fronteras entre las unidades. En este caso, como hay un mayor número de unidades geometalúrgicas, existe un mayor contacto (número de fronteras) entre unidades, por lo que aumenta la variabilidad de clasificación de

una simulación a otra, resultando en un mayor número de observaciones mal clasificadas. Este último comentario se desarrolla en detalle en las conclusiones.

3.4. Caso de estudio real: Geoquímica sector Colchane, I Región de Tarapacá

3.4.1. Descripción de la zona de estudio

La zona de estudio se sitúa en Colchane, comuna y pueblo de Chile de la Provincia del Tamarugal, correspondiente a la I Región de Tarapacá (Figura 3.11). Corresponde a un sector cordillerano, donde se desarrollan proyectos de pequeña y mediana minería tales como Santa Rita, San Enrique, Cerro Colorado, entre otros. Estos proyectos son en su mayoría a cielo abierto, de tipo vetiforme con menas primarias de cobre, molibdeno y plata, de edades máximas correspondientes al paleoceno eoceno inferior, compuesto por vetas y stockwork de alteraciones Fílica, Argílica y Propilítica. Además de estos, se encuentran en fase de estudio otros proyectos de tipo pórfido cobre-molibdeno, con minerales de tipo Calcopirita, Molibdenita, Calcosina, Crisocola y Blenda, entre otros. Se identifican tres zonas de alteración hidrotermal en andesita: Clorita-Epidota-Turmalina, Pirita periférica y Sericita-cuarzo-arcillas intermedia gradando hacia el interior.

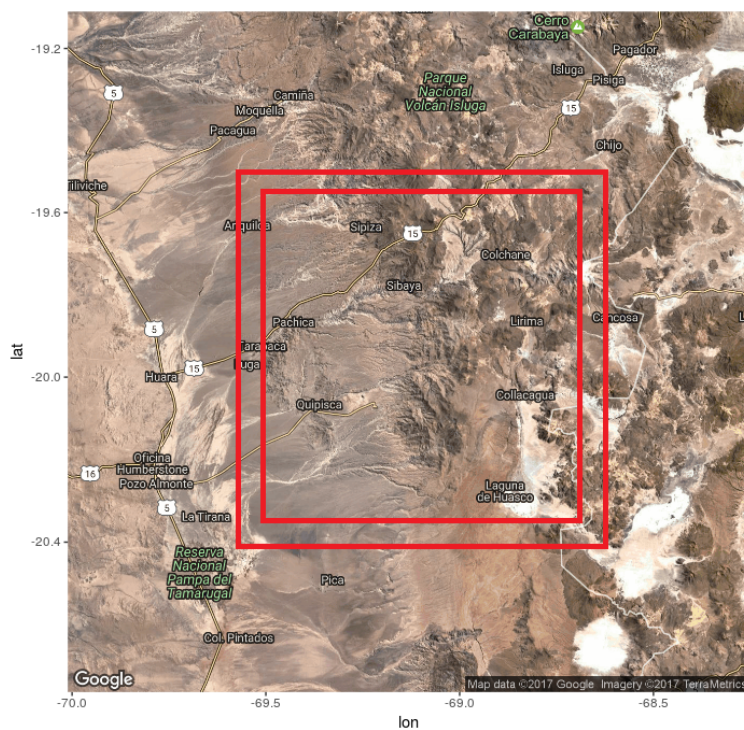


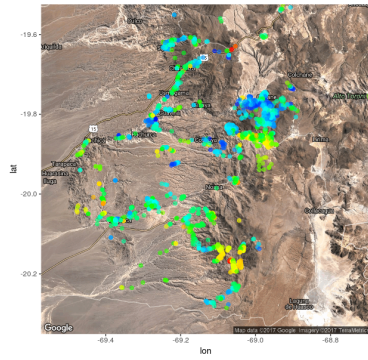
Figura 3.11: Zona de estudio, sector Colchane, I Región de Tarapacá.

3.4.2. Descripción de la base de datos

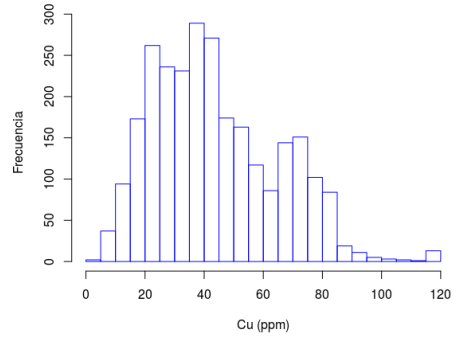
La base de datos corresponde a zonas de muestreo superficial de variables geoquímicas, acompañadas de estudios magnetométricos y gravimétricos. Entre las variables geoquímicas se encuentran los contenidos de Cobre, Molibdeno, Plata, Arsénico y Plomo en partes por millón. Se cuenta además con el contenido de Oro en partes por billón, valores RTP provenientes de la magnetometría y valores Bouger provenientes de la gravimetría del suelo, tal y como se muestra en la Tabla 3.3. El “valor RTP ” (Método de Reducción al Polo) cooresponde a un filtro aplicado para minimizar el ruido direccional causado por la latitud geomagnética baja (Phillips (1997)). Como el área bajo estudio se encuentra en una región de baja latitud, los valores positivos indican la posición de fuentes más fuertemente magnéticas. El análisis del campo magnético anómalo reducido al polo muestra los conjuntos de fuentes en su posición física verdadera. El “valor Bouger” corresponde a una corrección sobre la anomalía gravitatoria in situ. Una anomalía gravitatoria se define como la diferencia entre el valor de gravedad observado en un determinado lugar del planeta y la gravedad teórica (Hofmann-Wellenhof y Moritz (2006)). Una anomalía positiva de gravedad indica la presencia de un cuerpo con exceso de masa respecto a la masa del modelo de referencia. Es posible apreciar que existen distribuciones unimodales, como en el caso del Molibdeno (Fig. 3.13) y el Plomo (Fig. 3.16), bimodales como en el caso del Cobre (Fig. 3.12), Plata (Fig. 3.14) y el Arsénico (Fig. 3.15) y distribuciones multimodales como en el caso del contenido de Oro (Fig. 3.17), los valores RTP y Bouger (Fig. 3.18 y 3.19 respectivamente). Este hecho justifica la necesidad de realizar un análisis de clúster y definición de unidades geológicas con características específicas.

	Este	Norte	Cu (ppm)	Mo (ppm)	Ag (ppm)
Mínimo	450100	7757000	2.6010	0.9547	0.0467
1er Cuartil	471800	7782000	27.4000	3.6470	0.1437
2do Cuartil	490200	7801000	40.2200	4.2820	0.1735
Media	486000	7795000	43.6700	4.2520	0.1776
3er Cuartil	500800	7808000	58.0800	4.8330	0.2034
Máximo	511600	7843000	118.7000	8.2370	0.3119
Des. est.	15460.6295	17208.2129	20.8551	0.9540	0.0531
	As (ppm)	Pb (ppm)	Au (ppb)	Valor RTP	Valor Bouger
Mínimo	4.0560	1.1830	0.6841	-308.5000	-30.3800
1er Cuartil	17.7500	7.2200	3.3460	-103.0000	-8.0850
2do Cuartil	23.9700	9.4620	5.3680	-74.6400	-0.3709
Media	23.0900	9.6290	5.3050	-51.7200	0.8132
3er Cuartil	28.5500	11.4900	6.9460	-10.2900	7.9000
Máximo	44.0600	22.0000	10.3000	121.6000	51.3400
Des. est.	7.3987	3.1484	2.2129	65.8534	13.1579

Tabla 3.3: Estadísticas descriptivas para variables en la base de datos.

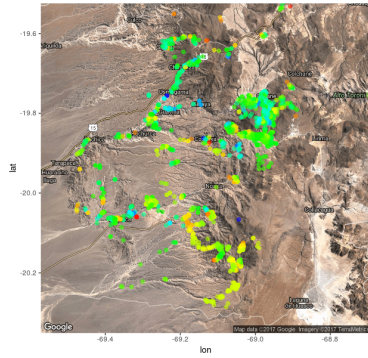


(a) Distribución espacial para la variable Cu (ppm)

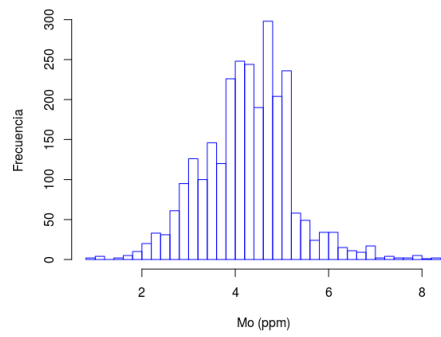


(b) Distribución frecuencial para la variable Cu (ppm)

Figura 3.12: Distribución espacial y frecuencial para la variable Cu (ppm).

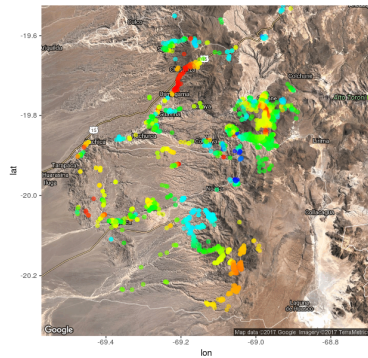


(a) Distribución espacial para la variable Mo (ppm)

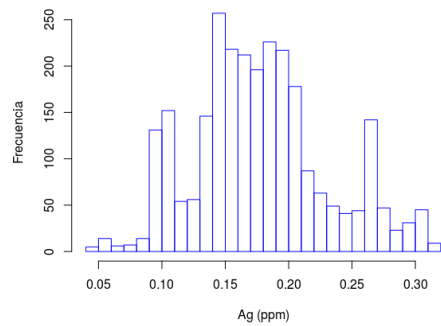


(b) Distribución frecuencial para la variable Mo (ppm)

Figura 3.13: Distribución espacial y frecuencial para la variable Mo (ppm).

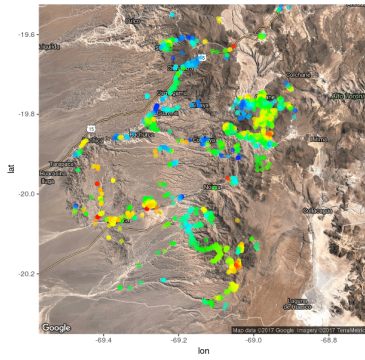


(a) Distribución espacial para la variable Ag (ppm)

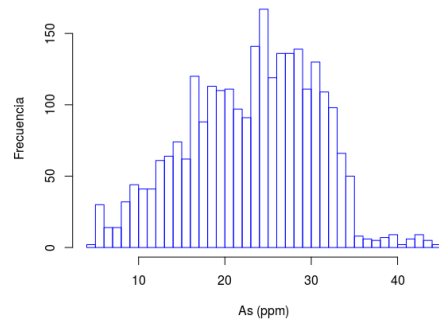


(b) Distribución frecuencial para la variable Ag (ppm)

Figura 3.14: Distribución espacial y frecuencial para la variable Ag (ppm).

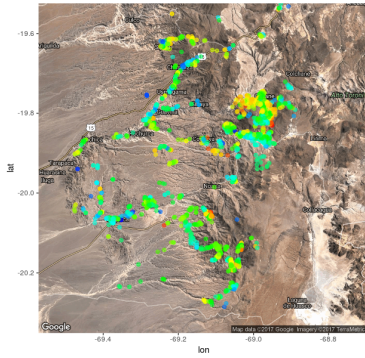


(a) Distribución espacial para la variable As (ppm)

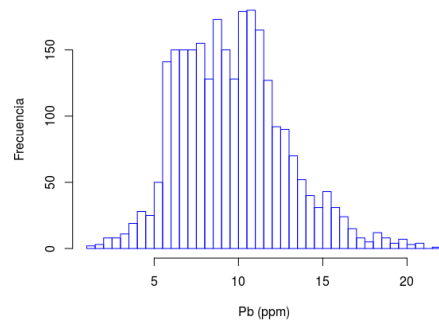


(b) Distribución frecuencial para la variable As (ppm)

Figura 3.15: Distribución espacial y frecuencial para la variable As (ppm).

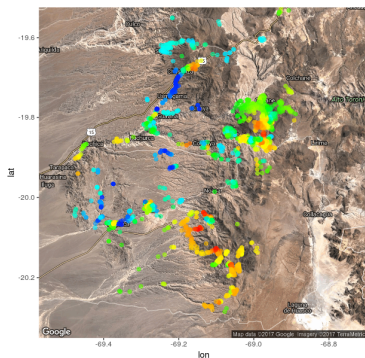


(a) Distribución espacial para la variable Pb (ppm)

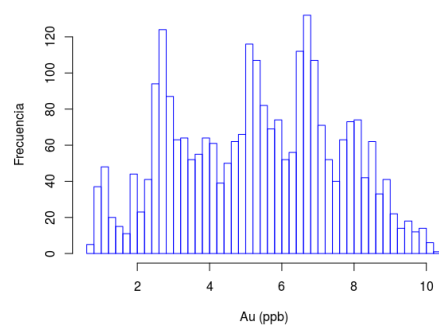


(b) Distribución frecuencial para la variable Pb (ppm)

Figura 3.16: Distribución espacial y frecuencial para la variable Pb (ppm).

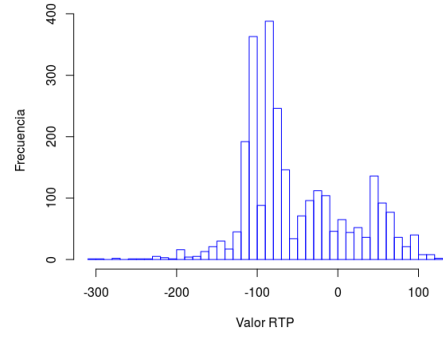
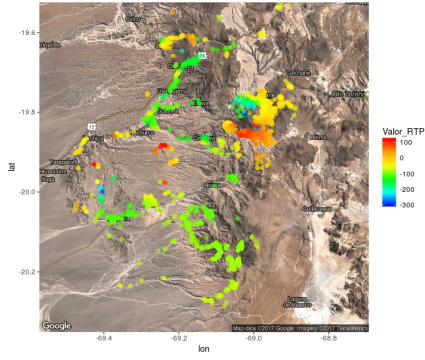


(a) Distribución espacial para la variable Au (ppb)



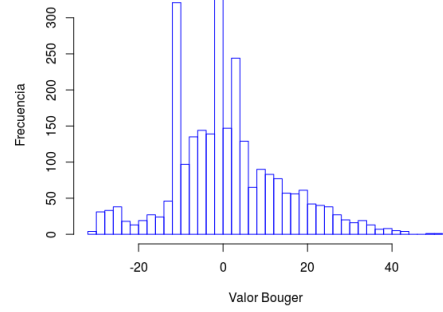
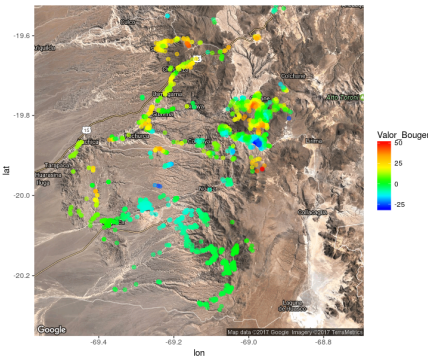
(b) Distribución frecuencial para la variable Au (ppb)

Figura 3.17: Distribución espacial y frecuencial para la variable Au (ppb).



(a) Distribución espacial para la variable RTP (b) Distribución frecuencial para la variable RTP

Figura 3.18: Distribución espacial y frecuencial para la variable RTP.



(a) Distribución espacial para la variable Bouger (b) Distribución frecuencial para la variable Bouger

Figura 3.19: Distribución espacial y frecuencial para la variable Bouger.

3.4.3. Conocimiento geológico del área

Se cuenta con información diversa acerca de características geológicas de la zona de estudio: Período de formación (Figura 3.20), Época geológica (Figura 3.21) y Tipos de roca (Figura 3.22). Estas descripciones servirán para validar los resultados obtenidos con el algoritmo de Clustering Geoestadístico propuesto al tratar de encontrar alguna relación entre los clústers formados utilizando las variables geoquímicas presentes.

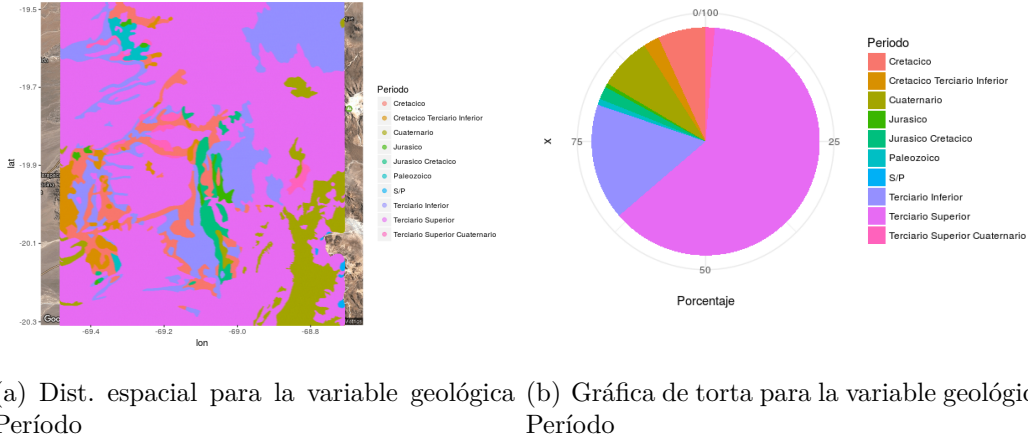
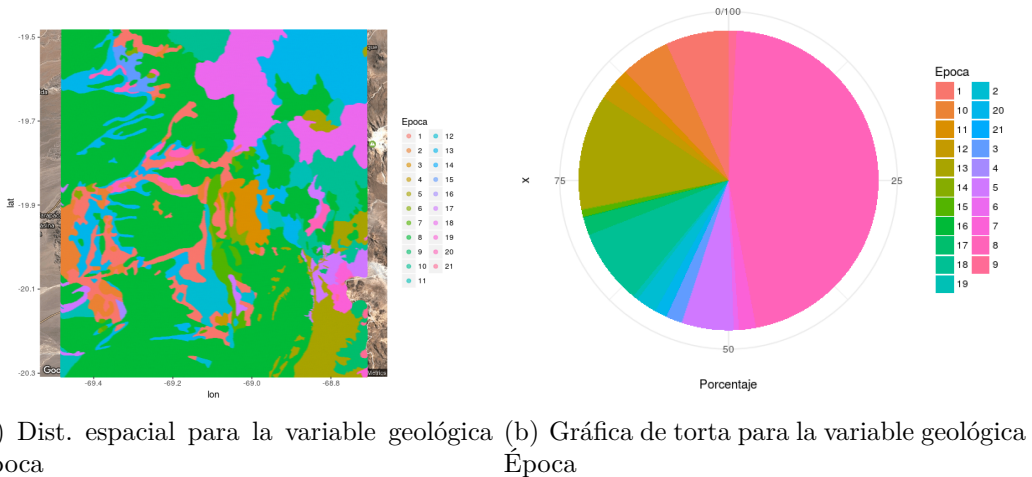
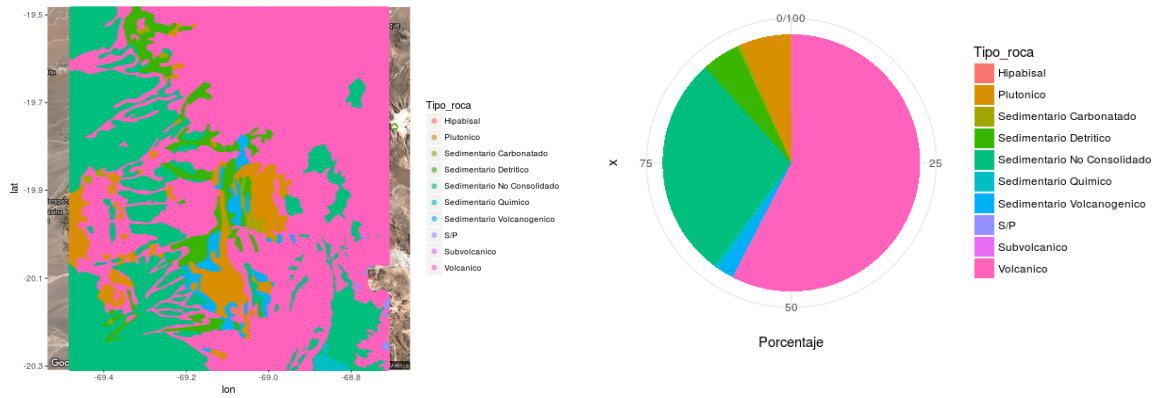


Figura 3.20: Distribución espacial y frecuencial para la variable geológica Período.



- | | | |
|--|---|---|
| [1] Cretacico Superior | [2] Cretacico Superior Paleoceno Inferior | [3] Eoceno Superior |
| [4] Eoceno Superior Oligoceno Inferior | [5] Holoceno | [6] Juracico Inferior Juracico Superior |
| [7] Juracico Superior Cretacico Inferior | [8] Mioceno Inferior Mioceno Medio | [9] Mioceno Medio |
| [10] Mioceno Superior | [11] Mioceno Superior Plioceno Inferior | [12] Oligoceno Inferior |
| [13] Oligoceno Superior Mioceno Inferior | [14] Paleoceno Superior | [15] Paleozoico Inferior |
| [16] Paleozoico Superior | [17] Pleistoceno | [18] Plioceno |
| [19] Plioceno Inferior | [20] Plioceno Superior Pleistoceno | [21] SE |

Figura 3.21: Distribución espacial y frecuencial para la variable geológica Época.



(a) Dist. espacial para la variable geológica Tipo de roca (b) Gráfica de torta para la variable geológica Tipo de roca

Figura 3.22: Distribución espacial y frecuencial para la variable geológica Tipo de roca.

3.4.4. Aplicación

Para iniciar el algoritmo de Clustering Geoestadístico es necesario realizar un análisis variográfico de las variables geoquímicas (Cu, Mo, Ag, As, Pb, Au), magnetométricas (RTP) y gravimétricas (valor Bouger), el que servirá como punto de partida para calcular la medida de disimilaridad que utilizaremos en el algoritmo jerárquico. En este análisis es necesario incluir los variogramas directos y cruzados (Figura 3.23), para asegurar el máximo uso de la información disponible.

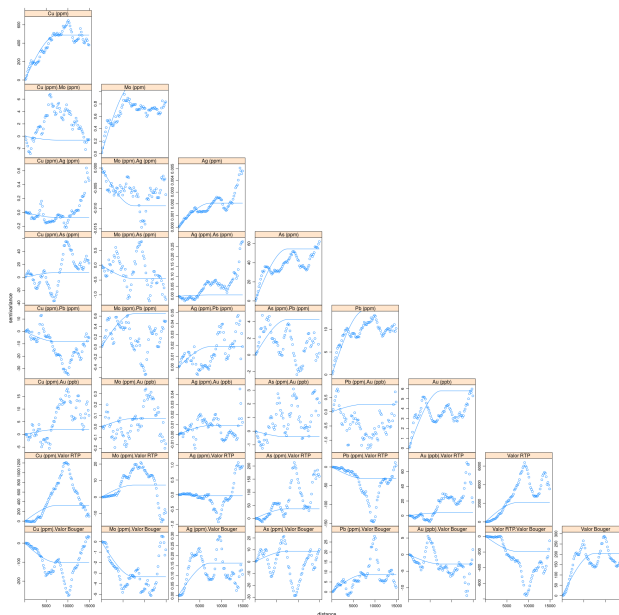


Figura 3.23: Cálculo y ajuste del modelo de semivariograma.

Luego de realizado este proceso, se calculan las disimilitudes para todos los pares de datos, utilizando la distancia geoestadística propuesta, lo que origina una matriz que toma en cuenta la correlación espacial de las variables. Se propone formar los clústers utilizando el criterio de Ward en primera instancia, debido a que en la práctica fue el que presentó resultados coherentes con las descripciones geológicas y mostró resultados valiosos en términos de conectividad y continuidad espacial. En el Anexo A se encuentran los mismos resultados utilizando el procedimiento de la máxima distancia para realizar la comparación en caso de ser requerido. Es posible observar observando la Figura 3.24, que los resultados poseen un grado de conectividad espacial bajo, independiente del número de clústers formados. Este resultado se condice con la distribución irregular de las variables geológicas, las que presentan formas diseminadas en el espacio.

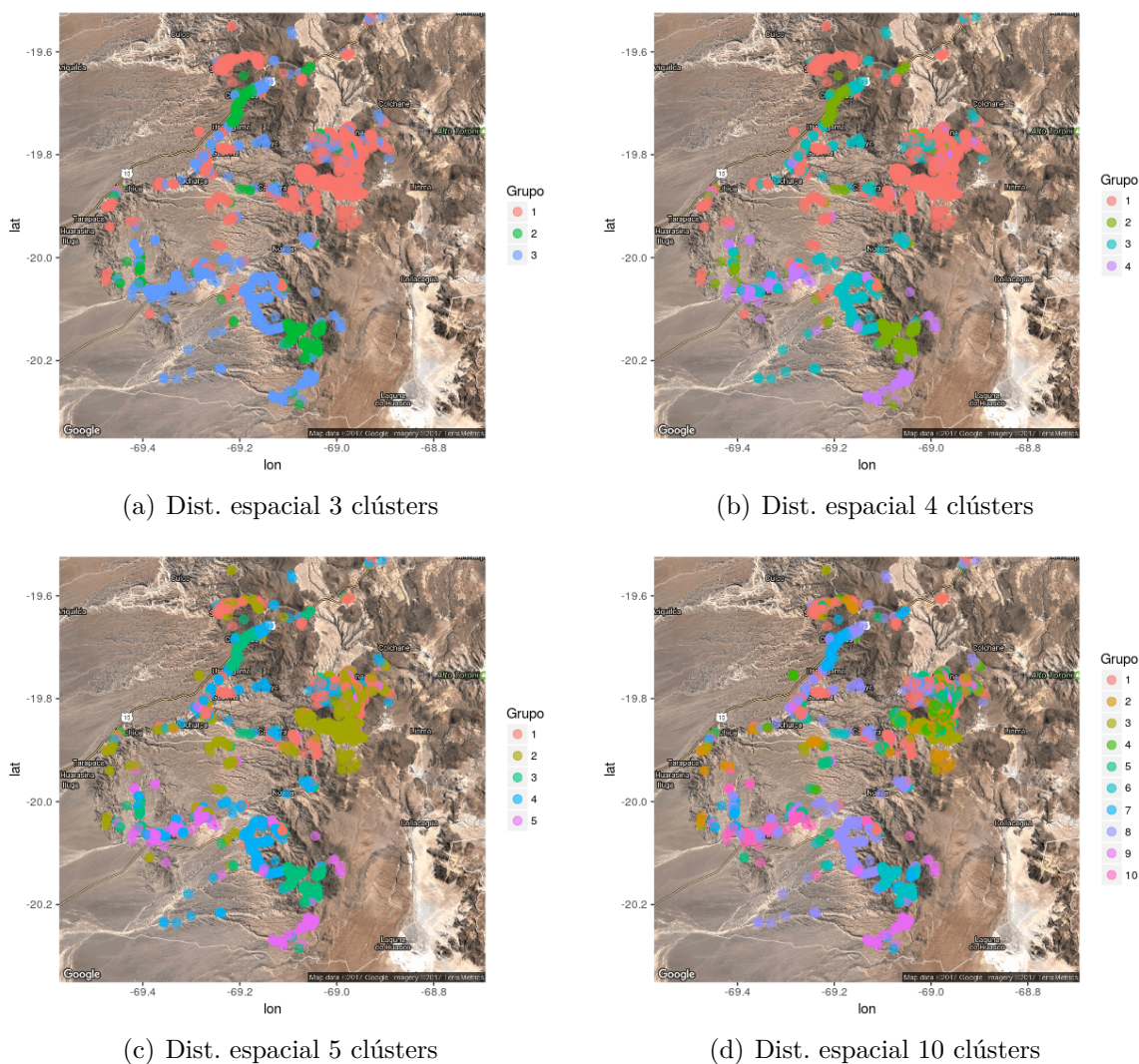


Figura 3.24: Distribución espacial para los clústers formados por el Método de Ward.

3.4.5. Resumen de resultados

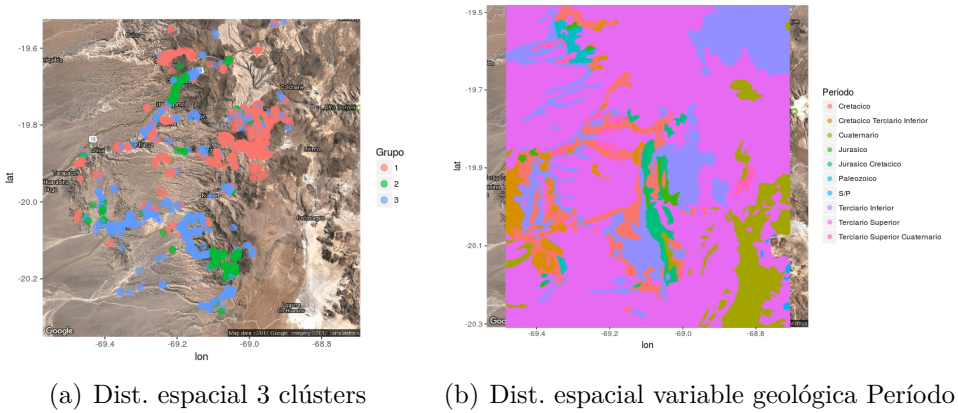


Figura 3.25: Asociación de conglomerados con el Período geológico.

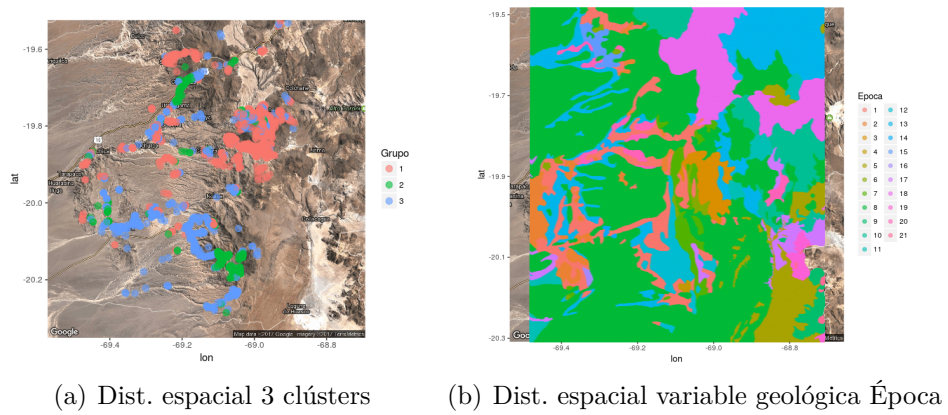


Figura 3.26: Asociación de conglomerados con la Época geológica.

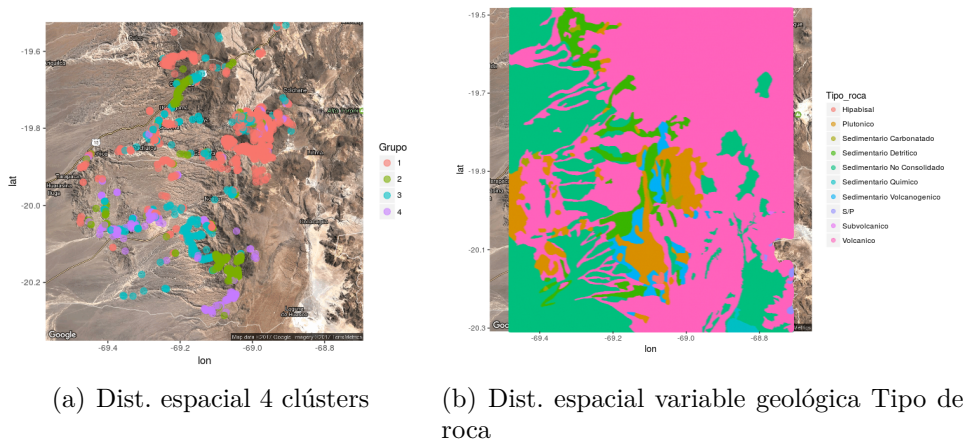
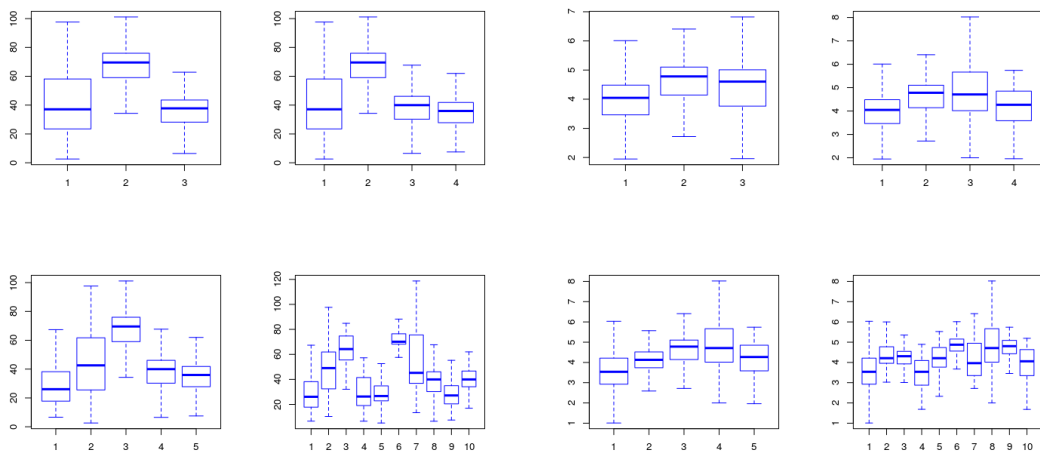


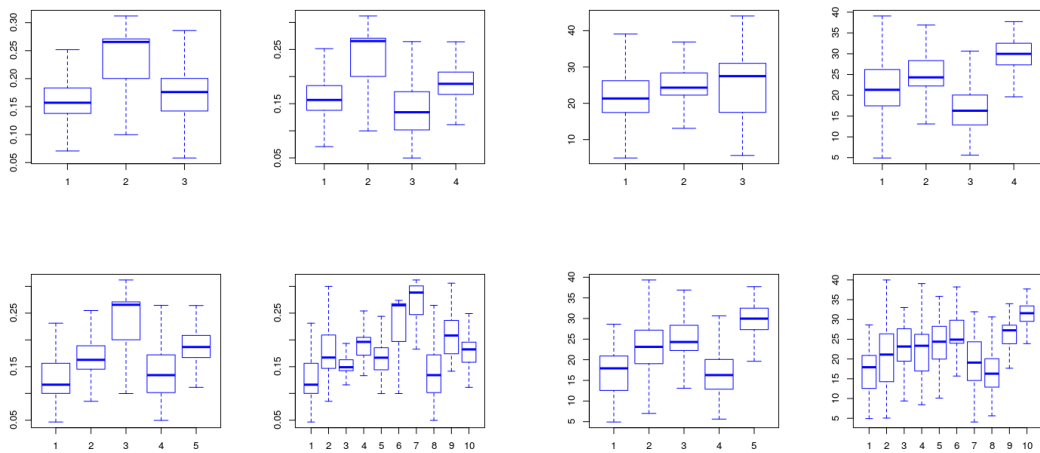
Figura 3.27: Asociación de conglomerados con el Tipo de roca.

Es posible apreciar una correspondencia entre los 3 conglomerados formados y el Período geológico de la zona de estudio (Figura 3.25). El conglomerado 1 se asocia en gran parte con el período Terciario inferior, el conglomerado 2 con el Cretácico y el conglomerado 3 con el Terciario superior y superior cuaternario en su mayoría. De la misma forma los 3 conglomerados formados se asocian con la Época geológica (Figura 3.26), donde es posible observar la correspondencia entre el conglomerado 1 con las épocas del Eoceno Superior ([3]) y del Eoceno Superior Oligoceno Inferior ([4]). El conglomerado 2 se asocia con el Cretacico Superior ([1]) y el conglomerado 3 con las épocas del Oligoceno Superior Mioceno Inferior ([13]) y el Paleoceno Superior ([14]).



(a) Distribución del Cu (ppm) en los conglomerados formados

(b) Distribución del Mo (ppm) en los conglomerados formados

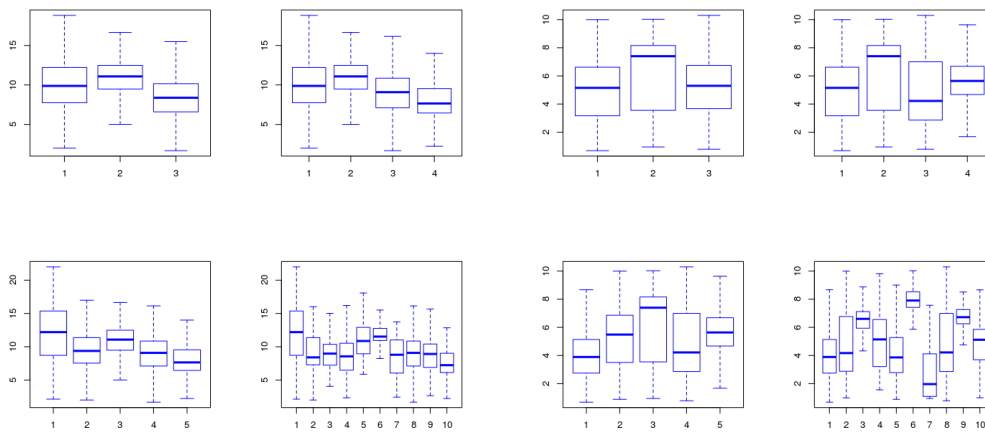


(c) Distribución del Ag (ppm) en los conglomerados formados

(d) Distribución del As (ppm) en los conglomerados formados

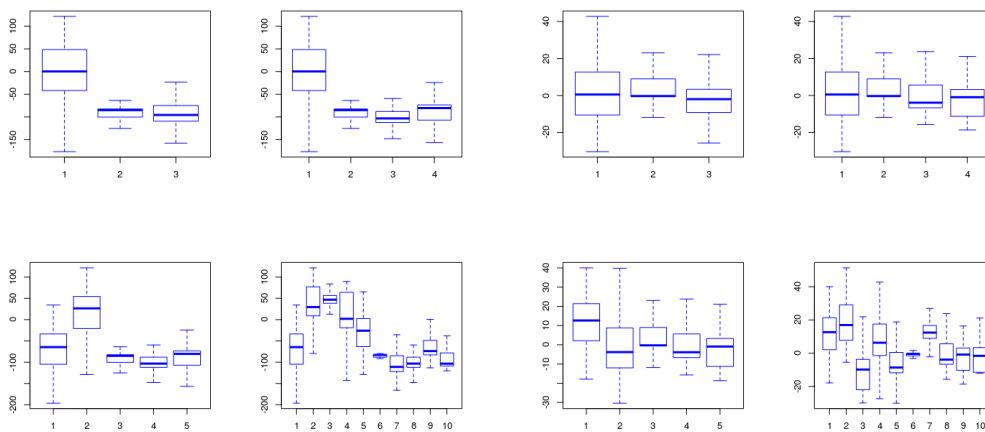
Figura 3.28: Variables geoquímicas Cu (ppm), Mo (ppm), Ag (ppm) y As (ppm) por conglomerado formado.

Para la variable geológica Tipo de roca (Figura 3.27), podemos observar una correspondencia entre los tipos: Plutónico con el conglomerado 1, Sedimentario Detrítico con el conglomerado 2, Sedimentario Volcanogénico con el conglomerado 3 y Volcánico con el conglomerado 4. Si bien estas asociaciones no están claramente demarcadas, es posible observar un comportamiento general entre los conglomerados y las variables geológicas disponibles. La Figura 3.28 refleja el comportamiento de las variables geoquímicas Cu (ppm), Mo (ppm), Ag (ppm) y As (ppm) para cada conglomerado formado. El posible apreciar que mientras menor sea número de conglomerados, las diferencias entre las cajas es menos significativa, por lo que se debe hacer un estudio detallado del número óptimo de conglomerados a formar, teniendo en cuenta la distribución frecuencial y espacial de tales.



(a) Distribución del Pb (ppm) en los conglomerados formados

(b) Distribución del Au (ppb) en los conglomerados formados



(c) Distribución del Valor RTP en los conglomerados formados

(d) Distribución del Valor Bouguer en los conglomerados formados

Figura 3.29: Variables geoquímicas Pb (ppm) y Au (ppb), magnetométrica RTP y gravimétrica Bouguer por conglomerado formado.

Para las variables Cu (ppm) (Figura 3.28(a)) y Ag (ppm) (Figura 3.28(c)) se observan diferencias significativas de las distribuciones para cada conglomerado, no así para las variables geoquímicas Mo (ppm) (Figura 3.28(b)) y As (ppm) (Figura 3.28(d)). Para las variables Pb (ppm), Au (ppb) y valor Bouger (Figuras 3.29(a), 3.29(b) y 3.29(d) respectivamente) no se observan diferencias significativas en las distribuciones tanto para 3, 4 e incluso 5 conglomerados. La única variable que presenta una diferencia notable para cada conglomerado es la variable magnetométrica RTP (Figura 3.29(c)).

3.5. Conclusiones parciales

El Análisis de Conglomerados Geoestadístico permitió descubrir las Unidades Geológicas presentes en la región de estudio. Es aconsejable la inclusión de las variables geológicas de manera directa en la construcción de la Distancia Geoestadística, la que permitirá obtener mayor conocimiento y control sobre la forma en cómo se generan las unidades por el método de Clustering Jerárquico. También es posible incorporar este tipo de información en términos de variogramas de indicadores. La pregunta acerca del número ideal de unidades sigue sin respuesta única, número tal que el experto (geólogo o geometalurgista) sea capaz de interpretar las unidades descubiertas y relacionarlas con características tales como litología, mineralización o alteración. Se puede incluir las coordenadas como variables o excluir las coordenadas y perder información usando los algoritmos tradicionales de agrupamiento. En el caso del Análisis de Conglomerados Geoestadístico, las coordenadas están incluidas en el cálculo de la matriz de distancias, via los variogramas directos y cruzados, razón por la cual la partición formada presentó cierta continuidad espacial, que fue era uno de los objetivos de la propuesta. La existencia de continuidad espacial, la falta de información exhaustiva que provoca incertidumbre y la interpretación de variables regionalizadas en términos de funciones aleatorias, hacen que se desaconseje el uso de métodos tradicionales de clustering. Observando los resultados del caso sintético, nos damos cuenta que existe un gran número de fuentes de variabilidad que condicionan los resultados. Todo parte con el proceso de simulación, pues es ahí donde se genera el supuesto de independencia entre las unidades geometalúrgicas. Cada una tiene una estructura de correlación espacial, que se repite en todas. Esto no es algo que sea 100 % realista, pues es posible que ciertas áreas del dominio tenga comportamientos, alcances y comportamientos isótropos o anisótropos diferentes. Es recomendable en el futuro considerar este punto y asignar distintas estructuras de correlación espacial, una por cada grupo o unidad. También existe una gran fuente de variabilidad en el proceso de ajuste de los variogramas directos y cruzados por el modelo de correogionalización lineal, que se relaciona con las distancias entre cada par de observaciones usando la distancia geoestadística propuesta. El concepto de uso de conocimiento geológico, hace referencia al uso de información de tipo cualitativa, ya sea nominal u ordinal. Se usó como conocimiento y a la vez como restricción, una distancia geográfica máxima de pertenencia a la misma unidad geometalúrgica en el caso de datos simulados, concepto empleado por ejemplo, cuando se delimitan dominios geológicos u otras variables relacionadas con las características del dominio. Si bien es posible hacer la modificación de la matriz de distancias como se propuso durante el proyecto, es aconsejable la inclusión de variables categóricas, binarias o con más de una clase, para poder construir una nueva distancia de disimilaridad s_{ij} , la que permitirá

obtener mayor conocimiento y control sobre la forma de cómo se generan las unidades por el método de clustering jerárquico. Con respecto a la distancia d_{max} presentada en los resultados, podemos notar que tanto para el caso inicial como secundario, se obtuvieron los mejores resultados en términos de conectividad e identificación de las unidades geometalúrgicas originales para una distancia $d_{max} = 200$. Los resultados, nos hacen pensar que el parámetro d_{max} está relacionado con el rango de las coordenadas y con el número de unidades que se quieren descubrir. También podemos concluir que a mayor número de fronteras en común, mayor será el error de clasificación a la hora de definir las fronteras que separan las unidades geometalúrgicas.

Capítulo 4

Conglomerados basados en mezcla de distribuciones

4.1. Introducción

Las mezclas de distribuciones han servido para resolver problemas de identificabilidad entre sujetos, los que son caracterizados por una función de probabilidad propia de cada grupo y que se rige por parámetros propios (Bishop (2006)). Dependiendo del enfoque, estos modelos de mezclas pueden ser utilizados para resolver problemas de clasificación en caso de contar con datos ausentes, los que son imputados en base al conjunto de correlaciones existentes, o para resolver problemas de agrupamiento en caso de querer encontrar fronteras no observables que separen individuos en base a sus características.

Existen dos enfoques principales cuando se trabaja con modelos de mezclas de distribuciones: enfoque clásico y enfoque bayesiano (Diebolt y Robert (1994)). Para el primero de ellos el objetivo es encontrar la configuración óptima de asignación de pertenencia de cada individuo a cada grupo que maximice la función de verosimilitud del modelo. Para esto, es necesario encontrar los estimadores de los parámetros de cada mezcla, lo que requiere un procedimiento iterativo de estimación y maximización, lo que usualmente recae sobre variantes del algoritmo EM (Bilmes y cols. (1998)). El segundo enfoque se basa en la idea de que los parámetros que describen cada mezcla son considerados como variables aleatorias, las que deben ser encontradas para luego interpretarlas en base a sus momentos muestrales. Este proceso es de carácter iterativo al igual que en el caso anterior; la diferencia fundamental recae sobre el método empleado para tal objetivo, pues en este caso se emplean algoritmos tales como muestreador de Gibbs (Kozumi y Kobayashi (2011)) o Metrópolis Hastings (Chib y Greenberg (1995)). Estos algoritmos producen una muestra condicional de los parámetros que rigen cada mezcla, que es utilizada para construir una muestra aleatoria de la distribución de probabilidad de cada parámetro de interés a posteriori.

Los modelos de mezclas son sencillos de implementar, pues no requieren operaciones complejas ni tampoco funciones especiales para su desarrollo. En el caso de los métodos basados en verosimilitud, para el caso Gaussiano, estos son sencillos de obtener y tienen una forma exacta, lo que hace que la carga computacional se vea reducida, entregando soluciones en corto tiempo.

Como desventaja, el uso de otras distribuciones (Ji y cols. (2005), MacEachern y Müller (1998)) se encuentra reducido por el hecho de que no es posible obtener los estimadores de máxima verosimilitud de manera analítica, por lo que usualmente se utilizan técnicas de optimización numérica para maximizar la función de verosimilitud. Estos modelos están condicionados al supuesto de independencia entre los elementos de la muestra, a los que llamamos observaciones o sujetos, supuesto que no afecta a la relación existente entre los atributos medidos para estos elementos. Por lo que en el caso de no contar con una muestra aleatoria de una población, se hace imposible el uso de los modelos de mezclas y una aplicación indebida puede llevar a conclusiones erróneas.

Dentro del contexto de variables regionalizadas, la correlación entre los elementos que componen la muestra (localidades indexadas por el espacio con atributos medidos) es un supuesto fundamental, sin el cual no tendría sentido asumir una estructura de correlación espacial. Esta misma estructura impide trabajar con el supuesto de independencia entre los individuos que componen la muestra, lo que corresponde al gran obstáculo al utilizar modelos de mezclas en presencia de variables regionalizadas.

Una de las características de los algoritmos de mezclas es que la asignación resultante de los individuos a cada población es discreta, por lo que a pesar de haber un proceso iterativo de convergencia, la respuesta obtenida es única y sin una medida de error o incertidumbre en la asignación (Vallejos y cols. (2015)). Un resultado ideal es aquella respuesta que cuenta con esta incertidumbre o probabilidad de pertenencia, para de esa forma poder generar diferentes escenarios en base al grado de error que se está dispuesto a cometer.

En lo que sigue, se va a optar por el método de mezclas Gaussianas bajo el enfoque bayesiano, pues éste permite introducir una medida del error en la asignación tanto de los parámetros que rigen la mezcla como de la pertenencia de los individuos. Utilizando este método bajo el enfoque tradicional, se presentan restricciones al trabajar con variables regionalizadas, por lo que se deberá generar una propuesta que considere este aspecto de la manera más fidedigna y eficiente posible.

4.2. Metodología propuesta

En los modelos geoestadísticos la función aleatoria asociada a la variable regionalizada, por lo general, asume después de una eventual transformación monótona, una densidad normal

multivariante $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Ecuación 4.2.1) en las localidades x_1, \dots, x_n , donde $\boldsymbol{\mu}$ corresponde a la esperanza de la función aleatoria y $\boldsymbol{\Sigma}$ es la matriz de varianzas-covarianzas, la que presenta una estructura definida por el variograma teórico que describe la continuidad espacial de la variable regionalizada (Ribeiro Jr y cols. (2001)).

$$f(\mathbf{Y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu})\right)$$

$$\mathbf{Y} = [\mathbf{Y}(x_1), \dots, \mathbf{Y}(x_n)]^\top \quad \mathbf{Y}(x_i) \in \mathbb{R}, \forall i = 1, \dots, n \quad f(\mathbf{Y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^+ \quad (4.2.1)$$

$$\boldsymbol{\mu} = [E\{\mathbf{Y}(x_1)\}, \dots, E\{\mathbf{Y}(x_n)\}]^\top \quad \boldsymbol{\Sigma} = \begin{pmatrix} C(x_1 - x_1 | \theta) & \dots & C(x_1 - x_n | \theta) \\ \vdots & & \vdots \\ C(x_n - x_1 | \theta) & \dots & C(x_n - x_n | \theta) \end{pmatrix}$$

Bajo esta definición, la matriz de varianzas-covarianzas puede expresarse como $\boldsymbol{\Sigma} = C_Y(h | \theta)$, con $\theta = (\sigma^2, \phi)$, donde $h = x_i - x_j$ representa la distancia de separación espacial entre las localidades x_i y x_j , σ^2 corresponde a la varianza del proceso espacial y ϕ corresponde a un factor de escala del modelo de variograma (típicamente, el alcance o rango de correlación). En la Figura 4.1 es posible apreciar un ejemplo de un proceso geoestadístico con esta distribución. En el contexto de los procesos geoestadísticos, un aspecto necesario para poder caracterizar la continuidad espacial es el número de individuos¹ que componen la muestra (Kravchenko (2003)). Si se posee un número adecuado de mediciones se podrá construir una estimación confiable e insesgada del variograma (Figura 4.2). Sin embargo, a pesar que un gran número de individuos contribuye a caracterizar el proceso espacial, los algoritmos de mezclas de distribuciones se hacen menos eficientes a medida que se aumenta el número de observaciones y parámetros en la muestra, independiente del enfoque utilizado. Los algoritmos empleados involucran un gran número de operaciones matriciales por cada iteración, cuya dificultad aumenta conjuntamente con la dimensionalidad del problema.

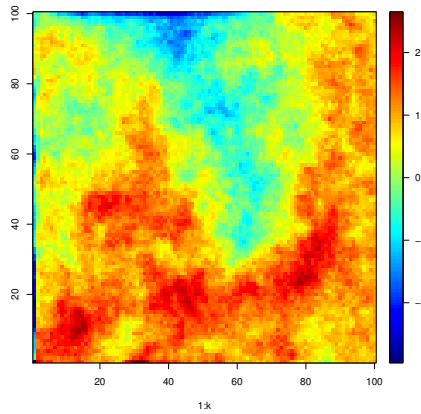
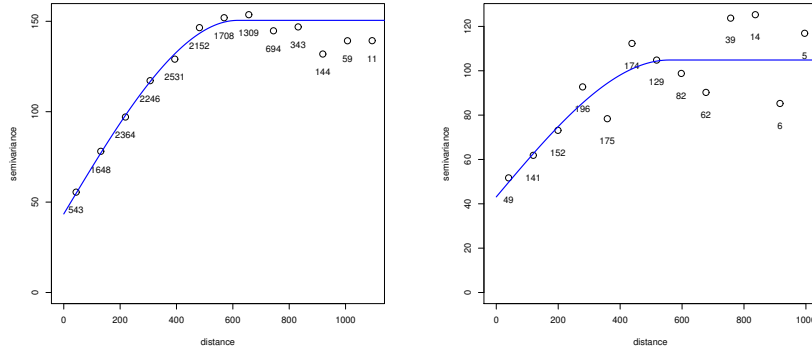


Figura 4.1: Proceso Gaussiano con media $\boldsymbol{\mu} = \mathbf{1}$, varianza del proceso espacial $\sigma^2 = 1$, alcance $\phi = 40$ y covarianza exponencial.

¹Recordar que un individuo, medición o elemento en el contexto geoestadístico corresponde a una localidad del espacio en la que se han medido un número determinado de atributos



(a) Base de datos original

(b) Muestra aleatoria de base de datos original

Figura 4.2: Influencia del número de pares de muestras en la construcción del variograma experimental.

Este impedimento hace que se requieran algoritmos mucho más eficientes, los que no se vean afectados por la dimensionalidad del proceso y que al mismo tiempo, utilicen la mayor cantidad de información disponible para describir de buena forma el proceso espacial.

Denotaremos por $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$, donde $Y_i = Y(x_i)$ corresponde a la variable aleatoria presente en la localidad $x_i \in D \subset \mathbb{R}^p$, que toma valores en el eje real. Asumiremos que este vector puede ser descompuesto en K particiones, todas ellas independientes y disjuntas

$$\begin{aligned} \mathbf{V}_1 \cup \dots \cup \mathbf{V}_K &= \mathbf{Y} \\ \mathbf{V}_1 \cap \dots \cap \mathbf{V}_K &= \emptyset \end{aligned} \quad (4.2.2)$$

con tamaños n_1, \dots, n_K y $\sum_{i=1}^K n_i = n$, las que pueden ser descritas por sus momentos muestrales de la forma

$$E[Y(x)] = \mu_k, \quad \forall Y(x) \in \mathbf{V}_k, \quad k = 1, \dots, K \quad (4.2.3)$$

$$C_k(h) = E[Y(x+h)Y(x)] - \mu_k^2, \quad \text{tal que } Y(x+h), Y(x) \in \mathbf{V}_k$$

Adicionalmente, es posible considerar un modelo de regresión con la siguiente estructura para el primer momento muestral

$$\mu_k = \mathbf{r}^\top \psi_k + \varepsilon_k \quad (4.2.4)$$

con $\mathbf{r}^\top = (1, r_1, \dots, r_p)$ el vector que contiene p variables explicativas relacionadas con la media de la variable regionalizada, ψ_k el vector de coeficientes lineales y ε_k un error aleatorio para el proceso k -ésimo. En el caso de la estructura de covarianzas, $C_k(h)$ acepta la representación $C_k(h) = \sigma_k^2 \rho(h|\phi_k)$, donde σ_k^2 corresponde a la varianza del proceso espacial y ϕ_k es el rango del proceso e indica la distancia a partir de la cual dos variables no están correlacionadas para la partición k -ésima. De esta forma, cada función aleatoria tendrá la siguiente distribución para $k = 1, \dots, K$

$$\mathbf{Y}|\mathbf{V}_k = k \sim N_{n_k}(\boldsymbol{\mu}_k, C_k), \quad k = 1, \dots, K \quad (4.2.5)$$

y la verosimilitud de la mezcla puede escribirse como

$$L(\Theta|\mathbf{Y}) = \prod_{k=1}^K f_k(\mathbf{Y}|\mathbf{V}_k = k, \boldsymbol{\theta}_k), \quad \Theta = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\} \quad (4.2.6)$$

donde $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, C_k)$ corresponde al vector de parámetros desconocidos, los cuales corresponden a las cantidades de interés a estimar (Allard y Guillot (2000)). Este supuesto se sustenta en el hecho de que se trabaja con K procesos independientes en el espacio, pero con una estructura de correlación interna.

El teorema de Bayes (Box y Tiao (2011)) refleja el conocimiento actualizado sobre los parámetros desconocidos condicionado a los datos observados y su distribución. Debido a que la densidad marginal de los datos no tiene muchas veces una forma definida, es posible hacer la simplificación

$$\pi(\Theta|\mathbf{Y}) = \frac{f(\mathbf{Y}, \Theta)}{f(\mathbf{Y})} = \frac{f(\mathbf{Y}|\Theta)\pi(\Theta)}{f(\mathbf{Y})} \implies \pi(\Theta|\mathbf{Y}) \propto L(\Theta|\mathbf{Y})\pi(\Theta) \quad (4.2.7)$$

que constituye el núcleo de la distribución a posteriori, ya que la densidad marginal $f(\mathbf{Y})$ no depende de Θ y puede ser considerada como una constante regularizadora. La cantidad $\pi(\Theta)$ corresponde a la distribución a priori de los parámetros desconocidos y representa el conocimiento preliminar que se tiene acerca de éstos. De esta forma, la cantidad a posteriori de la propuesta queda

$$\pi(\Theta|\mathbf{Y}) \propto \prod_{k=1}^K f_k(\mathbf{Y}|\mathbf{V}_k = k, \boldsymbol{\theta}_k)\pi(\boldsymbol{\theta}_k) \quad (4.2.8)$$

El objetivo del enfoque bayesiano es poder muestrear valores provenientes de esta distribución a posteriori, para de esa forma, poder tener las distribuciones a posteriori de los parámetros en ella y calcular momentos muestrales tales como la media, varianza, mediana, entre otros. Sin embargo, existe la dificultad de que para construir esta cantidad a posteriori de interés, es necesario saber a qué mezcla pertenece cada uno de los individuos de la muestra, lo que constituye el problema del agrupamiento (François y cols. (2006)).

Para lograr incorporar esta fuente de incertidumbre en el modelo, la vía directa sería agregar tantos parámetros desconocidos como datos al modelo, los que van a indicar la pertenencia de cada dato a cada mezcla, quedando una distribución a posteriori de la forma

$$\pi(\Theta|\mathbf{Y}) \propto \prod_{k=1}^K f_k(\mathbf{Y}|\mathbf{V}_k = k, \boldsymbol{\theta}_k)\pi(\boldsymbol{\theta}_k) \quad (4.2.9)$$

con $\Theta = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$ y $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, C_k, I_{1k}, \dots, I_{nk})$ para $k = 1, \dots, K$. El valor I_{ik} , para $i = 1, \dots, n$ y $k = 1, \dots, K$, indica la pertenencia de la i -ésima variable a la mezcla k -ésima y corresponde a una variable de tipo binaria (0 o 1). Esto incrementa de manera considerable el número de parámetros y hace poco eficiente cualquier algoritmo que se plantee usar. Para reducir el número de parámetros en el modelo e incorporar esta fuente de incertidumbre de manera indirecta, se propone utilizar un criterio de agrupamiento de los individuos $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ basados en algoritmos de clustering tradicionales, para de esa forma generar

un nuevo vector $\mathbf{Z} = (Z_1, \dots, Z_P)^\top$ con $P \ll n$, donde cada Z_p corresponde a un conglomerado que agrupa un número determinado de elementos del vector \mathbf{Y} . Esta idea ha sido aplicada de forma similar en otros trabajos utilizando la triangulación de Delaunay (Romary y cols. (2012)). En la Figura 4.3 es posible apreciar la forma en la que opera este criterio de agrupamiento. Estos conglomerados serán los individuos que ingresarán al modelo y a los que se evaluará su pertenencia a las mezclas $\mathbf{V}_1, \dots, \mathbf{V}_K$, quedando la nueva distribución a posteriori de la forma

$$\pi(\Theta|\mathbf{Z}) \propto \prod_{k=1}^K f_k \left(\bigcup_{p=1}^P Z_p | \mathbf{V}_k = k \right) \pi(\theta_k) \quad (4.2.10)$$

La elección del valor de P queda restringido a la expertiz del investigador, el que deberá escoger un valor suficiente para alcanzar una resolución adecuada sobre el dominio de estudio. Los conglomerados Z_p serán creados a través del algoritmo K-Medias utilizando para ello sólo la información de las coordenadas geográficas. De esta forma, los conglomerados Z_p resultantes serán dominios conexos e irregulares en el espacio.

El algoritmo K-medias depende del número de conglomerados que se quiera formar y de los centroides iniciales. De esta forma, para un mismo número de conglomerados a encontrar, es posible obtener diferentes resultados en base al punto de partida. No obstante, es posible que la definición de estos conglomerados se mantenga constante, restringiendo el número de configuraciones posibles que maximicen la verosimilitud. Para evitar este inconveniente y garantizar una correcta aleatorización de las particiones, la transformación del vector $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ en $\mathbf{Z} = (Z_1, \dots, Z_P)^\top$ será realizada en base a una grilla que cubra todo el dominio de estudio, de tipo irregular y variable en número de conglomerados (Figura 4.4(b)).

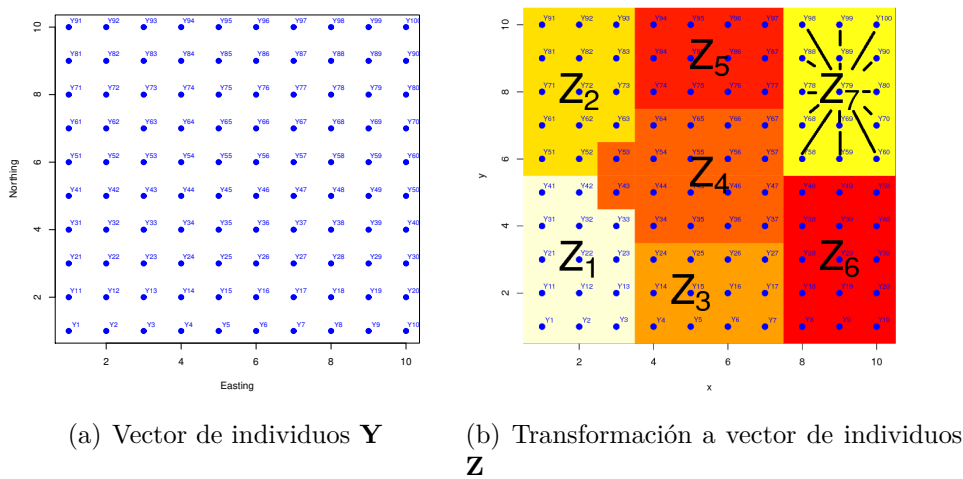


Figura 4.3: Paso preliminar utilizando algoritmo K-Medias en base a las coordenadas geográficas

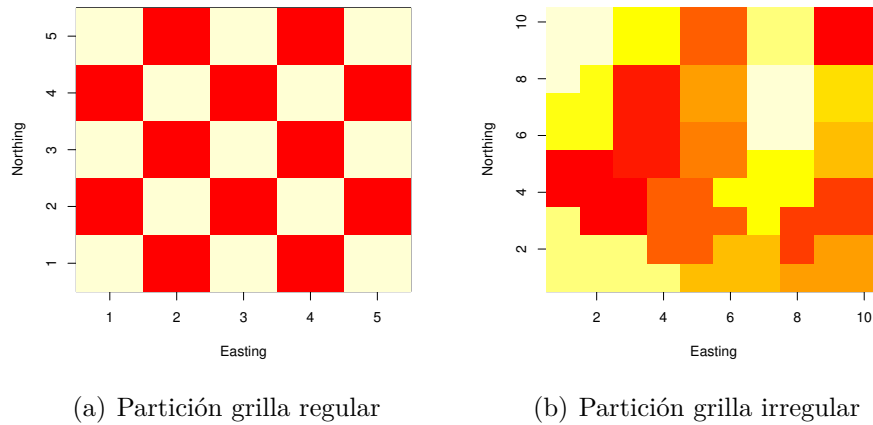


Figura 4.4: Tipos de mallado para definir las particiones $\mathbf{Z} = (Z_1, \dots, Z_P)^\top$

La grilla de tipo regular (Figura 4.4(a)) es descartada pues limita las formas posibles de las fronteras naturales entre las mezclas $\mathbf{V}_1 \dots \mathbf{V}_K$, haciendo posible una correcta caracterización sólo en caso de encontrarse fronteras regulares. En las primeras iteraciones del algoritmo escogido para maximizar la verosimilitud, se escogerá un número reducido de particiones que abarquen la totalidad de la zona de estudio, lo que permitirá reducir el número de cálculos por iteración, para luego ir aumentando con el paso de las iteraciones, para de esa forma tener un agrupamiento lo más exacto posible y mejorar la resolución de las particiones, tal y como se puede apreciar en la Figura 4.5. El procedimiento para muestrear valores de la distribución a posteriori es el algoritmo Metrópolis Hastings basado en Cadenas de Markov (Chib y Greenberg (1995)). En la propuesta se trata de simular una cadena de Markov sobre Θ de tal modo que $\{\Theta^{(1)}, \Theta^{(2)}, \dots, \Theta^{(t)}, \Theta^{(t+1)}, \dots\}$ posea una distribución estacionaria $\pi(\Theta|\mathbf{Z})$. El Algoritmo 2 muestra la forma general para realizar este procedimiento.

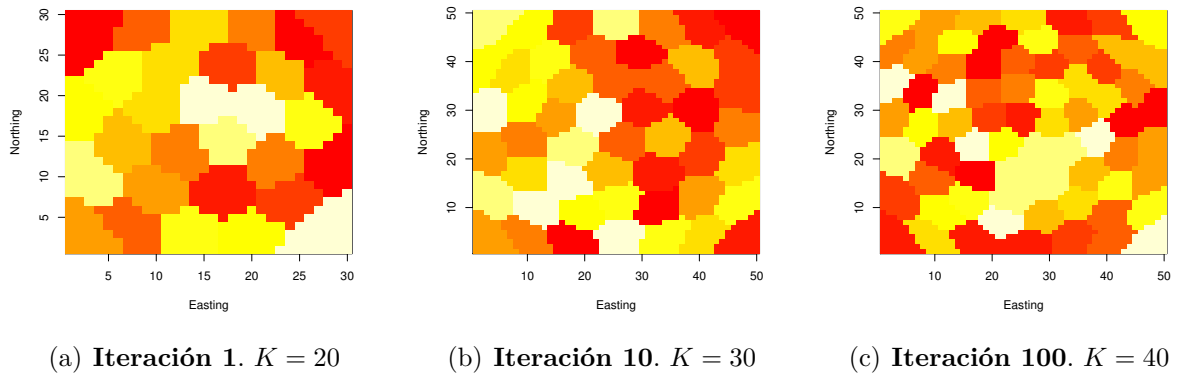


Figura 4.5: Evolución de grilla irregular durante algoritmo iterativo

Valores iniciales $\Theta^{(0)}$

⋮

Paso i Para obtener $\Theta^{(i)}$ se genera un candidato de Θ' a partir de $q(\cdot|\Theta^{(i-1)})$ y se actualiza según

$$\begin{cases} \Theta^{(i)} = \Theta' & \text{con probabilidad } \alpha(\Theta^{(i-1)}, \Theta') \\ \Theta^{(i)} = \Theta^{(i-1)} & \text{con probabilidad } 1 - \alpha(\Theta^{(i-1)}, \Theta') \end{cases}$$

Donde

$$\alpha(\Theta^{(i-1)}, \Theta') = \min \left\{ 1, \frac{\pi(\Theta'|\mathbf{Z}) q(\Theta^{(i-1)}|\Theta')}{\pi(\Theta^{(i-1)}|\mathbf{Z}) q(\Theta'|\Theta^{(i)})} \right\}.$$

Algoritmo 2: Algoritmo Metrópolis Hastings

Es decir, una vez calculada la cantidad $\alpha(\Theta^{(i-1)}, \Theta')$, se muestrea un valor de $u \sim \text{Uniforme}(0, 1)$ y

$$\begin{cases} u < \alpha(\Theta^{(i-1)}, \Theta') \implies \Theta^{(i)} = \Theta' \\ u > \alpha(\Theta^{(i-1)}, \Theta') \implies \Theta^{(i)} = \Theta^{(i-1)} \end{cases} \quad (4.2.11)$$

Cabe destacar que si la distribución candidato generadora ($q(\cdot)$) es de tipo simétrica, entonces el problema se reduce a

$$\alpha(\Theta^{(i-1)}, \Theta') = \min \left\{ 1, \frac{\pi(\Theta'|\mathbf{Z})}{\pi(\Theta^{(i-1)}|\mathbf{Z})} \right\} \quad (4.2.12)$$

siendo este último llamado Algoritmo de Metrópolis, un caso particular de Metrópolis Hastings. Con este procedimiento se puede generar una muestra aleatoria de la densidad a posteriori conjunta $\pi(\boldsymbol{\theta}|\mathbf{Z})$.

Para resolver el problema de mezclas de procesos aleatorios Gaussianos (en el caso de variables regionalizadas) se postula el Algoritmo 3, el que define a \mathbf{Y} como un proceso aleatorio Gaussiano de dimensión n y $\mathbf{V}_1 \dots \mathbf{V}_K$ como un conjunto de mezclas que componen a este proceso (con $K \ll n$). Se define $\mathbf{Z}^{(0)} = (Z_1^{(0)}, \dots, Z_P^{(0)})$ como una transformación de \mathbf{Y} según la Figura 4.3, $\mathbf{M}^{(0)} = (M_1^{(0)}, \dots, M_P^{(0)})$ como un vector numérico donde $M_p^{(0)}$ es el resultado de una variable aleatoria multinomial que indica la pertenencia del elemento $Z_p^{(0)}$ a la k -ésima mezcla. Se define además el vector numérico de tipo secuencia $\boldsymbol{\delta} = (\delta_{min}, \dots, \delta_{max})$ de largo igual al número de iteraciones del siguiente algoritmo. Definimos además $\Theta^{(0)} = \{\boldsymbol{\theta}_1^{(0)}, \dots, \boldsymbol{\theta}_K^{(0)}\}$, donde $\boldsymbol{\theta}_k^{(0)} = \{\mu_k^{(0)}, \sigma_k^{2(0)}, \phi_k^{(0)}\}$.

para ii de 1 a $iter$ **hacer**

Evaluar $\pi(\Theta^{(ii-1)}|\mathbf{Z}) = \prod_{k=1}^K f_k \left(\bigcup_{p=1}^P Z_p^{(ii-1)} | M_p^{(ii-1)} = k \right) \pi(\theta_k^{(ii-1)});$

Hacer $P = \delta^{(ii)}$ y crear la partici3n $\mathbf{Z}^{(ii)} = (Z_1^{(ii)}, \dots, Z_P^{(ii)})$ junto a $\mathbf{M}^{(ii)} = (M_1^{(ii)}, \dots, M_P^{(ii)});$

para jj de 1 a P **hacer**

Simular $M^* \sim Multinomial(1, K)$ (grupo de pertenencia aleatorio)
y crear $\mathbf{M}^* = (M_1^{(ii)}, \dots, M_P^{(ii)})$ con $M_{jj}^{*(ii)} = M^*$

Evaluar $\pi(\Theta|\mathbf{Z})^* = \prod_{k=1}^K f_k \left(\bigcup_{p=1}^P Z_p^{(ii)} | M_p^{*(ii)} = k \right) \pi(\theta_k^{(ii-1)}),$

calcular $\alpha(\Theta^{(ii-1)}, \Theta^*) = \min \left\{ 1, \frac{\pi(\Theta|\mathbf{Z})^*}{\pi(\Theta^{(ii-1)}|\mathbf{Z})} \right\}$ y simular un valor $u \sim U(0, 1)$

si $\alpha(\Theta^{(ii-1)}, \Theta^*) > u$ **entonces**

| $\mathbf{M}^{(ii)} = \mathbf{M}^*$

fin

fin

para jj de 1 a K **hacer**

Generar el candidato $\mu^* \sim Normal(\mu_{jj}^{(ii-1)}, \psi)$ y crear

$\Theta^* = \{\theta_1^{(ii-1)}, \dots, \theta_K^{(ii-1)}\}$ con $\theta_{jj}^{*(ii-1)} = \{\mu^*, \sigma_{jj}^{2(ii-1)}, \phi_{jj}^{(ii-1)}\}$

Evaluar $\pi(\Theta^*|\mathbf{Z}) = \prod_{k=1}^K f_k \left(\bigcup_{p=1}^P Z_p^{(ii)} | M_p^{(ii)} = k \right) \pi(\theta_k^*),$

calcular $\alpha(\Theta^{(ii-1)}, \Theta^*) = \min \left\{ 1, \frac{\pi(\Theta^*|\mathbf{Z})}{\pi(\Theta^{(ii-1)}|\mathbf{Z})} \right\}$ y simular un valor $u \sim U(0, 1)$

si $\alpha(\Theta^{(ii-1)}, \Theta^*) > u$ **entonces**

| $\theta_{jj}^{(ii)} = \theta_{jj}^{*(ii-1)}$

en otro caso

| $\theta_{jj}^{(ii)} = \theta_{jj}^{(ii-1)}$

fin

Calcular el variograma experimental con $\bigcup_{p=1}^P Z_p^{(ii)} | M_p^{(ii)} = jj$

Ajustar el variograma para actualizar $\Theta^{(ii-1)}$ a $\Theta^{(ii)}$
($\sigma_{jj}^{2(ii-1)} \rightarrow \sigma_{jj}^{2(ii)}$ y $\phi_{jj}^{(ii-1)} \rightarrow \phi_{jj}^{(ii)}$)

fin

fin

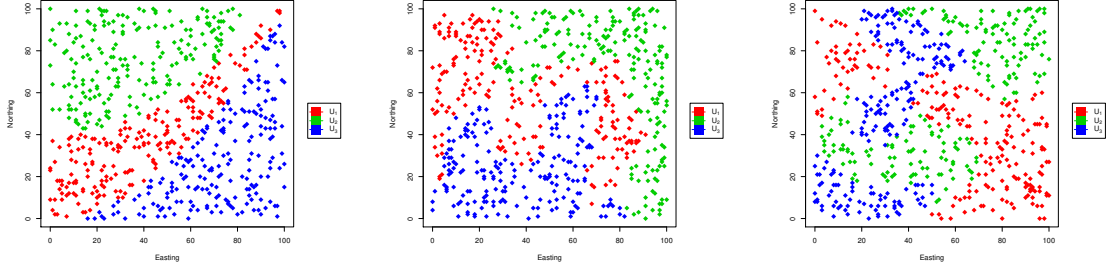
Algoritmo 3: Algoritmo Metr3polis Hastings para mezclas de procesos aleatorios Gausianos

Como resultado se espera obtener una cadena de Markov del tipo $\{\Theta^{(1)}, \Theta^{(2)}, \dots, \Theta^{(t)}, \Theta^{(t+1)}, \dots\}$ la que en las primeras iteraciones estará altamente autocorrelacionada. Como el objetivo es extraer de esta cadena una muestra aleatoria de la distribución estacionaria de $\pi(\Theta|\mathbf{Z})$, se realiza un proceso descarte de las primeras iteraciones con el fin de eliminar esta autocorrelación, proceso denominado como “burnin” (Raftery y Lewis (1996)). Adicionalmente, buscando lograr el mismo objetivo, se realiza una selección sistemática de muestras provenientes de la cadena, proceso denominado como “thinin”. Como desventaja, estos procedimientos de descarte inicial y sistemático de valores provenientes de la cadena, aumentan el número necesario de simulaciones para realizar los análisis de la distribución a posteriori. Por último, se podrá extraer de la distribución a posteriori los parámetros $\theta_k = \{\mu_k, \sigma_k^2, \phi_k\}$ junto a una cadena que almacena los valores $\mathbf{M} = (M_1, \dots, M_P)$, que indican la pertenencia de cada conglomerado a cada mezcla. Con esta última información, se podrá obtener una clasificación de tipo probabilística para cada individuo del conjunto de datos. El Algoritmo Metrópolis Hastings para mezclas de procesos aleatorios Gaussianos se asemeja con los propuestos en Allard y Guillot (2000) y Romary y cols. (2012), siendo la mayor diferencia la elección del método de maximización de la función de verosimilitud y la creación de la partición auxiliar respectivamente. El método utilizado en Allard y Guillot (2000) es el algoritmo “EC-M” para conglomerados espaciales, un derivado del algoritmo “CEM” para clasificación propuesto en Celeux y Govaert (1992). En el caso de Romary y cols. (2012), la grilla o partición auxiliar es creada a partir de la triangulación de Delaunay, herramienta utilizada para estructurar los datos con respecto a una medida de proximidad. Para la mezcla de procesos aleatorios Gaussianos, se utilizó el enfoque bayesiano y el algoritmo MCMC Metrópolis Hastings para lograr maximizar la verosimilitud; y el algoritmo K-Medias para crear la partición auxiliar.

4.3. Caso de estudio sintético

4.3.1. Metodología de simulación

Las geometrías posibles de las unidades son muy variadas, van desde unidades regulares en el espacio con fronteras marcadas hasta unidades disgregadas y complejas. La capacidad de identificación de fronteras entre unidades toma una importancia relevante a la hora de realizar un análisis de conglomerados, por lo que esta cualidad se pone a prueba en el estudio de simulación siguiente. Se plantean tres casos de estudio, los que se muestran en la Figura 4.6. El primero corresponde a un caso donde la frontera entre las unidades está plenamente definida y presenta bordes suaves. El segundo caso de estudio corresponde a unidades intrusivas y poco regulares, conservando la propiedad de conexidad. El tercer caso corresponde a unidades disconexas y poco regulares, con fronteras intrusivas y geometría compleja. La dificultad para el algoritmo aumenta con cada caso de datos simulados. Aquí se pone a prueba la eficacia del algoritmo bayesiano y la metodología empleada. En cada caso de estudio, las unidades son generadas a partir de un proceso estacionario de segundo orden de 200 observaciones cada uno, con una estructura de covarianza dada por $(\sigma^2, \phi)^\top = (0.1, 50.0)$ y un covariograma de tipo exponencial, quedando un total de 600 observaciones por caso.



(a) Caso 1: Fronteras duras (b) Caso 2: Fronteras intrusivas (c) Caso 3: Fronteras desconexas

Figura 4.6: Casos de estudio con diferencias en la definición de sus fronteras

$$\begin{cases} \mathbf{Y}|\mathbf{U}_1 \sim N_n(\boldsymbol{\mu}_1, \mathbf{C}) \\ \mathbf{Y}|\mathbf{U}_2 \sim N_n(\boldsymbol{\mu}_2, \mathbf{C}) \\ \mathbf{Y}|\mathbf{U}_3 \sim N_n(\boldsymbol{\mu}_3, \mathbf{C}) \end{cases} \quad \begin{aligned} n &= 200 \\ \mathbf{C} &= \mathbf{C}(h) = \sigma^2 \rho(h|\phi) \end{aligned} \quad (4.3.1)$$

La diferencia entre unidades radica en sus medias, siendo $(\mu_1, \mu_2, \mu_3)^\top = (1.0, 3.0, 5.0)$ los valores escogidos. Las prioris utilizadas fueron distribuciones Gaussianas, las que son simétricas y conducen a la aplicación del Algoritmo de Metrópolis, un caso particular de Metrópolis Hastings (Celeux y cols. (2000)).

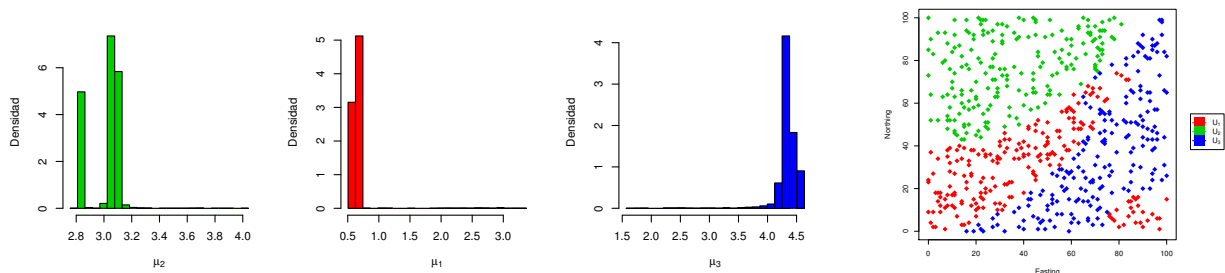
4.3.2. Resumen de resultados

En la Tabla (4.1) se muestra el resumen de las medias a posteriori para los parámetros de medias poblacionales obtenidos para las 100 simulaciones. Destaca el valor alto de la desviación estandar para la unidad U3, en el caso de estudio con fronteras duras y desconexas. Este valor se condice con el sesgo obtenido, el que mide el error en la estimación del parámetro. Como las distribuciones para todos los parámetros son asimétricas, se recomienda utilizar como estimador la mediana en vez de la media.

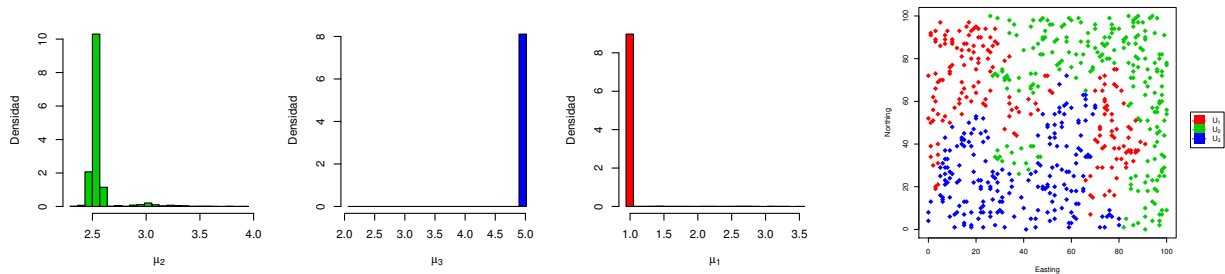
Fronteras		Valor real	$E\{\mu \mathbf{Y}\}$	$\sigma_{\mu \mathbf{Y}}$	Cuantiles					Sesgo
		μ			0%	25%	50%	75%	100%	
Duras	\mathbf{U}_1	1.0000	1.0103	0.4530	0.0000	0.8064	1.0585	1.2333	1.9968	-0.0103
	\mathbf{U}_2	3.0000	2.9487	0.4602	2.0450	2.7069	2.9886	3.1285	3.9993	0.0171
	\mathbf{U}_3	5.0000	3.9699	1.8406	0.0000	4.1606	4.7490	5.0048	5.6593	0.2060
Intrusivas	\mathbf{U}_1	1.0000	1.0609	0.3458	0.0000	0.8914	1.0163	1.2203	1.9820	-0.0609
	\mathbf{U}_2	3.0000	2.9538	0.7281	0.0000	2.7228	3.0917	3.4175	3.7892	0.0154
	\mathbf{U}_3	5.0000	4.8411	0.3059	4.1043	4.6480	4.8563	5.0632	5.6309	0.0318
Desconexas	\mathbf{U}_1	1.0000	1.0711	0.4031	0.0000	0.8209	1.0369	1.2577	1.9378	-0.0711
	\mathbf{U}_2	3.0000	3.1739	0.6776	0.0000	2.9804	3.1551	3.5722	3.9931	-0.0580
	\mathbf{U}_3	5.0000	4.3757	1.3372	0.0000	4.4091	4.7451	4.9962	5.4010	0.1249

Tabla 4.1: Resumen de medias a posteriori para los casos de estudio en las 100 simulaciones

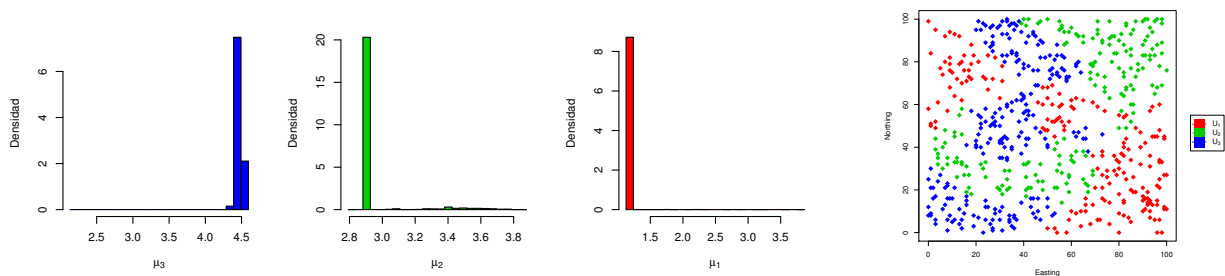
En la Figura (4.7) es posible apreciar los resultados obtenidos para una simulación en cada caso de estudio. Por un lado, se observan las distribuciones a posteriori de los parámetros μ que corresponden a las medias poblacionales de cada unidad, donde se destaca la forma leptocúrtica de las distribuciones. Además se observa la distribución de las unidades en base a la frecuencia máxima de asignación para cada localidad en cada una de las simulaciones. En cada caso de estudio se observan pequeñas variaciones con respecto a la configuración real. Sin embargo, la estrategia de asignación en base a la frecuencia máxima reproduce las fronteras entre unidades. En la Figura (4.8) es posible observar las distribuciones de las medias a posteriori para las 100 simulaciones, las que presentan una clara tendencia hacia los valores reales de los parámetros y se destacan por su baja dispersión.



(a) Simulación #24 en el caso de estudio con fronteras duras

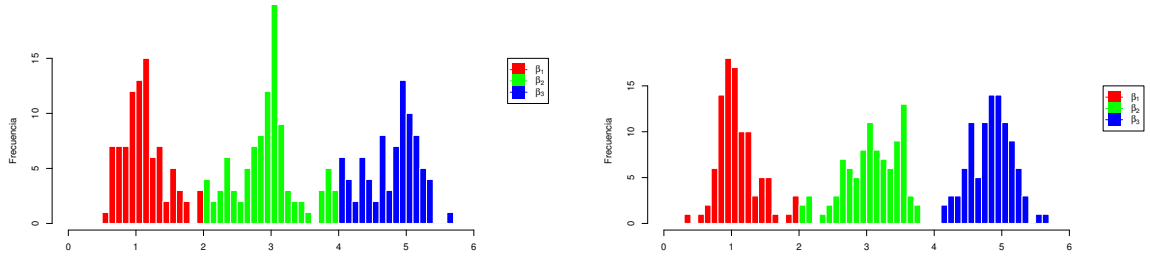


(b) Simulación #12 en el caso de estudio con fronteras intrusivas



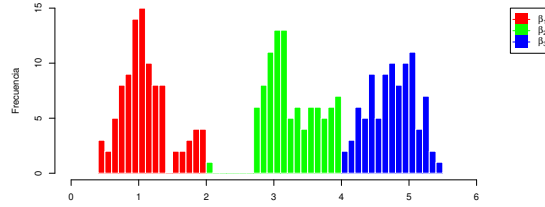
(c) Simulación #7 en el caso de estudio con fronteras disconexas

Figura 4.7: Distribución a posteriori de las medias poblacionales y gráfico de pertenencia máxima para los casos de estudio en una de las simulaciones



(a) Distribución a posteriori para el caso de fronteras duras

(b) Distribución a posteriori para el caso de fronteras intrusivas



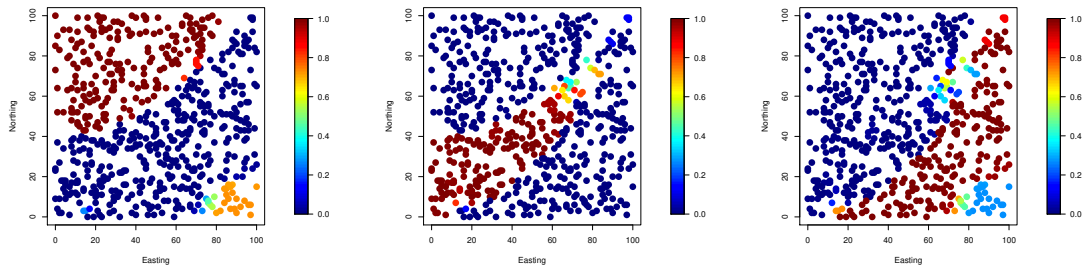
(c) Distribución a posteriori para el caso de fronteras disconexas

Figura 4.8: Distribución de medias a posteriori para los casos de estudio en las 100 simulaciones

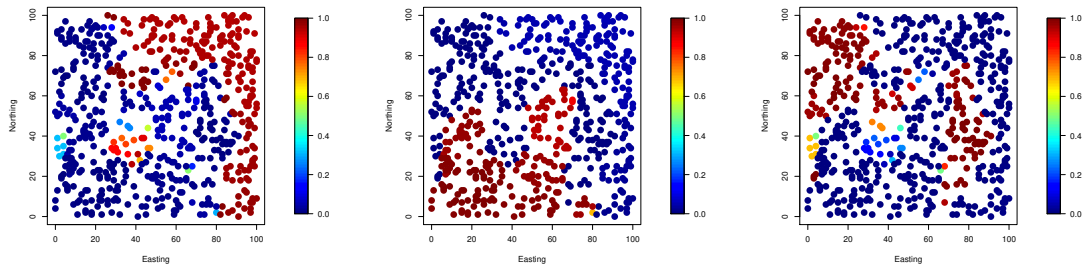
Fronteras	Matriz de confusión			
		\mathbf{U}_1	\mathbf{U}_2	\mathbf{U}_3
Duras	\mathbf{U}_1	0.83025(0.3042)	0.01195(0.0999)	0.01515(0.0245)
	\mathbf{U}_2	0.14520(0.3099)	0.72410(0.4409)	0.07060(0.2046)
	\mathbf{U}_3	0.02455(0.0389)	0.26395(0.4353)	0.91425(0.2033)
Intrusivas	\mathbf{U}_1	0.92590(0.1320)	0.05405(0.1270)	0.01260(0.0330)
	\mathbf{U}_2	0.06125(0.1338)	0.87625(0.1900)	0.09865(0.1188)
	\mathbf{U}_3	0.01285(0.0162)	0.06970(0.1576)	0.88875(0.1262)
Disconexas	\mathbf{U}_1	0.95810(0.0549)	0.12660(0.3006)	0.03660(0.1362)
	\mathbf{U}_2	0.02185(0.0256)	0.71420(0.3681)	0.11775(0.1738)
	\mathbf{U}_3	0.02005(0.0540)	0.15920(0.2773)	0.84565(0.2059)

Tabla 4.2: Resumen de pertenencia y porcentaje de asignación a posteriori para las 100 simulaciones

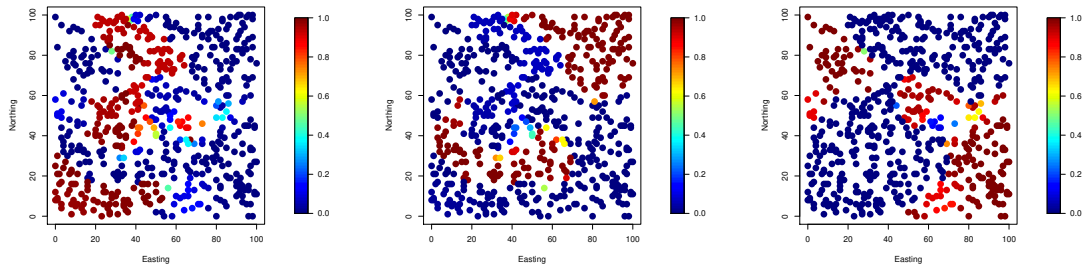
En la Tabla (4.2) se observa el resumen de pertenencia y porcentaje de asignación a posteriori, con su respectiva medida de desviación estándar entre paréntesis. Se destaca el mayor porcentaje de asignación correcta para la unidad 1 en el caso de fronteras disconexas y el menor porcentaje de asignación correcta para la unidad 2 en el mismo caso de estudio. En el resto de las posibles asignaciones erróneas se observa que en promedio, éstas no superan el 27% de los casos (unidad 3, fronteras duras), destacando el caso de fronteras intrusivas como el que tuvo un mejor desempeño.



(a) Simulación #24 en el caso de estudio de fronteras duras



(b) Simulación #12 en el caso de estudio de fronteras intrusivas

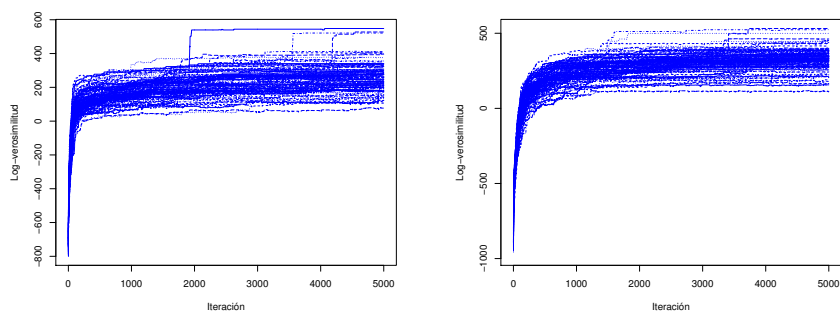


(c) Simulación #7 en el caso de estudio de fronteras disconexas

Figura 4.9: Probabilidades de pertenencia a posteriori para los casos de estudio en una de las simulaciones

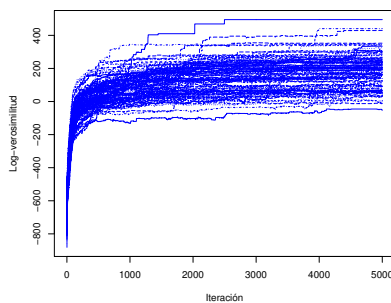
La Figura 4.9 muestra lo que son las probabilidades de pertenencia a posteriori para los casos de estudio en una de las simulaciones, factor clave y motivante en la propuesta, el que es obtenido gracias al enfoque bayesiano del algoritmo empleado. Con este tipo de resultados, es posible obtener un mapa de probabilidad de pertenencia para cada simulación, para luego poder hacer un resumen de simulaciones. Los dominios formados en cada una de las figuras hacen referencia a las unidades y la probabilidad de encontrarlas en cada individuo en el espacio. El algoritmo empleado (Algoritmo 3) se basa en una maximización iterativa de la función de verosimilitud ponderada por las distribuciones a priori y las distribuciones candidato generadoras. En base a este criterio, el mejor modelo es el que presenta una log-

verosimilitud mayor en comparación al resto. Como los casos de estudio no son comparables, no es posible aplicar los criterios de AIC, BIC o EBIC, conocidos en el ámbito del análisis bayesiano. No obstante, podemos identificar aquellas simulaciones que presenten un mayor valor de log-verosimilitud alcanzado y suponer que la configuración respectiva es la óptima. En base a la figura (4.10), podemos identificar un valor promedio alcanzado de 242.8, 315.8 y 167.3 para las simulaciones en los casos de estudio de fronteras duras, intrusivas y disconexas respectivamente, valores presentes en la tabla (4.3). En los casos de estudio se observa una dispersión de la log-verosimilitud similar, la que indicaría un estado de convergencia a la distribución estacionaria de los parámetros y la configuración, lo que es un requisito para construir una muestra aleatoria que permita hacer la inferencia. En la Figura (4.11) se observa la distribución discreta a posteriori para tres de los individuos escogidos aleatoriamente en los diferentes casos de estudio. En la Figura (4.11(a)) se distingue una clara definición de la asignación para los individuos 346 y 518 a las unidades 2 y 3 respectivamente. Para el caso del individuo 53, en vista que se encuentra localizado cercano a la frontera de las unidades 1 y 3, tiene como resultado una distribución de frecuencias repartidas entre estas unidades, no obstante, se inclina por una mayor asignación o pertenencia a la unidad 3. La misma situación podemos observar en la distribución discreta a posteriori de la Figura (4.11(b)) en el caso de estudio de fronteras intrusivas. El individuo 344 presenta una distribución repartida entre las unidades 1 y 3 debido a su cercanía a la frontera entre éstas.



(a) Fronteras duras

(b) Fronteras intrusivas

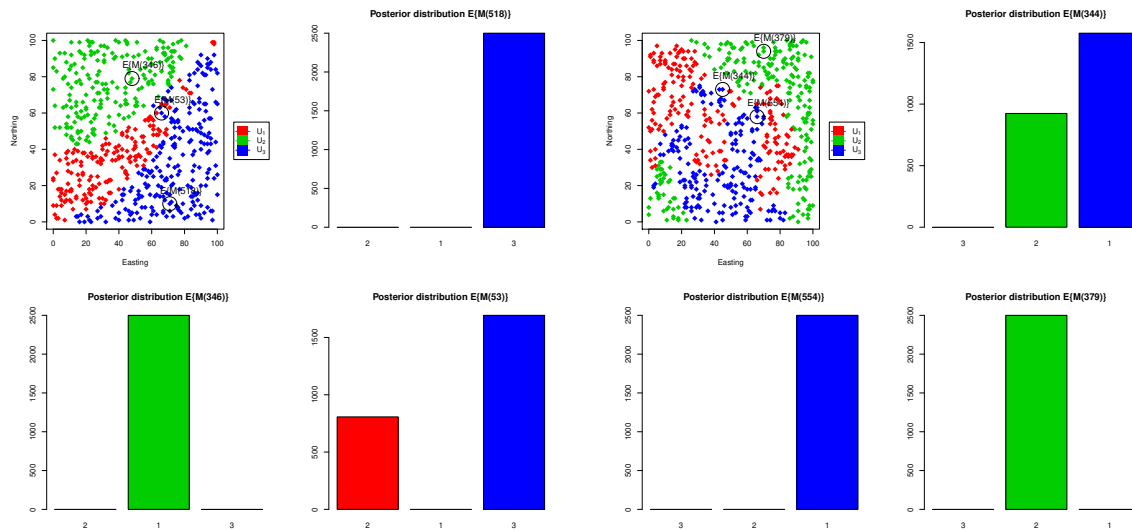


(c) Fronteras disconexas

Figura 4.10: Resumen de Log-verosimilitudes para las 100 simulaciones en cada caso de estudio

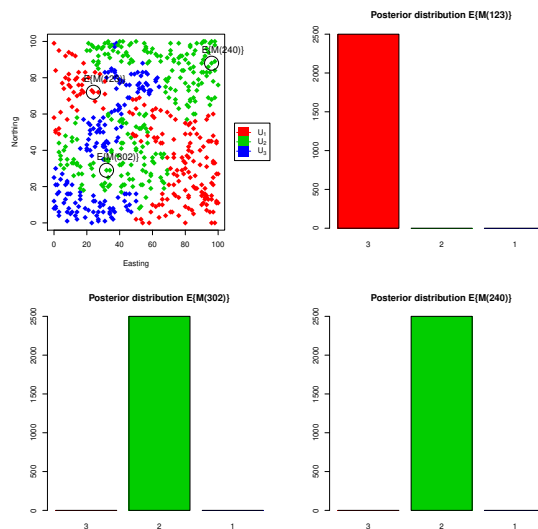
Fronteras	Media	Des. Est.	Cuantiles				
			0 %	25 %	50 %	75 %	100 %
Duras	242.8	12.01	217.4	233.1	242.4	254.1	260.2
Intrusivas	315.8	12.50	288.3	306.7	318.3	328.0	334.2
Disconexas	167.3	15.09	137.6	155.0	169.4	180.6	188.7

Tabla 4.3: Resumen de log-verosimilitud para las 100 simulaciones



(a) Simulación #19 en el caso de estudio de fronteras duras

(b) Simulación #34 en el caso de estudio de fronteras intrusivas



(c) Simulación #25 en el caso de estudio de fronteras disconexas

Figura 4.11: Distribución de pertenencia a posteriori en los casos de estudio para una simulación

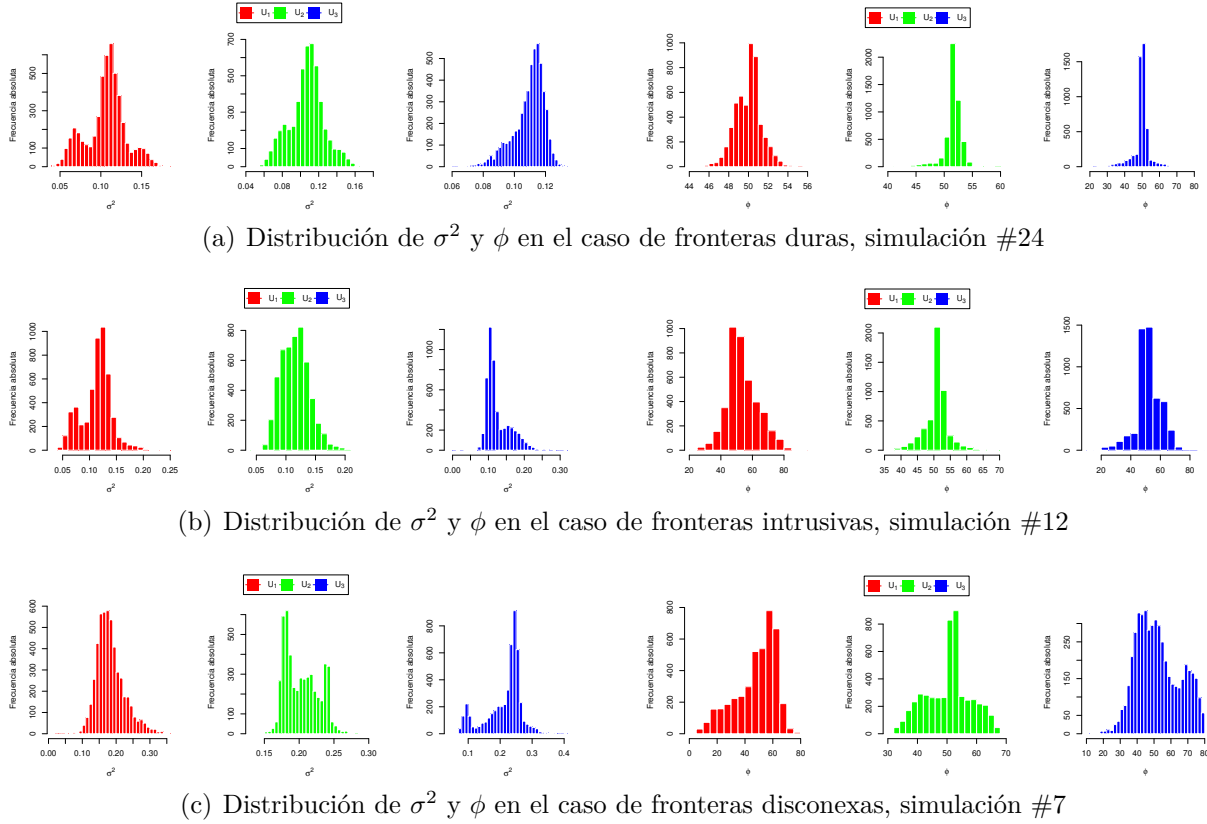


Figura 4.12: Distribución de los parámetros espaciales para los casos de estudio en una de las simulaciones (valores reales: $\sigma^2 = 0,1$ y $\phi = 50$)

En estos casos se asume que aún el algoritmo no ha alcanzado la distribución estacionaria, por lo que es necesario un número mayor de iteraciones para alcanzar la convergencia. Para la Figura (4.11(c)) correspondiente al caso de estudio con fronteras disconexas se observa un grado de pertenencia absoluto en los individuos seleccionados debido a que estos se encuentran cercanos al centro de cada unidad. Por último, la Figura (4.12) muestra los resultados obtenidos y la distribución de los parámetros espaciales ajustados en cada iteración del algoritmo bayesiano. Destacamos una sobrestimación del parámetro σ^2 en el caso de fronteras disconexas para una de las simulaciones. Este comportamiento se presenta en uno de los casos más complejos en el estudio de simulación y es posible que con el número de iteraciones utilizadas no se haya alcanzado la convergencia de la cadena. Éstos parámetros no fueron considerados como parámetros para ajustar la distribución a posteriori en base al criterio bayesiano, sino que fueron obtenidos utilizando el ajuste por mínimos cuadrados en cada iteración. Las formas de las distribuciones de estos parámetros están condicionadas a la configuración de las unidades a medida que se maximiza la verosimilitud, por lo que el ideal es que una vez alcanzada la distribución estacionaria, estos parámetros debiesen tener una forma acampanada y con una baja dispersión en su distribución de frecuencias.

4.4. Caso de estudio real: Minera Escondida

4.4.1. Descripción de la base de datos

Minera Escondida es un yacimiento tipo pórfido Cu-Mo ubicado en la II Región de Antofagasta. Sus instalaciones principales, minas y plantas se encuentran a 170 Km al SE de la ciudad de Antofagasta a una altitud de 3100 metros sobre el nivel del mar. Pertenecen en un 57.5% a BHP Billiton Ltds., 30% a Rio Tinto Plc, 10% al consorcio corporativo JECO el cual incluye Mitsubishi, Nippon Mining and Metals Ltd., y 2.5% a International Finance Corp. Cuenta con dos faenas principales a rajo abierto: Escondida y Escondida Norte; siendo la primera de estas la zona de interés (Figura 4.13). Estas faenas se encuentran dentro del dominio del Sistema de Fallas de Domeyko en la franja metalogénica del Eoceno superior-Oligoceno inferior. El yacimiento de Escondida se caracteriza por una secuencia de rocas volcánicas y sedimentarias intruidas por varios cuerpos magmáticos responsables de la mineralización económica (Fig. 4.14). La alteración hidrotermal de Escondida se asocia a la intrusión del pórfido cuarzo monzonítico a granodiorítico del Eoceno - Oligoceno emplazado en lavas andesíticas Paleocenas. La distribución espacial de las alteraciones corresponde a una zonación típica de un sistema tipo pórfido cuprífero (Fig. 4.15).

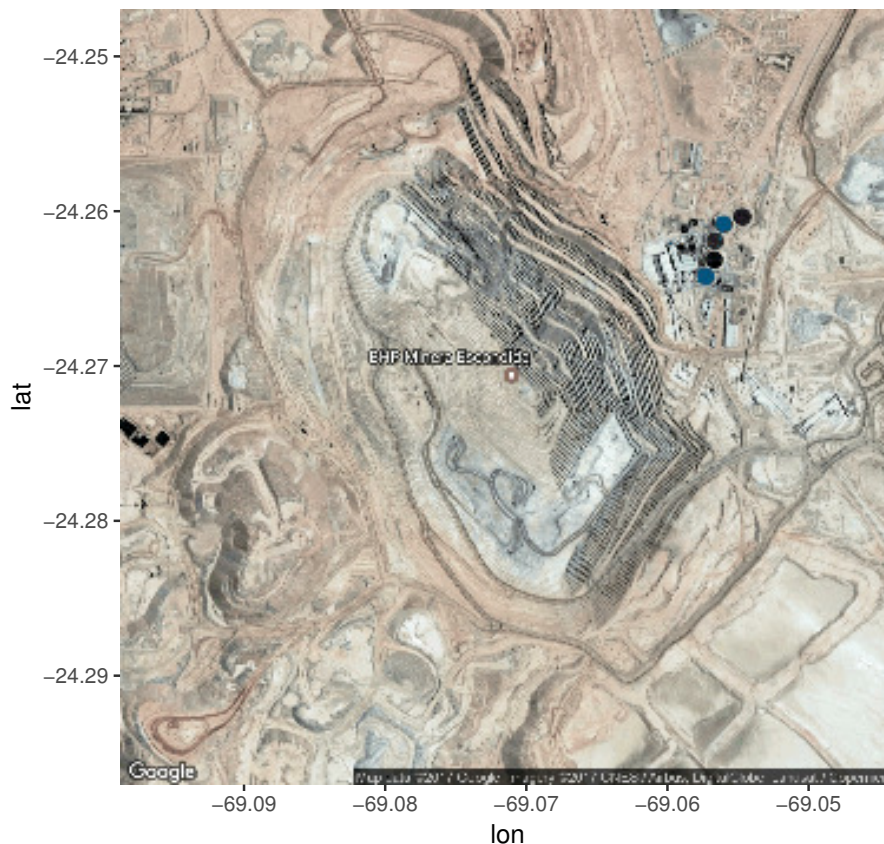


Figura 4.13: Zona de estudio, Mina Escondida, II Región de Antofagasta

	Este	Norte	Elevación	Ley de Cu Total	Ley de Fe Total
Mínimo	14731	106000	1335	0.0100	0.2600
1er Cuartil	15926	107050	2612	0.5000	1.3250
2do Cuartil	16386	107553	2747	0.7250	1.7900
Media	16332	107714	2722	0.8175	1.9950
3er Cuartil	16780	108260	2846	1.0100	2.5350
Máximo	17945	109777	3172	2.5000	5.0000
Des.est.	643.91	828.76	181.1	0.4379	0.9266

Tabla 4.4: Estadísticas descriptivas para variables en la base de datos

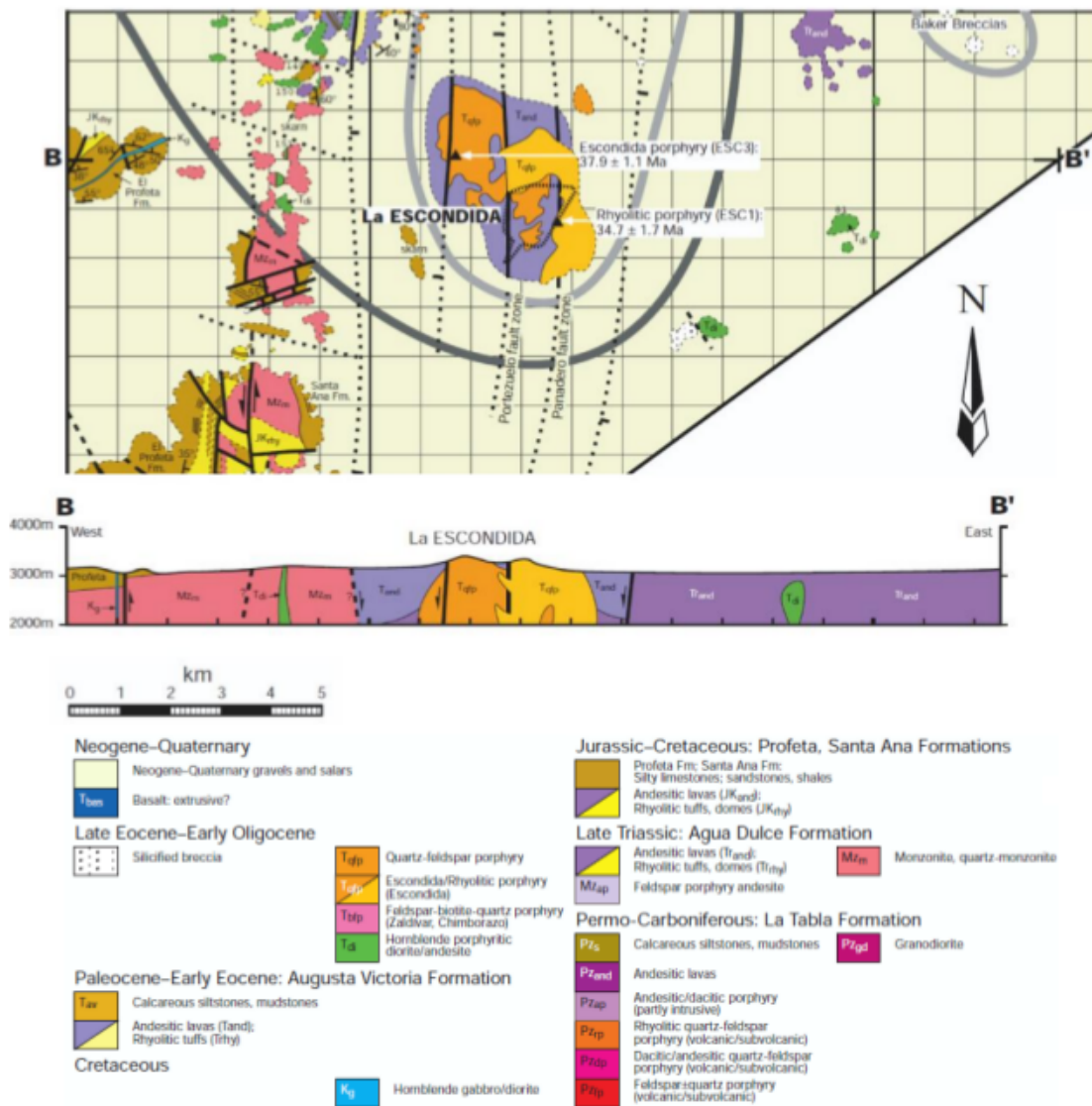


Figura 4.14: Mapa geológico del yacimiento Escondida (Richards y cols. (2001))

En la base de datos contamos con información acerca de la ley de cobre (Cu) total y la ley de hierro (Fe) total distribuidas en los sondeos de exploración (Tabla 4.4). Estas leyes presentan un factor de correlación lineal de -0.1295 , el que resulta ser significativo ($p\text{-valor} < 2.2e-16$), pero muy débil. Este hecho impide que al utilizar herramientas estadísticas tradicionales, estas variables sean analizadas de manera conjunta (Fig. 4.16). Para ambas variables, se observa un comportamiento distribucional frecuencial unimodal (Figura 4.17(a) y 4.17(b)), donde resaltan las concentraciones altas de leyes de Cu Total al centro del dominio (Figura 4.17(c)) y concentraciones altas de Fe Total por el borde externo de la zona de estudio (Figura 4.17(d)). Gracias a las vistas de distribuciones de sondeos (Figura 4.17(e) y 4.17(f)) es posible apreciar una disparidad en el rango de las coordenadas, lo que se refleja en sondeos que cubren una mayor extensión en planta y que poseen una menor longitud en la vertical.

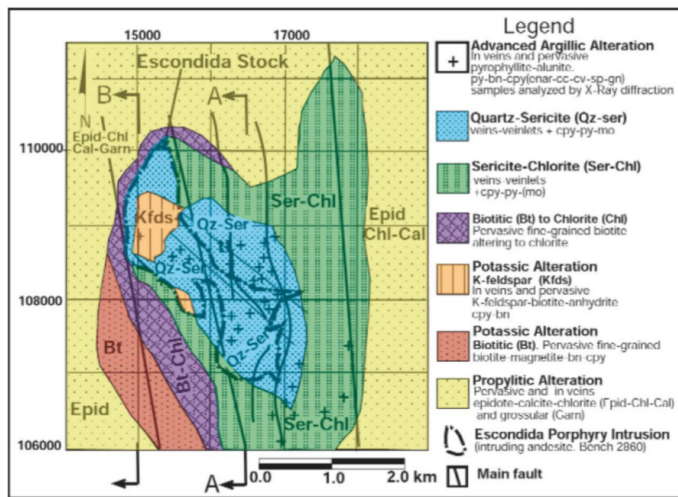


Figura 4.15: Mapa de alteraciones hidrotermales del yacimiento Escondida (Padilla y cols. (2001))

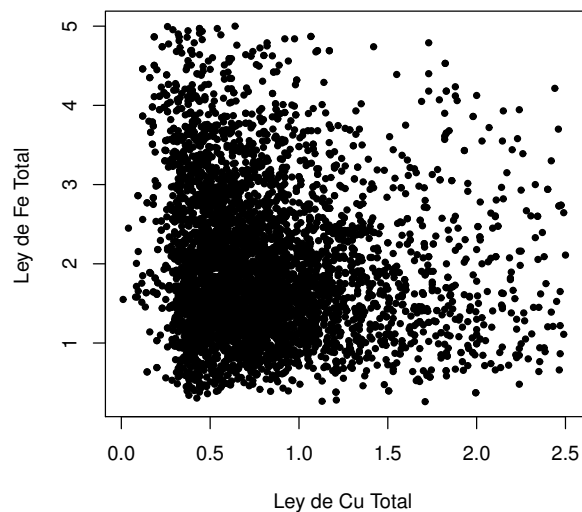
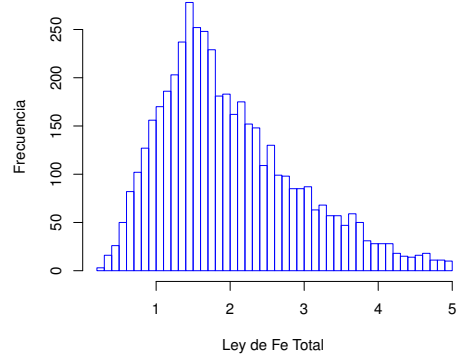
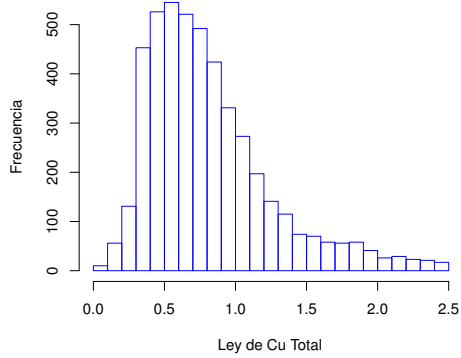
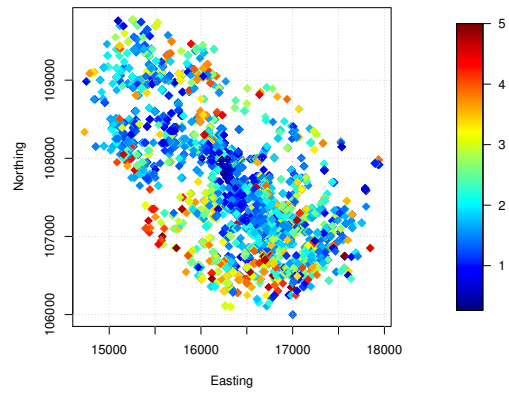
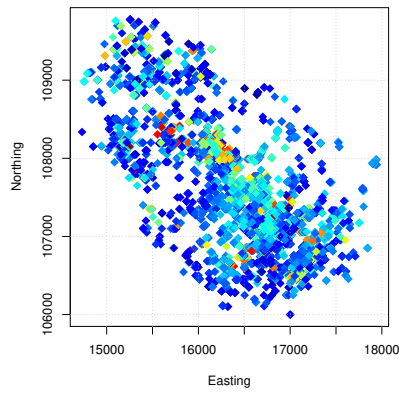


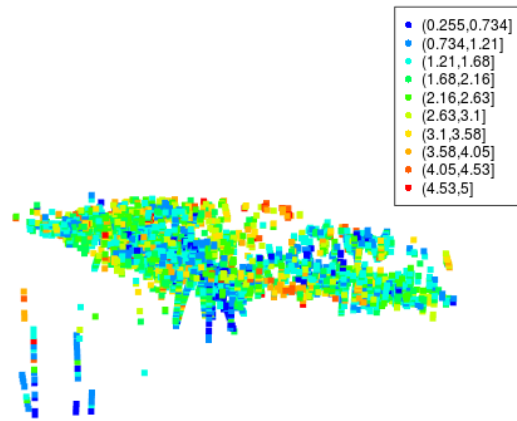
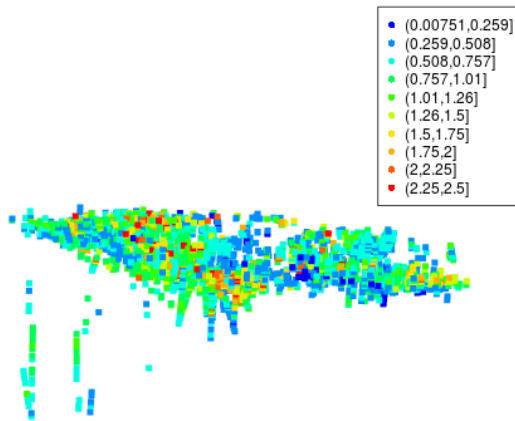
Figura 4.16: Distribución de las leyes de Cu Total y Fe Total



(a) Distribución frecuencial de la Ley de Cu Total (b) Distribución frecuencial de la Ley de Fe Total



(c) Distribución en planta de la Ley de Cu Total (d) Distribución en planta de la Ley de Fe Total



(e) Distribución espacial de la Ley de Cu Total (f) Distribución espacial de la Ley de Fe Total

Figura 4.17: Distribución frecuencial, en planta y espacial para las variables Ley de Cu Total y Ley de Fe Total

4.4.2. Aplicación

El objetivo fue aplicar la propuesta de Mezcla de Distribuciones Geoestadísticas para identificar de manera independiente unidades de ley de Cu Total y ley de Fe Total, cuya correlación a pesar de ser significativa (p -valor $< 2,2e - 16$), resultó ser baja, con un valor de -0.1295 según el test de correlación de Pearson. Estas unidades, debiesen tener características similares y además deben ser gobernadas por una estructura de correlación propia, por lo que los variogramas experimentales son calculados en forma omnidireccional y ajustados por un modelo exponencial en las distintas iteraciones (i.e., no se calcula ni ajusta el variograma cruzado). Las Figuras 4.18(a) y 4.18(c) muestran los resultados obtenidos al aplicar la propuesta en el caso de la Ley de Cu Total, para tres unidades previamente escogidas. De igual manera, las Figuras 4.18(b) y 4.18(d) muestran los resultados para la variable Ley de Fe Total. La Tabla 4.5 muestra las distribuciones de los valores medios de cada unidad descubierta para las variables Ley de Cu Total y Ley de Fe Total.

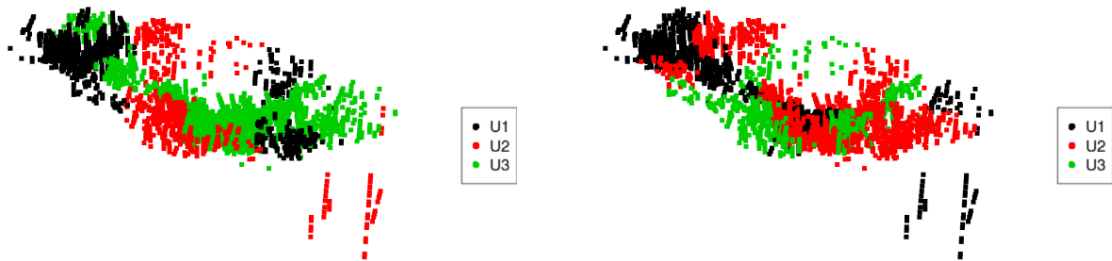
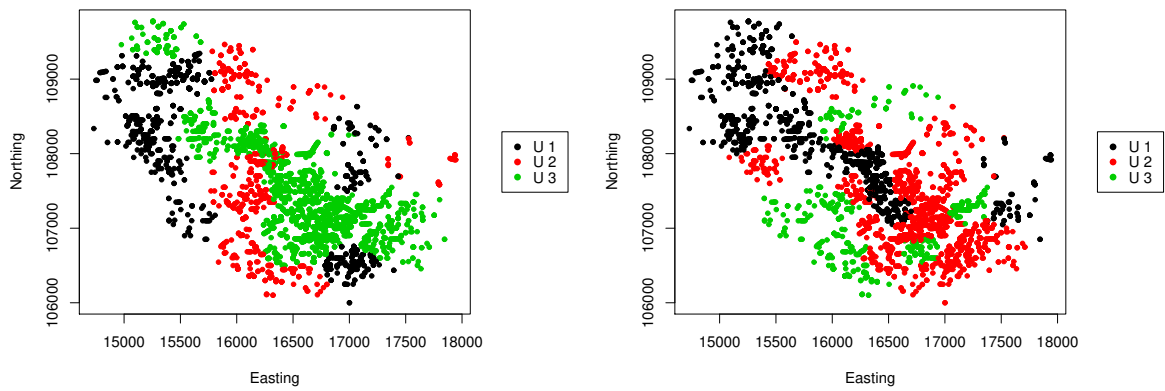
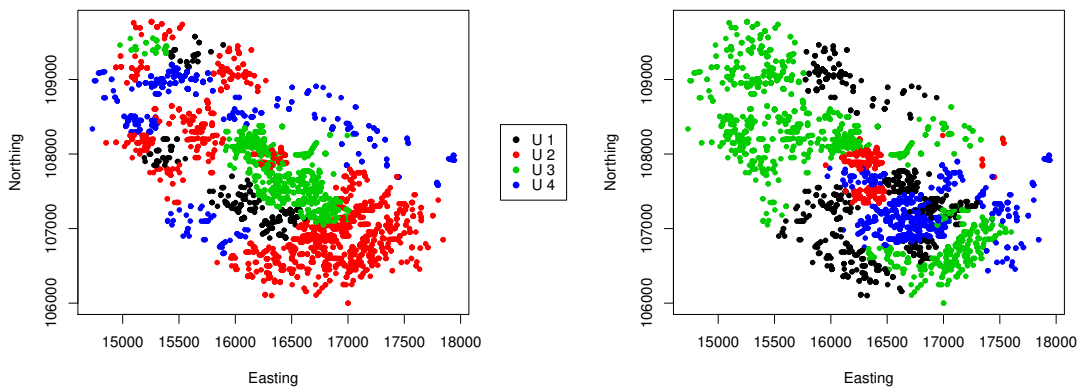


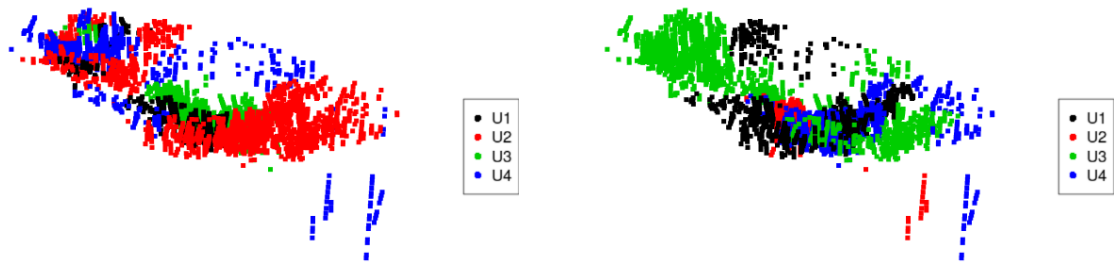
Figura 4.18: Distribución en planta y espacial por mezcla para la Ley de Cu Total y la Ley de Fe Total para tres unidades descubiertas

	μ Ley de Cu Total			μ Ley de Fe Total		
	U_1	U_2	U_3	U_1	U_2	U_3
Mínimo	0.64	0.60	0.90	1.62	1.98	2.20
1er Cuartil	0.70	0.60	0.95	1.62	2.00	2.79
Mediana	0.70	0.60	0.95	1.71	2.06	2.79
Media	0.70	0.61	0.96	1.72	2.05	2.75
3er Cuartil	0.71	0.62	0.96	1.80	2.09	2.79
Máximo	0.74	0.74	0.99	2.00	2.10	2.79
Des. est.	0.02	0.04	0.01	0.10	0.04	0.12

Tabla 4.5: Descriptivos de los parámetros μ para la Ley de Cu Total y Ley de Fe Total por unidad descubierta



(a) Distribución en planta de las 4 mezclas para la Ley de Cu Total (b) Distribución en planta de las 4 mezclas para la Ley de Fe Total



(c) Distribución espacial de las 4 mezclas para la Ley de Cu Total (d) Distribución espacial de las 4 mezclas para la Ley de Fe Total

Figura 4.19: Distribución en planta y espacial por mezcla para la Ley de Cu Total y la Ley de Fe Total para cuatro unidades descubiertas

La Tabla 4.5 nos muestra la forma en la que los valores medios de las leyes de Cu Total y Fe Total se distribuyen en las unidades o mezclas encontradas por la propuesta. Para todas las unidades destacamos el bajo valor de la desviación estándar encontrada, lo que muestra la convergencia del algoritmo, lo que era un aspecto importante dentro de la propuesta. En

cuanto a los valores medios, que corresponden a las medias poblacionales estimadas de cada mezcla, se obtuvieron cantidades que muestran zonas con concentraciones medias diferentes. Como caso complementario, se realizó el mismo ejercicio en el caso de querer descubrir cuatro unidades o mezclas. Estos resultados se pueden apreciar en la Figura 4.19 y en la Tabla 4.6, para las variables Ley de Cu Total y Ley de Fe Total respectivamente. Al tener diferentes escenarios, podremos utilizar el conocimiento geológico que se tiene de la zona de estudio para validar los resultados y escoger el número óptimo de unidades a descubrir, por ejemplo, identificando las unidades descubiertas con unidades mineralógicas, litológicas o de alteración.

	μ Ley de Cu Total				μ Ley de Fe Total			
	U_1	U_2	U_3	U_4	U_1	U_2	U_3	U_4
Mínimo	0.69	0.74	0.95	0.60	2.34	1.19	1.99	1.76
1er Cuartil	0.72	0.79	1.06	0.61	2.54	1.19	2.04	1.90
Mediana	0.72	0.80	1.06	0.61	2.57	1.19	2.04	1.90
Media	0.72	0.80	1.05	0.61	2.55	1.30	2.05	1.89
3er Cuartil	0.75	0.80	1.06	0.61	2.57	1.35	2.05	1.90
Máximo	0.77	0.81	1.07	0.70	2.58	1.78	2.12	1.93
Des. est.	0.02	0.01	0.02	0.02	0.05	0.18	0.02	0.04

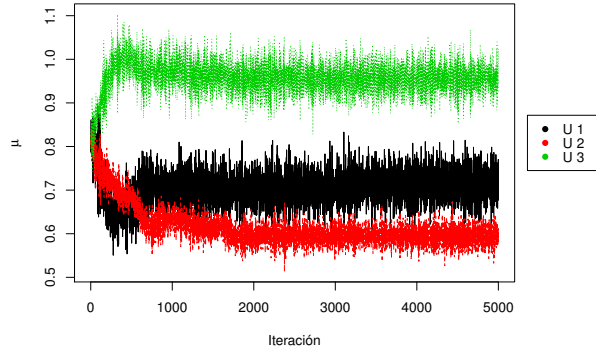
Tabla 4.6: Descriptivos de los parámetros μ para la Ley de Cu Total y Ley de Fe Total por unidad descubierta

4.4.3. Resumen de resultados

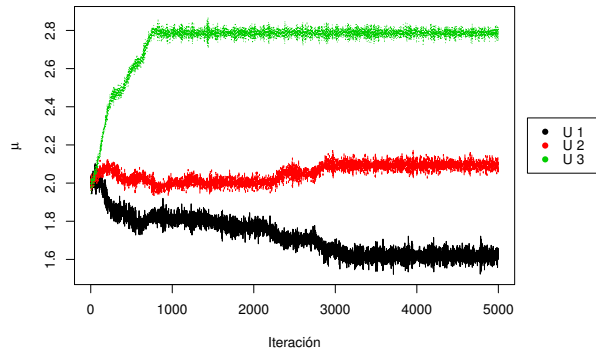
La Figura 4.20 muestra la forma en la que la cadena de valores propuestos para ajustar los parámetros de medias poblacionales va convergiendo. En el caso de la Figura 4.20(a), podemos destacar la desviación estándar de cada cadena, siendo ésta una cantidad amplia, pero controlada. Este carácter demuestra la convergencia del algoritmo y la ausencia de autocorrelación, condiciones necesarias para obtener una muestra de la distribución a posteriori de cada parámetro. Para la variable Ley de Fe Total, a diferencia de la anterior, la Figura 4.20(b) muestra un bajo valor de desviación estándar y un comportamiento convergente.

La Tabla 4.7 muestra las verosimilitudes obtenidas para cada mezcla por variable de estudio. Un valor mayor en la verosimilitud indica un acercamiento al verdadero valor medio, por lo que en ese sentido, es posible ordenar en función de credibilidad las mezclas encontradas, destacando en primer lugar la mezcla 2, luego la 1 y por último la 3 en el caso de la Ley de Cu Total. La mezcla 3 es la que contiene las leyes más altas de Cu Total en la zona de estudio (Figuras 4.17(c) y 4.18(a)), por lo que se debiese tener un mayor grado de credibilidad en la definición de esta zona. El valor de la verosimilitud se ve influenciado por valores extremos que están presentes en esa zona (Figura 4.17(c)).

El orden de credibilidad para la Ley de Fe Total resulta ser mezcla 3, mezcla 1 y por último mezcla 2. La mezcla 3 es la que contiene los valores más altos para la Ley de Fe Total, la mezcla 1 es la que contiene los valores más bajos para la Ley de Fe Total y la mezcla 2 es la que contiene los valores intermedios (Figuras 4.17(d) y 4.18(b)). Este comportamiento es predecible pues es con los valores medios donde se presenta la mayor incertidumbre en éste y cualquier otro algoritmo.



(a) Parámetro μ de cada mezcla para la Ley de Cu Total



(b) Parámetro μ de cada mezcla para la Ley de Cu Total

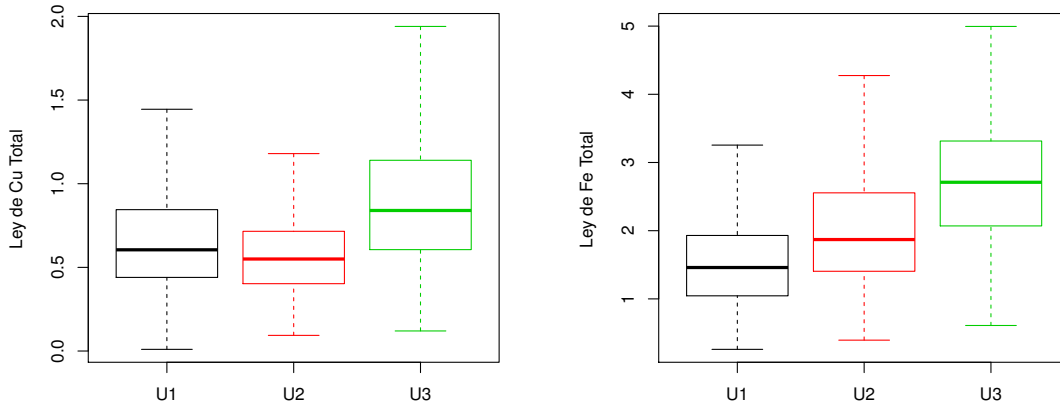
Figura 4.20: Análisis de convergencia de los parámetros μ (medias) de cada mezcla para la Ley de Cu Total y la Ley de Fe Total

	Verosimilitud para la Ley de Cu Total			Verosimilitud para la Ley de Fe Total		
	U_1	U_2	U_3	U_1	U_2	U_3
Mínimo	0.32940000	1.31400000	0.00539800	0.00000163	0.00000008	0.00000626
1er Cuartil	2.08900000	6.40300000	0.03181000	0.00001152	0.00000018	0.00082490
Mediana	2.10800000	6.67500000	0.03464000	0.00007370	0.00000019	0.00082760
Media	2.03300000	6.25900000	0.03287000	0.00010540	0.00000100	0.00072820
3er Cuartil	2.20100000	6.67500000	0.03612000	0.00020120	0.00000076	0.00082990
Máximo	2.20100000	7.34600000	0.03612000	0.00020120	0.00001166	0.00082990
Des. est.	0.31937794	1.23887411	0.00574053	0.00008518	0.00000212	0.00026230

Tabla 4.7: Descriptivos para las verosimilitudes en el caso de 3 unidades descubiertas

La Tabla 4.8 entrega los descriptivos sobre los parámetros de correlación espacial, la varianza del proceso espacial σ^2 y el alcance ρ para cada mezcla. En el caso de la variable Ley de Cu Total, las mesetas entre las mezclas 1 y 3 son semejantes (0.1864 y 0.1931), no así sus alcances que son completamente diferentes (720.3 metros y 237.4 metros). Se destaca la mezcla 2 en

este caso al tener una meseta menor y un alcance intermedio. Para la variable Ley de Fe Total todas las mezclas poseen alcances diferentes y en cuanto a la varianza del proceso espacial, las mezclas 3 y 2 tienen valores semejantes, las que corresponden a unidades con concentraciones altas y medias de ley de Fe, respectivamente (Figuras 4.17(d) y 4.18(b)).



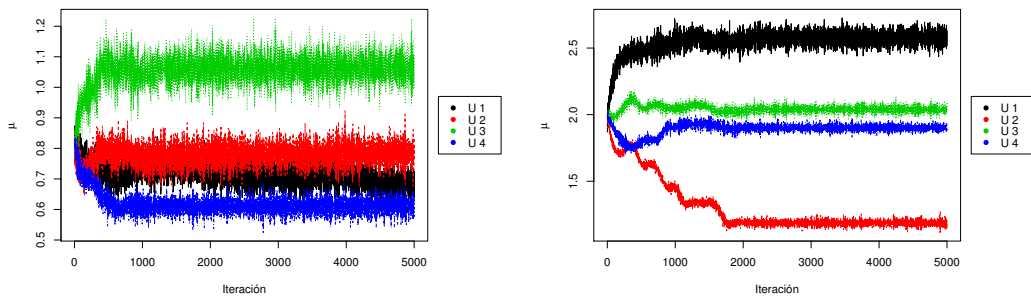
(a) Distribución a posteriori de μ de cada mezcla para la Ley de Cu Total
 (b) Distribución a posteriori de μ de cada mezcla para la Ley de Fe Total

Figura 4.21: Distribuciones a posteriori de los parámetros μ (medias) de cada mezcla para la Ley de Cu Total y la Ley de Fe Total

Parámetros de correlación espacial para la Ley de Cu Total						
	$\sigma^2 \mathbf{U}_1$	$\rho \mathbf{U}_1$	$\sigma^2 \mathbf{U}_2$	$\rho \mathbf{U}_2$	$\sigma^2 \mathbf{U}_3$	$\rho \mathbf{U}_3$
Mínimo	0.10950000	146.70000000	0.10020000	248.00000000	0.18570000	229.80000000
1er Cuartil	0.13360000	196.90000000	0.10630000	283.00000000	0.18680000	234.50000000
Mediana	0.17060000	359.60000000	0.13650000	404.30000000	0.18710000	236.10000000
Media	0.18640000	720.30000000	0.14550000	439.60000000	0.19310000	237.40000000
3er Cuartil	0.24700000	1332.00000000	0.18650000	562.50000000	0.19340000	238.90000000
Máximo	0.25330000	1396.00000000	0.19370000	1024.00000000	0.24480000	268.90000000
Des. est.	0.05610397	559.04060477	0.04065528	182.77904076	0.01272884	6.44688967
Parámetros de correlación espacial para la Ley de Fe Total						
	$\sigma^2 \mathbf{U}_1$	$\rho \mathbf{U}_1$	$\sigma^2 \mathbf{U}_2$	$\rho \mathbf{U}_2$	$\sigma^2 \mathbf{U}_3$	$\rho \mathbf{U}_3$
Mínimo	0.51770000	119.60000000	0.76810000	218.80000000	0.68720000	167.00000000
1er Cuartil	0.56800000	231.20000000	0.77770000	221.80000000	0.85010000	168.40000000
Mediana	0.62860000	249.30000000	0.80630000	358.80000000	0.85080000	189.00000000
Media	0.60690000	267.10000000	0.81780000	382.80000000	0.86730000	191.40000000
3er Cuartil	0.65040000	331.20000000	0.83590000	525.00000000	0.89880000	189.70000000
Máximo	0.69080000	340.90000000	1.05800000	732.80000000	0.96930000	300.40000000
Des. est.	0.05553523	67.99429680	0.06370714	157.95709017	0.04896562	34.89621466

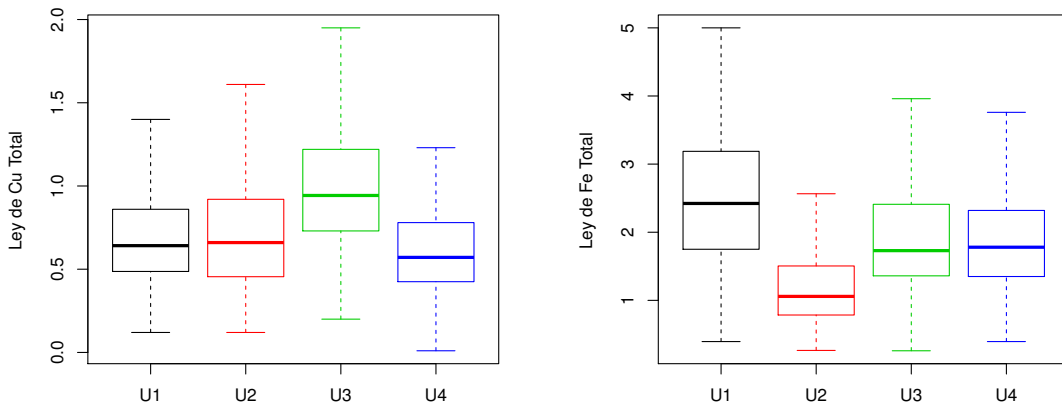
Tabla 4.8: Descriptivos de los parámetros de correlación espacial en el caso de tres unidades descubiertas

En las Figuras 4.21 y 4.23 se observan gráficos de caja los cuales reflejan la forma de las distribuciones a posteriori de los parámetros μ (medias) de cada mezcla para la Ley de Cu Total y la Ley de Fe Total en los casos de tres y cuatro unidades descubiertas respectivamente. Resultados igualmente informativos se obtienen para las cuatro mezclas, donde se destaca la rápida convergencia del algoritmo y la baja autocorrelación de las cadenas (Figuras 4.22), los diferentes grados de credibilidad de las mezclas encontradas (Tabla 4.9) y las diferencias entre mesetas y alcances para cada mezcla o unidad descubierta (Tabla 4.10). En cuanto al uso del conocimiento geológico del área de estudio como herramienta de validación y elección del número de mezclas óptimo, podemos encontrar relaciones directas entre los tipos de alteraciones hidrotermales y los resultados obtenidos para tres o cuatro mezclas con la variable Ley de Cu Total (Figuras 4.24(a), 4.24(c) y 4.24(f)).



(a) Parámetro μ de cada mezcla para la Ley de Cu Total (b) Parámetro μ de cada mezcla para la Ley de Fe Total

Figura 4.22: Análisis de convergencia de los parámetros μ (medias) de cada mezcla para la Ley de Cu Total y la Ley de Fe Total



(a) Distribución a posteriori de μ de cada mezcla para la Ley de Cu Total (b) Distribución a posteriori de μ de cada mezcla para la Ley de Fe Total

Figura 4.23: Distribuciones a posteriori de los parámetros μ (medias) de cada mezcla para la Ley de Cu Total y la Ley de Fe Total

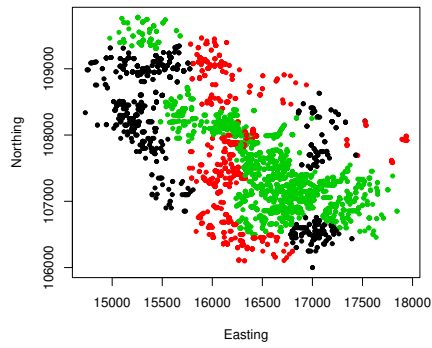
	Verosimilitud para la Ley de Cu Total				Verosimilitud para la Ley de Fe Total			
	U_1	U_2	U_3	U_4	U_1	U_2	U_3	U_4
Min.	0.20590000	0.13760000	0.04307000	0.53500000	0.00000241	0.00048330	0.00000268	0.00004270
1st Qu.	2.58700000	0.14770000	0.23160000	48.14000000	0.00012040	0.02028000	0.00000297	0.00090880
Median	3.60600000	0.14770000	0.23580000	48.14000000	0.00026070	0.12840000	0.00000297	0.00090880
Mean	3.04900000	0.17580000	0.20750000	45.73000000	0.00019940	0.09098000	0.00003425	0.00099850
3rd Qu.	3.66600000	0.17780000	0.23710000	53.86000000	0.00026070	0.12840000	0.00001915	0.00090880
Max.	3.66600000	0.35870000	0.23710000	54.01000000	0.00027760	0.12840000	0.00059220	0.00215200
sd	0.89600016	0.05425937	0.05809666	14.14595920	0.00008992	0.05479499	0.00009810	0.00042708

Tabla 4.9: Descriptivos para las verosimilitudes en el caso de cuatro unidades descubiertas

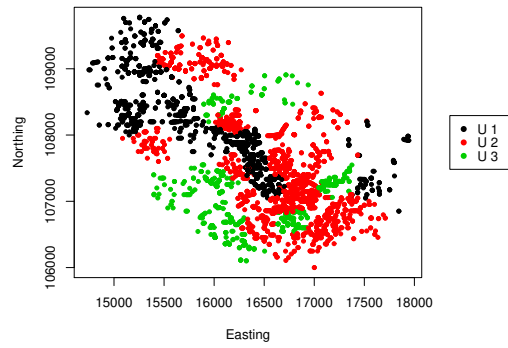
	Parámetros de correlación espacial para la Ley de Cu Total							
	$\sigma^2 U_1$	ρU_1	$\sigma^2 U_2$	ρU_2	$\sigma^2 U_3$	ρU_3	$\sigma^2 U_4$	ρU_4
Min.	0.07201000	80.58000000	0.17910000	235.50000000	0.19740000	215.80000000	0.07160000	96.51000000
1st Qu.	0.10210000	211.30000000	0.19310000	239.10000000	0.19820000	220.60000000	0.07198000	110.70000000
Median	0.10740000	382.70000000	0.19420000	240.30000000	0.19830000	222.20000000	0.07236000	178.10000000
Mean	0.10800000	432.90000000	0.19270000	247.80000000	0.20670000	229.60000000	0.07785000	167.60000000
3rd Qu.	0.11790000	694.80000000	0.19460000	243.10000000	0.19880000	224.00000000	0.07311000	186.90000000
Max.	0.16940000	803.30000000	0.20460000	358.00000000	0.25430000	322.80000000	0.16090000	354.80000000
sd	0.01858837	267.92445306	0.00502519	22.26002252	0.01909024	19.89603839	0.01890103	68.72072324
	Parámetros de correlación espacial para la Ley de Fe Total							
	$\sigma^2 U_1$	ρU_1	$\sigma^2 U_2$	ρU_2	$\sigma^2 U_3$	ρU_3	$\sigma^2 U_4$	ρU_4
Min.	0.83340000	101.80000000	0.43240000	255.90000000	0.69470000	204.70000000	0.53000000	134.40000000
1st Qu.	0.90590000	169.40000000	0.43320000	257.00000000	0.74090000	221.60000000	0.55320000	188.10000000
Median	0.91540000	171.20000000	0.47520000	267.00000000	0.74160000	223.10000000	0.56120000	198.70000000
Mean	0.91290000	180.00000000	0.50590000	292.20000000	0.74280000	224.60000000	0.57230000	196.20000000
3rd Qu.	0.92180000	201.70000000	0.52850000	283.90000000	0.74210000	224.90000000	0.57010000	199.60000000
Max.	0.94820000	273.70000000	1.08900000	900.70000000	0.81290000	279.50000000	0.80580000	344.60000000
sd	0.02615040	28.28258777	0.11641897	102.45363623	0.02001524	11.84600851	0.04622614	34.16891247

Tabla 4.10: Descriptivos de los parámetros de correlación espacial en el caso de cuatro unidades descubiertas

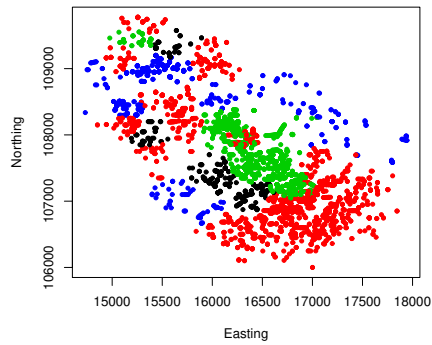
En el caso de utilizar los resultados con tres mezclas para la ley de cobre (Cu) total, se asocia la mezcla de ley media más alta (0.96 %Cu) con la alteración de tipo cuarzo-sericita y los de ley media más baja (0.61 %Cu) con la alteración de tipo sericita-clorita. Además, utilizando ya sea los resultados obtenidos con tres o cuatro mezclas, se ve reflejada la discontinuidad existente para la alteración de tipo cuarzo-sericita. En cuanto al mapa geológico del yacimiento, éste tiene una marcada relación con las tres mezclas descubiertas para la variable Ley de Fe Total (Figuras 4.24(b) y 4.24(e)). La unidad geológica 1 descubierta es aquella que presenta una menor ley media de hierro (Fe) total (1.72 %Fe) y está asociada al Complejo Intrusivo Feldespático Escondida (CIFE), que corresponde a un stock cuarzo monzonítico-granodiorítico y posee una forma elíptica (Padilla-Garza y cols. (2004)). La unidad geológica 2 descubierta es aquella que presenta una ley media de hierro (Fe) total (2.05 %Fe) intermedia y está asociada al Pórfido Riolítico de Escondida, que corresponde a un cuerpo intrusivo hipabisal que se distribuye en la zona NE y SE del rajo Escondida.



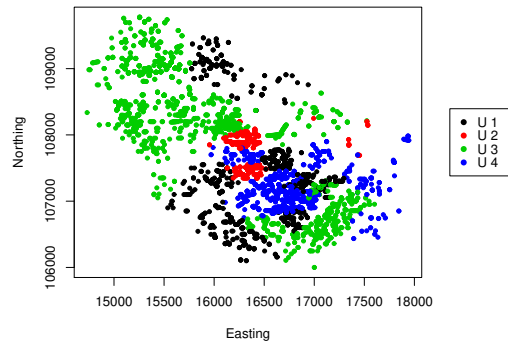
(a) Distribución en planta de las 3 mezclas para la Ley de Cu Total



(b) Distribución en planta de las 3 mezclas para la Ley de Fe Total



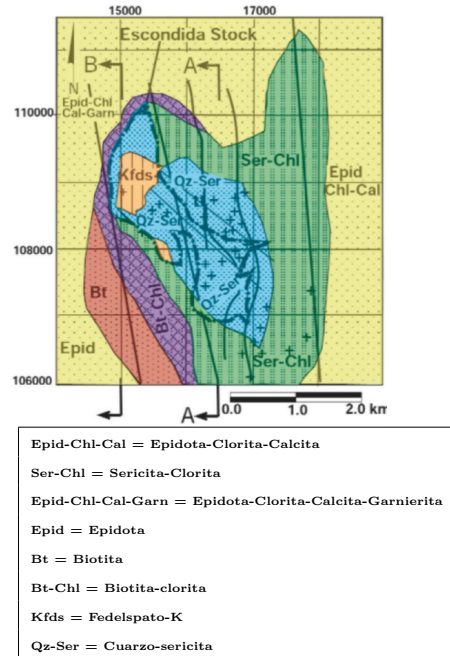
(c) Distribución en planta de las 4 mezclas para la Ley de Cu Total



(d) Distribución en planta de las 4 mezclas para la Ley de Fe Total



(e) Mapa geológico del yacimiento Escondida



(f) Mapa de alteraciones hidrotermales del yacimiento Escondida

Figura 4.24: Comparación de resultados con conocimiento geológico de la zona de estudio

La unidad geológica 3 descubierta es aquella que presenta una mayor ley media de hierro (Fe) total (2.75 %Fe) y corresponde en mayor porcentaje a lavas andesíticas. De esta forma, el número óptimo de mezclas o unidades descubiertas para la Ley de Cu Total, puede ser tres o cuatro, haciendo necesario el uso de más información para tener una decisión clara. En el caso de la Ley de Fe Total, las tres mezclas encontradas permiten describir de buena forma la distribución espacial de la variable y poseen una validación basada en el mapa geológico del yacimiento. Para más detalles acerca del caso de estudio y comparación de los resultados entre mezclas de distribuciones Gaussianas y métodos de análisis de conglomerados tradicionales se puede consultar el Anexo B

4.5. Conclusiones parciales

Refiriéndonos al caso de estudio de datos simulados, el interés fue demostrar si el algoritmo de Mezclas de Distribuciones Geoestadísticas era capaz de reproducir geometrías complejas de dominios con características similares y con una estructura de correlación espacial definida. Se plantearon tres casos de estudio, cada uno con una dificultad mayor al del anterior. Los resultados obtenidos fueron satisfactorios al momento de identificar las fronteras entre las mezclas para los tres casos, no mostrando dificultades vistas con otros algoritmos, por ejemplo, en el caso de dominios disconexos e irregulares. En cuanto al ajuste de las medias poblacionales de cada mezcla, se lograron resultados sobresalientes en el caso de estudio de dominios intrusivos, seguido del caso de dominios disconexos y finalmente el de dominios con fronteras duras. En el caso de datos simulados se obtuvieron buenos resultados, con porcentajes de clasificación correcta a cada unidad por sobre el 70 % en el peor de los casos. Se destaca además, la flexibilidad de este algoritmo en cuanto a la gran cantidad de conocimiento que es posible obtener. Es posible pensar que los resultados pueden ser condicionados a la dimensión de la partición auxiliar con la que trabaja el método, pero como ésta tiene un carácter aleatorio a lo largo de toda la cadena, este aspecto queda sobrellevado de buena forma. Con respecto al caso de estudio de datos reales, el conocimiento geológico es un aspecto fundamental a la hora de determinar la validez de los resultados y definir un número óptimo de mezclas a descubrir. Las unidades descubiertas cumplieron con el objetivo de contener muestras con características similares y con diferentes estructuras de correlación espacial, tanto para la Ley de Cu Total como para la Ley de Fe Total. La convergencia del algoritmo, la ausencia de autocorrelación, y la velocidad de convergencia son aspectos estudiados en el área de los procesos bayesianos. La metodología empleada permitió que estas características se cumplieren de buena manera, por lo que es posible extender esta idea a otras distribuciones diferentes de la Gaussiana en un trabajo futuro.

Capítulo 5

Discusiones generales

El Clustering Geoestadístico (la primera propuesta de esta tesis) corresponde a un algoritmo basado en los métodos tradicionales de clustering jerárquico, por lo que su implementación es relativamente sencilla. El aspecto desafiante del algoritmo consiste en el cálculo de la matriz de disimilitud, donde cada disimilitud entre pares de datos debe ir acompañada por una inversión de la matriz variograma evaluada en la distancia entre los pares, aspecto que resulta ser intensivo en el caso de estudios complejos. Esta matriz variograma proviene de la variografía del total de las muestras, las que a pesar de pertenecer a unidades diferentes según la hipótesis que origina este estudio, son tratadas como un sólo dominio. La variografía es un aspecto crucial dentro del algoritmo, que si no es realizada de manera prudente puede llevar a resultados erróneos. La incorporación de diversos tipos de variables resulta ser natural gracias al uso de variogramas cruzados e indicadores en la variografía. Es por esta razón que el algoritmo propuesto puede incorporar diversas fuentes de conocimiento para la formación de unidades geometalúrgicas. A pesar de todas las bondades previamente descritas, el problema de la sensibilidad ante valores extremos en las distribuciones influye en el descubrimiento de las unidades geometalúrgicas. En casos extremos se originan distancias geoestadísticas excesivas, lo que afecta a los criterios de agrupamiento en la formación de las unidades, llevando a conclusiones alejadas de la realidad. Si nos enfocamos ahora en el método de agrupamiento, en el capítulo del estado del arte, se mencionan una serie de criterios, los que generan resultados muy diferentes si son aplicados de manera indiscriminada. Por este motivo, no es posible obtener una clasificación óptima, pues ésta dependerá del criterio escogido y el sentido geo-minero-metalúrgico que se le pueda entregar a los resultados. El hecho de que se use un solo modelo de variografía para descubrir todas las unidades involucradas es una ventaja y a la vez una desventaja, pues este hecho no garantiza que las unidades formadas tengan la estructura de correlación espacial original, ni tampoco sean parecidas entre ellas. Es necesario realizar un estudio posterior al descubrimiento de las unidades geometalúrgicas para caracterizar las nuevas estructuras de correlación espacial. El aporte de esta propuesta hacia el estado del arte corresponde a la introducción de una nueva medida de disimilitud, la que está enfocada en relacionar variables regionalizadas sin la necesidad de incluir las coordenadas espaciales como variables informativas, a diferencia de un algoritmo de clustering tradicional. Esta nueva medida de disimilitud pondera en función de la varianza del proceso estacionario, la distancia de separación espacial entre pares de datos y el valor de los atributos medidos,

por lo que utiliza de manera óptima toda la información disponible en la muestra o sondeos de exploración. Si bien en la propuesta se utilizó el clustering jerárquico para los estudios, cualquier otro enfoque que utilice alguna medida de disimilitud dentro de su funcionamiento puede utilizar la Disimilitud Geoestadística, por ejemplo, algoritmos de particionado recursivo equivalentes a K-Medias o incluso otros algoritmos con objetivos diferentes. Se destaca como ejemplo de esto último, el algoritmo de Escalamiento Multidimensional ([Derndorfer y Baierl \(2014\)](#); [Borg y Groenen \(2005\)](#)), herramienta que busca realizar una reducción de la dimensionalidad en base a una representación bidimensional de una serie de atributos. Este algoritmo involucra en su funcionamiento el cálculo de distancias entre muestras, las que son obtenidas en base a la distancia Euclideana y que podría incorporar la propuesta en el caso de trabajar con variables regionalizadas. Si se cuenta con una base de datos heterotópicos (es decir, una base de datos multivariantes, donde no todas las variables están observadas en todos los datos) la única forma en la que es posible aplicar la propuesta es realizando un proceso de imputación o simulación previa para sobrellevar esta falta de información. En el primero de los casos, se tendrá una única solución, la que dependerá de los valores predichos. En el caso de realizar una simulación, podremos contar con diversos escenarios que nos entregarán un mapa de posibles unidades a descubrir. Una forma de escoger la distribución de las unidades descubiertas puede basarse en la máxima frecuencia de asignación a cada unidad, de la misma forma en la que se trabajó el caso de datos simulados.

La Mezcla de Distribuciones Geoestadísticas (la segunda propuesta de esta tesis) corresponde a un algoritmo basado en la teoría de mezcla de distribuciones ([Bilmes y cols. \(1998\)](#)), incorporando la perspectiva Bayesiana y Geoestadística. El algoritmo de esta propuesta resulta ser intuitivo y rápido de implementar, pues las operaciones involucradas no revisten mayor complejidad y están directamente relacionadas con la dimensionalidad del problema. Este último aspecto, es sobrellevado de buena forma gracias a la propuesta de particionado inicial, la que reduce los tiempos y carga computacional, factores recurrentes que muchas veces impiden el desarrollo de nuevos procedimientos en Geoestadística Bayesiana ([Cornford y cols. \(2005\)](#); [Yan y cols. \(2007\)](#)). Con la primera propuesta, la forma de las unidades geometalúrgicas descubiertas depende del método de agrupamiento escogido. Debido a esto, es posible obtener desde unidades con una alta convexidad a unidades dispersas y poco regulares en el espacio. La Mezcla de Distribuciones Geoestadísticas no presenta esta condición, pues el principio con el que actúa depende única y exclusivamente de la maximización iterativa de la función de verosimilitud a posteriori, independiente de la configuración en el espacio de las muestras. Es por esta razón que, utilizando esta propuesta, es posible descubrir unidades geometalúrgicas de toda índole: regulares, dispersas, con fronteras duras, intrusivas, disconexas, etc. La capacidad en la definición de las fronteras depende del algoritmo de particionado auxiliar del dominio, planteado en la propuesta. Una de las diferencias principales entre las propuestas es la forma en la que se utiliza la estructura de correlación espacial de los datos. En el caso del Clustering Geoestadístico, esta estructura es única durante todo el algoritmo y de ella depende la medida de disimilitud entre los pares de datos. En el caso de las Mezclas de Distribuciones Geoestadísticas tendremos tantas estructuras de correlación espacial como mezclas queramos encontrar. Estas estructuras evolucionan durante el procedimiento iterativo de maximización de la verosimilitud y son estimadas vía mínimos cuadrados según el enfoque clásico de ajuste de variogramas. Tanto en el caso de estudio de datos simulados así como en el de datos reales se trabajó con una sola variable regionalizada. De esta forma, en cada

iteración se tuvo que realizar el ajuste de un solo variograma experimental para cada mezcla. El problema es fácilmente extrapolable al caso multivariable. En este caso, se trabajará con variogramas directos y cruzados y con una matriz de varianza-covarianza, en lugar de una varianza escalar. Se podría también generalizar el algoritmo a casos donde la parametrización del variograma es más compleja, por ejemplo, cuando se utilizan modelos anisótropos o modelos anidados, en cuyo caso existirían varios rangos de correlación y/o varianzas parciales. Este procedimiento no reviste mayores complejidades a la hora de realizar los cálculos y la literatura existente permite el desarrollo de los pasos necesarios para lograrlo (Emery (2010)). Ya que la propuesta tiene un carácter bayesiano, es posible incorporar muchos de los desarrollos logrados en la actualidad en términos de eficiencia computacional (Wilkinson (2006)). En el caso del cómputo en paralelo, debido a que por iteración se van actualizando las pertenencias de cada partición auxiliar de manera independiente, este proceso se puede paralelizar para lograr una eficiencia mayor. Además, es conveniente utilizar esta misma idea de cálculo en paralelo para generar diversas cadenas de manera simultánea, lo que permitiría reducir el número de iteraciones necesarias para alcanzar la convergencia del algoritmo. El método se basa en el supuesto de que las fronteras entre conglomerados son “duras”, es decir, no existe correlación entre los valores encontrados en un dominio y en otro. Este supuesto puede no cumplirse en ciertas aplicaciones reales, donde suele existir correlación entre conglomerados (fronteras “blandas” o difusas). Si bien, el algoritmo es computacionalmente eficiente y existe un rango de mejora sustancial, hay una rutina que es imposible de evitar y que consume un gran tiempo de cómputo por iteración: el cálculo del variograma experimental por iteración. Este cálculo es necesario realizarlo debido a que las pertenencias a una u otra mezcla van evolucionando con cada iteración, por lo que es necesario realizar este cálculo tantas veces como número de iteraciones, número de grillas auxiliares y número de mezclas. A diferencia del Clustering Geoestadístico, en el caso de la Mezcla de Distribuciones Geoestadísticas no es posible incorporar variables de tipo categórica, por lo que a la hora de aplicar la propuesta en el descubrimiento de unidades geometalúrgicas, éstas no pueden incorporar variables tan importantes como lo son: los tipos de roca, alteraciones, edades geológicas, entre otras. Los aportes de esta propuesta son variados y pertenecen a diversas áreas del estado del arte. Desde el punto de vista bayesiano, este algoritmo afronta el problema del clustering desde la perspectiva de utilizar todo el conocimiento previo que se tenga y actualizarlo en base a la información disponible. Existen trabajos previos (Heller y Ghahramani (2005)) que operan bajo este mismo enfoque, pero que por problemas de dimensionalidad se ven muchas veces frustrados. Aquí radica la virtud de la propuesta, la que utilizando un paso extra dentro del algoritmo tradicional, es capaz de sobrellevar esta compleja situación. Otra de las virtudes de la propuesta es que los resultados obtenidos bajo esta perspectiva, entregan una solución de tipo probabilística y no única. De esta forma, se puede tener un grado mayor de flexibilidad ante la carencia de información y realizar estudios de validación de resultados que cuantifiquen el riesgo a la hora de definir si una observación pertenece a una mezcla u otra. Una de las condiciones de la propuesta planteada fue el uso de las distribuciones Gaussianas para la formación de la verosimilitud. Si bien la propuesta se limitó al uso de esta distribución, esto no representa una restricción para el uso de otras distribuciones diferentes a la Gaussiana. Así como en el caso de las mezclas de distribuciones beta (Ji y cols. (2005)), es posible extender la propuesta para que utilice de manera adecuada estas nuevas distribuciones dentro de un contexto geoestadístico (Lagos-Álvarez y cols. (2016)).

Capítulo 6

Conclusiones

Propuesta 1: Clustering Geoestadístico

En cuanto a la propuesta de Clustering Geoestadístico, ésta permitió descubrir las unidades geometalúrgicas presentes en la región de estudio tanto para el caso de datos simulados como reales.

Existe un gran número de fuentes de variabilidad que condicionan los resultados. Por esta razón, se aconseja la inclusión de variables geológicas categóricas de manera directa en la construcción de la medida de disimilitud, la que permitirá obtener mayor conocimiento y control sobre la forma en cómo se generan las unidades por el método de Clustering Geoestadístico.

También podemos concluir que, a mayor número de fronteras en común, mayor será el error de clasificación a la hora de definir las fronteras que separan las unidades geometalúrgicas.

La pregunta acerca del número ideal de unidades sigue sin respuesta única, número tal que el experto (geólogo o geometalurgista) sea capaz de interpretar las unidades descubiertas y relacionarlas con características tales como litología, mineralización o alteración.

La existencia de continuidad espacial, la falta de información exhaustiva que provoca incertidumbre y la interpretación de variables regionalizadas en términos de funciones aleatorias, hacen que se desaconseje el uso de métodos tradicionales de clustering, donde las coordenadas son ignoradas, o son consideradas como si fueran variables adicionales.

Propuesta 2: Mezcla de Distribuciones Geoestadísticas

En cuanto al caso de estudio de datos simulados para el algoritmo de Mezclas de Distribuciones Geoestadísticas, se pudo demostrar que éste fue capaz de reproducir geometrías complejas de dominios con características similares y con una estructura de correlación espacial definida.

En cuanto al ajuste de las medias poblacionales de cada mezcla, se lograron resultados sobresalientes en el caso de estudio de dominios intrusivos, seguido del caso de dominios disconexos y finalmente el de dominios con fronteras duras, aspecto relacionado con el carácter bayesiano de la propuesta.

Se destaca además la flexibilidad de este algoritmo en cuanto a la gran cantidad de conocimiento que es posible incluir y obtener. Es posible pensar que los resultados pueden ser condicionados a la dimensión de la partición auxiliar con la que trabaja el método, pero como ésta tiene un carácter aleatorio a lo largo de toda la cadena, este aspecto queda sobrellevado de buena forma.

Con respecto al caso de estudio de datos reales, las unidades descubiertas cumplieron con el objetivo de contener muestras con características similares y con diferentes estructuras de correlación espacial. La convergencia del algoritmo, la ausencia de autocorrelación y la velocidad de convergencia son aspectos estudiados en el área de los procesos bayesianos.

La metodología empleada permitió que estas características se cumpliesen de buena manera, por lo que es posible extender esta idea a otras distribuciones diferentes de la Gaussiana en un trabajo futuro.

Perspectivas para futuras investigaciones

- Relacionadas con el Clustering Geoestadístico
 - Utilizar otro enfoque que utilice alguna medida de disimilitud dentro de su funcionamiento puede utilizar la Disimilitud Geoestadística, por ejemplo, algoritmos de particionado recursivo equivalentes a K-Medias o incluso otros algoritmos con objetivos diferentes. Se destaca como ejemplo de esto último, el algoritmo de Escalamiento Multidimensional (Derndorfer y Baierl (2014); Borg y Groenen (2005)).
- Relacionadas con la Mezcla de Distribuciones Geoestadísticas
 - Extrapolar al caso multivariable, donde se debiese ajustar en cada iteración un modelo de corregionalización lineal para cada mezcla, basándose en la literatura existente (Emery (2010); Goulard y Voltz (1992)).
 - Incorporar el caso del cómputo en paralelo, para generar diversas cadenas de manera simultánea, lo que permitiría reducir el número de iteraciones necesarias para alcanzar la convergencia del algoritmo.
 - Hacer uso de otras distribuciones diferentes a la Gaussiana para la construcción de la verosimilitud. Basarse como punto inicial en el caso de las mezclas de distribuciones Beta (Bouguila y cols. (2006); Ji y cols. (2005)), la que es posible extender a la propuesta utilizando de manera adecuada la definición de los parámetros de posición y escala de estas distribuciones dentro de un contexto geoestadístico.

Bibliografía

- Allard, D., y Guillot, G. (2000). Clustering geostatistical data. En *Proceedings of the sixth geostatistical conference*.
- Ambroise, C., Dang, M., y Govaert, G. (1997). Clustering of spatial data by the EM algorithm. En *geoENV I—Geostatistics for environmental applications* (pp. 493–504). Springer.
- Anderberg, M. R. (1973). *Cluster analysis for applications. Monographs and textbooks on probability and mathematical statistics*. Academic Press, Inc., New York.
- Bahoura, M., y Pelletier, C. (2004). Respiratory sounds classification using cepstral analysis and gaussian mixture models. En *Engineering in medicine and biology society, 2004. iembs'04. 26th annual international conference of the ieee* (Vol. 1, pp. 9–12).
- Bailey, K. (1994). *Typologies and Taxonomies: An Introduction to Classification Techniques*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-102. Thousand Oaks, CA: Sage.
- Baker, F. B., y Hubert, L. J. (1976). A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering. *Journal of the American Statistical Association*, 71(356), 870–878.
- Ball, G. H., y Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Behavioral science*, 12(2), 153–155.
- Balzarini, M., Macchiavelli, R., y Casanoves, F. (2004). Aplicaciones de modelos mixtos en agricultura y forestería. *Curso de Capacitacion Centro Agronomico Tropical de Investigación y Enseñanza-CATIE*.
- Batanero, C., y Batanero, M. C. D. (2008). *Análisis de datos con statgraphics*. Departamento de Didáctica de la Matemática, Universidad de Granada.
- Bilmes, J. A., y cols. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510), 126.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

- Bonner, R. E. (1964). On some clustering techniques. *IBM journal of research and development*, 8(1), 22–32.
- Borg, I., y Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.
- Bouguila, N., Ziou, D., y Monga, E. (2006). Practical bayesian estimation of a finite beta mixture through gibbs sampling and its applications. *Statistics and Computing*, 16(2), 215–225.
- Box, G. E., y Tiao, G. C. (2011). *Bayesian inference in statistical analysis* (Vol. 40). John Wiley & Sons.
- Bozdogan, H. (1987). Model selection and akaike’s information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3), 345–370.
- Burnham, K., y Anderson, D. (2002). Information and likelihood theory: a basis for model selection and inference. *Model selection and multimodel inference: a practical information-theoretic approach*, 2, 49–97.
- Calegario, N., Maestri, R., Leal, C. L., y Daniels, R. F. (2005). Estimativa do crescimento de povoamentos de eucalyptus baseada na teoria dos modelos não lineares em multinível de efeito misto. *Ciência Florestal*, 15(3), 285–292.
- Caliński, T., y Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1), 1–27.
- Cantrell, C. D. (2000). *Modern mathematical methods for physicists and engineers*. Cambridge University Press.
- Carrero, O. (2008). Ajuste de curvas de índice de sitio mediante modelos mixtos para plantaciones de eucalyptus urophylla en venezuela. *Interciencia*, 33(4), 265–272.
- Celeux, G., y Govaert, G. (1992). A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3), 315–332.
- Celeux, G., Hurn, M., y Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451), 957–970.
- Chen, S., y Gopalakrishnan, P. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. En *Proc. darpa broadcast news transcription and understanding workshop* (Vol. 8, pp. 127–132).
- Chen, Y., CAM, E., Welling, M., y EDU, U. (2014). Austerity in mcmc land: Cutting the metropolis-hastings budget.
- Chib, S., y Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4), 327–335.

- Chilès, J.-P., y Delfiner, P. (2012). Geostatistics: Modeling spatial uncertainty, second edition. *Wiley series in probability and statistics*, 705–714.
- Cornford, D., Csató, L., y Opper, M. (2005). Sequential, bayesian geostatistics: a principled method for large data sets. *Geographical Analysis*, 37(2), 183–199.
- Cressie, N. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, 17(5), 563–586.
- Dagnelie, P. (1975). *L'analyse statistique à plusieurs variables*. Gembloux, Gembloux: Presses agron.
- Dempster, A., Laird, N., y Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*. 39, 1-38.
- Derndorfer, E., y Baierl, A. (2014). Multidimensional scaling (mds). *Mathematical and Statistical Methods in Food Science and Technology*, 175–186.
- Deutsch, C. (2013). Geostatistical modelling of geometallurgical variables—problems and solution. En *GEOMET'2013: the Second AUSIMM International Geometallurgy Conference*.
- Deza, M. M., y Deza, E. (2009). Encyclopedia of distances. En *Encyclopedia of Distances* (pp. 1–583). Springer.
- Dhillon, I. S., y Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2), 143–175.
- Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297–302.
- Diebolt, J., y Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 363–375.
- Diggle, P. J., Menezes, R., y Su, T.-I. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2), 191–232.
- Dixon, W. (1953). Processing data for outliers. *Biometrics*, 9(1), 74–89.
- Do-Jong, K., Yong-Woon, P., y Dong-Jo, P. (2001). A novel validity index for determination of the optimal number of clusters. *IEICE Transactions on Information and Systems*, 84(2), 281–285.
- Duda, R. O., Hart, P. E., y Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.
- Edwards, A. L. (1976). An introduction to linear regression and correlation.
- Emery, X. (2010). Iterative algorithms for fitting a linear model of coregionalization. *Com-*

- puters & Geosciences*, 36(9), 1150–1160.
- Estivill-Castro, V., y Yang, J. (2000). Fast and robust general purpose clustering algorithms. En *Pacific Rim International Conference on Artificial Intelligence* (pp. 208–218).
- Everitt, B. (1974). *Cluster Analysis*. London: Heinemann Educ. Books.
- Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21, 768–769.
- Fortier, J., y Solomon, H. (1996). Clustering procedures. En *Proceedings of the Multivariate Analysis, '66, P.R. Krishnaiah (Ed.)* (pp. 493–506).
- Fraley, C., y Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer journal*, 41(8), 578–588.
- François, O., Ancelet, S., y Guillot, G. (2006). Bayesian clustering using hidden markov random fields in spatial population genetics. *Genetics*, 174(2), 805–816.
- Gelfand, A. E., y Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410), 398–409.
- Gelman, A., y Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.
- Gilks, W. R., Richardson, S., y Spiegelhalter, D. (1995). *Markov chain monte carlo in practice*. CRC press.
- Gilks, W. R., Richardson, S., y Spiegelhalter, D. J. (1996). Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice*, 1, 19.
- Gordon, A. (1999). Classification. *Monographs on Statistics and Applied Probability*, 82.
- Goulard, M., y Voltz, M. (1992). Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, 24(3), 269–286.
- Guha, S., Rastogi, R., y Shim, K. (1998). CURE: an efficient clustering algorithm for large databases. En *ACM SIGMOD Record* (Vol. 27, pp. 73–84).
- Hair, J. F., Anderson, R. E., Tatham, R. L., y William, C. (1998). *Black (1998), Multivariate data analysis*. Upper Saddle River, NJ: Prentice Hall.
- Hallewell, M. (2009). Geometallurgy for mine data. *Materials World*, 17(7), 48–50.
- Han, J. (2001). *Kamber., M.: Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers.
- Hardle, W., y Simar, L. (2007). *Applied Multivariate Statistical Analysis*. Springer Berlin Heidelberg.

- Hartigan, J. (1975). *Clustering algorithms*. John Wiley and Sons.
- Heller, K. A., y Ghahramani, Z. (2005). Bayesian hierarchical clustering. En *Proceedings of the 22nd international conference on machine learning* (pp. 297–304).
- Hervada-Sala, C., y Jarauta-Bragulat, E. (2004). A program to perform Ward’s clustering method on several regionalized variables. *Computers & Geosciences*, 30(8), 881–886.
- Hoal, K., Woodhead, J., y Smith, K. (2013). The importance of mineralogical input into geo-metallurgy programs. En *Proceedings of the Second AusIMM International Geometallurgy Conference. Brisbane, Australia* (pp. 17–26).
- Hofmann-Wellenhof, B., y Moritz, H. (2006). *Physical geodesy*. Springer Science & Business Media.
- Horn, P. S., Feng, L., Li, Y., y Pesce, A. J. (2001). Effect of outliers and nonhealthy individuals on reference interval estimation. *Clinical Chemistry*, 47(12), 2137–2145.
- Hubert, L., y Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1), 193–218.
- Hurvich, C. M., y Tsai, C.-L. (1991). Bias of the corrected| mathromaic criterion for under-fitted regression and time series models. *Biometrika*, 499–509.
- Isaaks, E. H., y Srivastava, R. M. (1989). An introduction to applied geostatistics. *New York, USA: Oxford University Press*, 23, 345–383.
- Jaccard, P. (1900). *Contribution au Probleme de L’Immigration Post-Glaciere de la Flore Alpine: Etude comporative de la flore alpine du massif du Wildhorn. du haut bassin du Trient et de la haute vallée de Bagnes*. Corbaz et Cie.
- Jaccard, P. (1901). *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz.
- Jaccard, P. (1908). *Nouvelles recherches sur la distribution florale*.
- Jain, A. K., Murty, M. N., y Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3), 264–323.
- Ji, Y., Wu, C., Liu, P., Wang, J., y Coombes, K. R. (2005). Applications of beta-mixture models in bioinformatics. *Bioinformatics*, 21(9), 2118–2122.
- Jolliffe, I. (1986). *Principal Component Analysis*. Springer-Verlag.
- King, B. (1967). Step-wise clustering procedures. *Journal of the American Statistical Association*, 62(317), 86–101.
- Kozumi, H., y Kobayashi, G. (2011). Gibbs sampling methods for bayesian quantile regression. *Journal of statistical computation and simulation*, 81(11), 1565–1578.

- Kravchenko, A. (2003). Influence of spatial structure on accuracy of interpolation methods. *Soil Science Society of America Journal*, 67(5), 1564–1571.
- Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1–27.
- Kuiper, F. K., y Fisher, L. (1975). 391: A Monte Carlo comparison of six clustering procedures. *Biometrics*, 777–783.
- Lagos-Álvarez, B. M., Fustos-Toribio, R., Figueroa-Zúñiga, J., y Mateu, J. (2016). Geostatistical mixed beta regression: a bayesian approach. *Stochastic Environmental Research and Risk Assessment*, 1–14.
- Lance, G., y Williams, W. (1967). A general theory of classification sorting strategies: 1= hierarchical systems, 2= clustering systems. *Computer Journal*, 9–10.
- Lantuéjoul, C. (2013). *Geostatistical simulation: models and algorithms*. Springer Science & Business Media.
- Legendre, P., y Legendre, L. F. (2012). *Numerical ecology* (Vol. 24). Elsevier.
- MacEachern, S. N., y Müller, P. (1998). Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2), 223–238.
- Macnaughton-Smith, P., Williams, W., Dale, M., y Mockett, L. (1964). Dissimilarity analysis: a new technique of hierarchical sub-division.
- MacQueen, J., y cols. (1967). Some methods for classification and analysis of multivariate observations. En *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297).
- Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2, 49–55.
- Mardia, K. V., y Kent, J. (1979). Bibby. JM Multivariate analysis. *London: Academic*.
- McRae, D. (1971). MIKCA: A FORTRAN IV Iterative K-means Cluster Analysis Program, CTB. *McGraw Hill*, 19, 70.
- Miller, F., Vandome, A., y John, M. (2010). *Geometallurgy*. VDM Publishing. Descargado de <https://books.google.cl/books?id=tqrobwAACAAJ>
- Moellering, H., y Tobler, W. (1972). Geographical variances. *Geographical Analysis*, 4(1), 34–50.
- Molinero, L. (2002). El método bayesiano en la investigación médica. *Liga española para la lucha contra la hipertensión arterial*, 3–10.
- Molinero, L. (2003). ¿ qué es el método de estimación de máxima verosimilitud y cómo se interpreta. *Liga Española para la lucha contra la Hipertensión Arterial*, 1.

- Møller, J., Pettitt, A. N., Reeves, R., y Berthelsen, K. K. (2006). An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 451–458.
- Mountford, M. (1962). An index of similarity and its application to classificatory problems. *Progress in soil zoology*, 43, 50.
- Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4), 354–359.
- Oliver, M., y Webster, R. (1989). A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology*, 21(1), 15–35.
- Orlóci, L. (2013). *Multivariate analysis in vegetation research*. Springer.
- Padilla, R., Titley, S., y Pimentel, F. (2001). Geology of the escondida porphyry copper deposit. *Antofagasta region, Chile: Economic Geology*, 96, 307–324.
- Padilla-Garza, R., Titley, S., y Eastoe, C. (2004). Hypogene evolution of the escondida porphyry copper deposit, chile. *Soc Econ Geol Spec Publ*, 11, 141–165.
- Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185, 71–110.
- Phillips, J. D. (1997). *Potential-field geophysical software for the pc, version 2.2*. Citeseer.
- Plummer, M., Best, N., Cowles, K., y Vines, K. (2006). Coda: convergence diagnosis and output analysis for mcmc. *R news*, 6(1), 7–11.
- Raftery, A. E., y Lewis, S. M. (1996). Implementing mcmc. *Markov chain Monte Carlo in practice*, 115–130.
- Rajesh, D. G., y Punithavalli, M. (2014). Wavelets and gaussian mixture model approach for gender classification using fingerprints. En *Current trends in engineering and technology (icctet), 2014 2nd international conference on* (pp. 522–525).
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336), 846–850.
- Ribeiro Jr, P. J., Diggle, P. J., y cols. (2001). geor: a package for geostatistical analysis. *R news*, 1(2), 14–18.
- Richards, J. P., Boyce, A. J., y Pringle, M. S. (2001). Geologic evolution of the escondida area, northern chile: a model for spatial and temporal localization of porphyry cu mineralization. *Economic Geology*, 96(2), 271–305.
- Rivoirard, J. (1987). Two key parameters when choosing the kriging neighborhood. *Mathematical geology*, 19(8), 851–856.
- Robert, C., y Casella, G. (1999). *Monte Carlo statistical methods*. Springer series in statistics.

Springer-Verlag. New York.

- Romary, T., Rivoirard, J., Deraisme, J., Quinones, C., y Freulon, X. (2012). Domaining by clustering multivariate geostatistical data. En *Geostatistics oslo 2012* (pp. 455–466). Springer.
- Rosales, D. (2012). *Implementación de metodología para determinar dominios geometalúrgicos de estimación* (Tesis de Master no publicada). Universidad de Chile.
- Seber, G. A. (2009). *Multivariate observations* (Vol. 252). John Wiley & Sons.
- Selim, S. Z., y Ismail, M. A. (1984). K-means-type algorithms: a generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence*(1), 81–87.
- Shaw, W., S., K., y A., W. (2013). Modelling Geometallurgical Variability – A Case Study in Managing Risk. En *The Second AUSIMM International Geometallurgy Conference*.
- Shental, N., Bar-Hillel, A., Hertz, T., y Weinshall, D. (2004). Computing gaussian mixture models with em using equivalence constraints. En *Advances in neural information processing systems* (pp. 465–472).
- Siegel, S. (1972). Diseño Experimental No Paramétrico: Las medidas de correlación y sus pruebas de significación. El coeficiente de correlación de rangos de Spearman. *Diseño Experimental No Paramétrico: Las medidas de correlación y sus pruebas de significación. El coeficiente de correlación de rangos de Spearman*.
- Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4), 35–43.
- Sneath, P. H., Sokal, R. R., y cols. (1973). *Numerical taxonomy. The principles and practice of numerical classification*.
- Strehl, A., y Ghosh, J. (2000). A scalable approach to balanced, high-dimensional clustering of market-baskets. En *International Conference on High-Performance Computing* (pp. 525–536).
- Strehl, A., Ghosh, J., y Mooney, R. (2000). Impact of similarity measures on web-page clustering. En *Workshop on Artificial Intelligence for Web Search (AAAI 2000)* (pp. 58–64).
- Tibshirani, R., Walther, G., y Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 411–423.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, 1701–1728.
- Tryon, R. (1970). *Bailey, De Cluster Analysis*. McGraw-Hill, New York.

- Vallejos, R., Mallea, A., Herrera, M., y Ojeda, S. (2015). A multivariate geostatistical approach for landscape classification from remotely sensed image data. *Stochastic environmental research and risk assessment*, 29(2), 369–378.
- Veysieres, M., y Plant, R. E. (1998). Identification of vegetation state and transition domains in California’s hardwood rangelands. *University of California*, 101.
- Wackernagel, H. (2003). *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media.
- Wang, X., y Yu, Q. (2001). *Estimate the number of clusters in web documents via gap statistic*. May.
- Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301), 236–244.
- Weldon, W. F. R. (1889). The variations occurring in certain decapod crustacea.–i. crangon vulgaris. *Proceedings of the Royal Society of London*, 47(286-291), 445–453.
- Wilkinson, D. J. (2006). Parallel bayesian computation. *Statistics Textbooks and Monographs*, 184, 477.
- Williams, G. (2012). *Linear algebra with applications*. Jones & Bartlett Publishers.
- Wolfe, J. H. (1967). *Normix: Computational methods for estimating the parameters of multivariate normal mixtures of distributions* (Inf. Téc.). DTIC Document.
- Yan, J., Cowles, M. K., Wang, S., y Armstrong, M. P. (2007). Parallelizing mcmc for bayesian spatiotemporal geostatistical models. *Statistics and Computing*, 17(4), 323–335.
- Zhang, T., Ramakrishnan, R., y Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. En *ACM Sigmod Record* (Vol. 25, pp. 103–114).
- Zhu, X. (2007). *Knowledge Discovery and Data Mining: Challenges and Realities: Challenges and Realities*. Igi Global.

Anexo A

Caso de estudio de geoquímica, sector Colchane, I Región de Tarapacá

Las Figuras A.1 y A.2 muestran los resultados del caso de estudio aplicando Conglomerados Geoestadísticos con el método de Ward, las Figuras A.3 y A.4 muestran los resultados del caso de estudio con el método de la distancia máxima y las Figuras A.5 y A.6 muestran los resultados del caso de estudio con el método de la distancia mínima.

A.1. Resultados del caso de estudio aplicando Conglomerados Geoestadísticos con el método de Ward

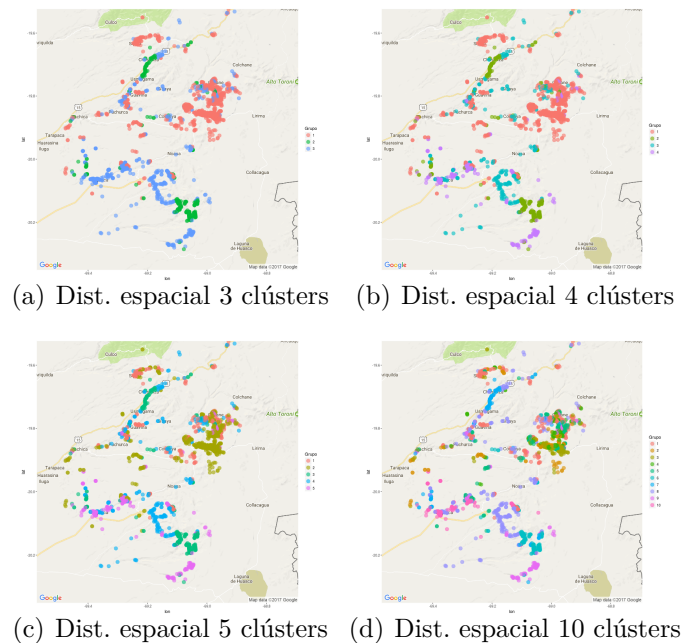
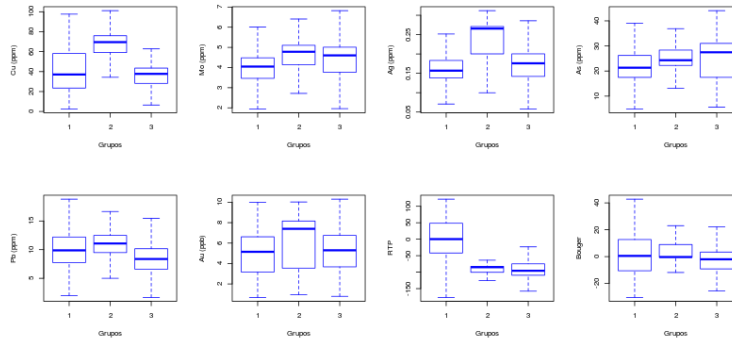
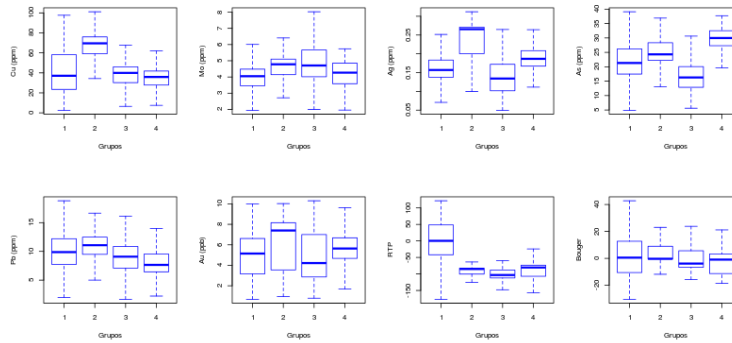


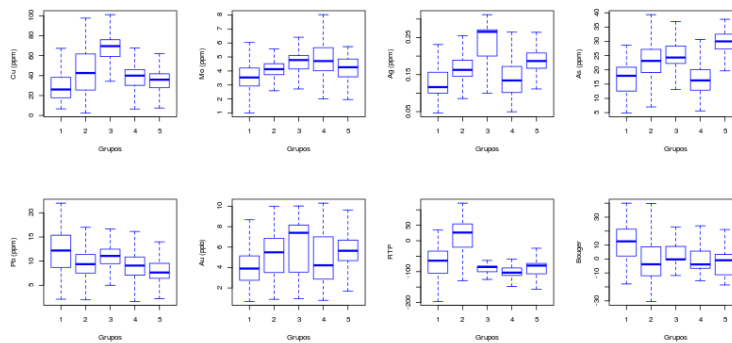
Figura A.1: Distribución espacial para los clústers formados por el Método de Ward



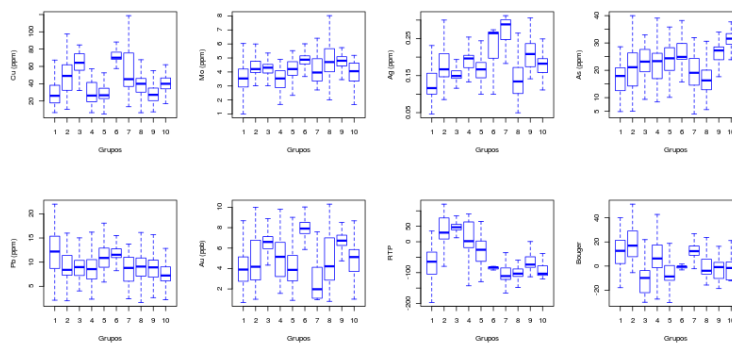
(a) Dist. de las variables para los 3 clústers



(b) Dist. de las variables para los 4 clústers



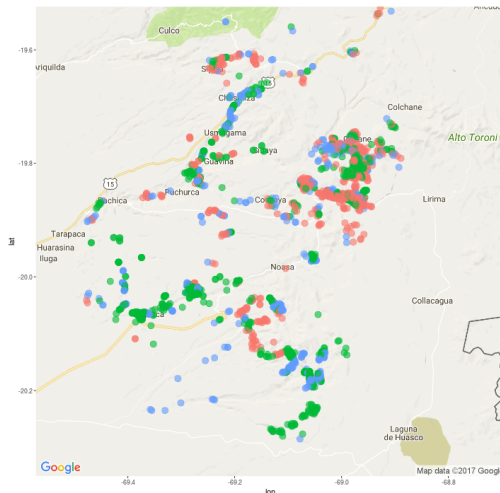
(c) Dist. de las variables para los 5 clústers



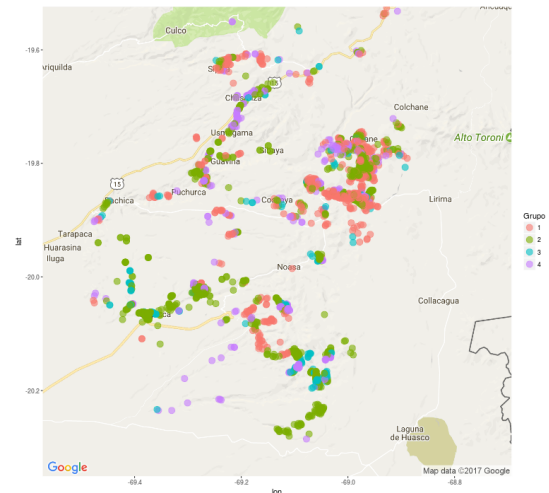
(d) Dist. de las variables para los 10 clústers

Figura A.2: Distribución de las variables de estudio en función de los clústers formados por el Método de Ward

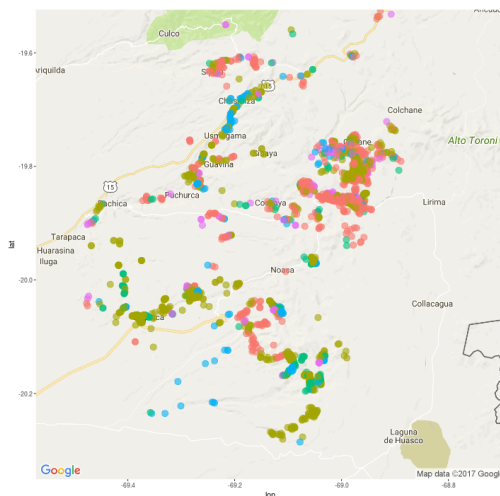
A.2. Resultados del caso de estudio aplicando Conglomerados Geoestadísticos con el método de la distancia máxima (Complete linkage)



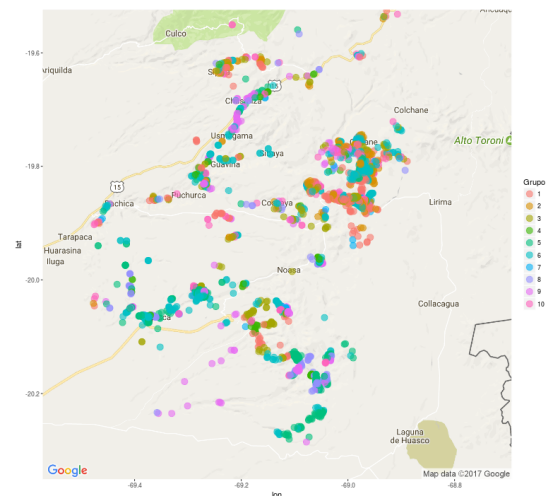
(a) Dist. espacial 3 clústers



(b) Dist. espacial 4 clústers

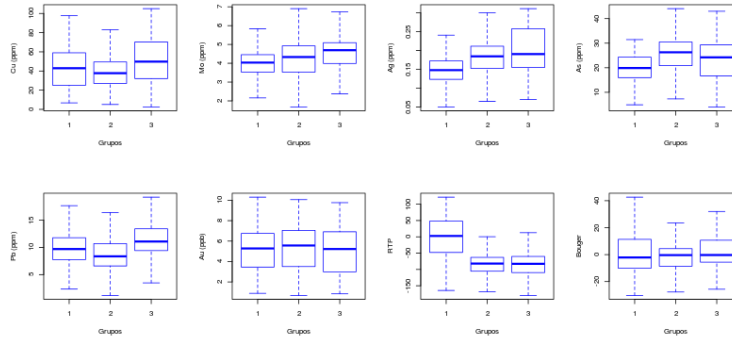


(c) Dist. espacial 5 clústers

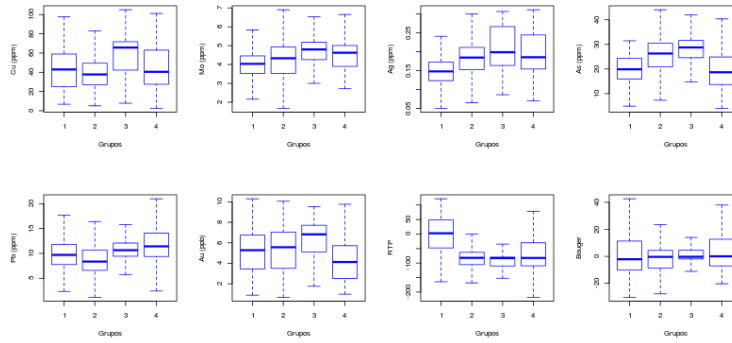


(d) Dist. espacial 10 clústers

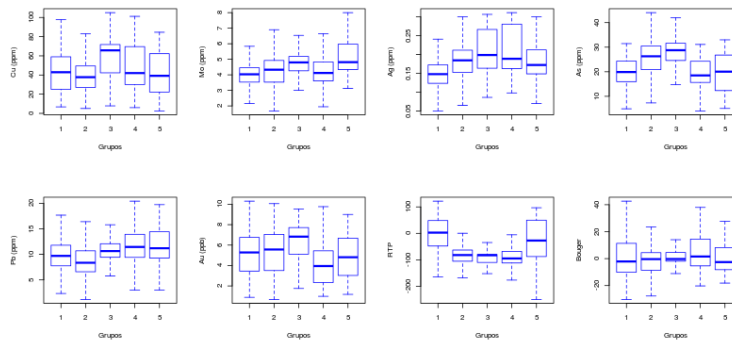
Figura A.3: Distribución espacial para los clústers formados por el método de la distancia máxima (Complete linkage)



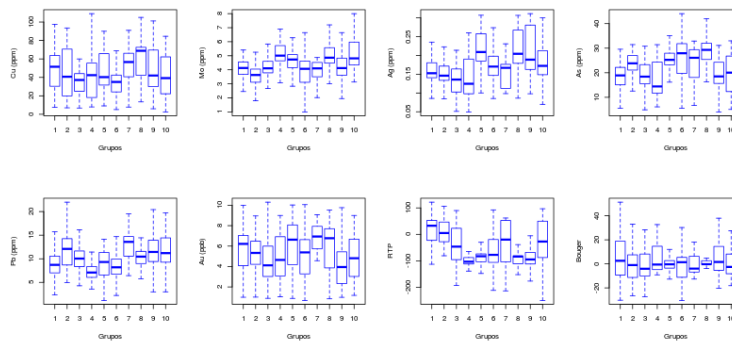
(a) Dist. de las variables para los 3 clústers



(b) Dist. de las variables para los 4 clústers



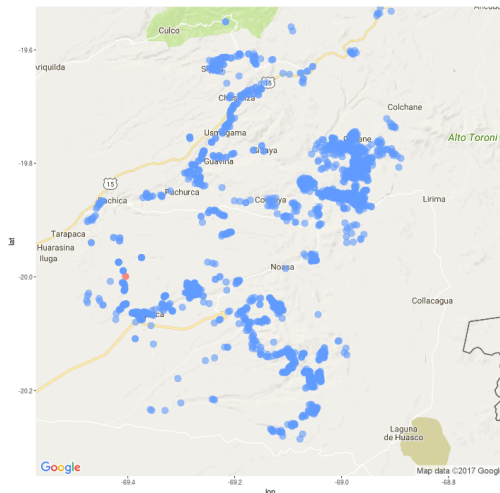
(c) Dist. de las variables para los 5 clústers



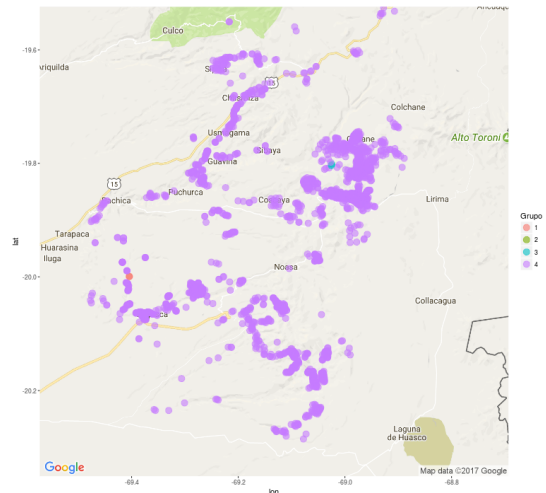
(d) Dist. de las variables para los 10 clústers

Figura A.4: Distribución de las variables de estudio en función de los clústers formados por el método de la distancia máxima (Complete linkage)

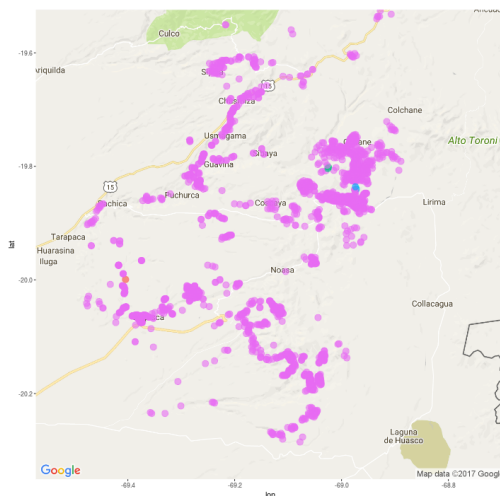
A.3. Resultados del caso de estudio aplicando Conglomerados Geoestadísticos con el método de la distancia mínima (Single linkage)



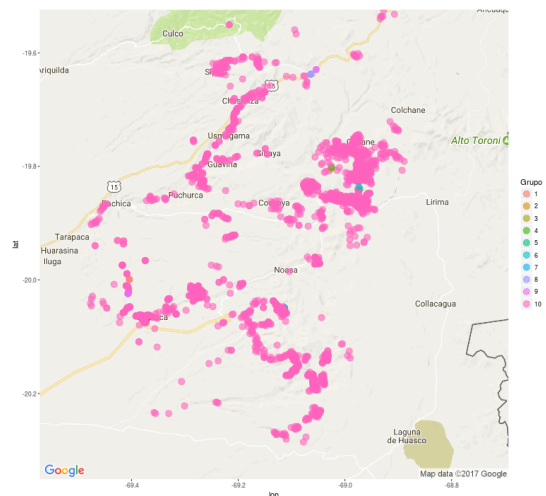
(a) Dist. espacial 3 clústers



(b) Dist. espacial 4 clústers

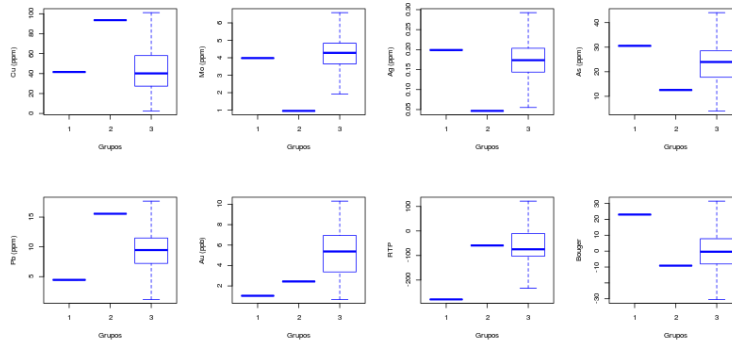


(c) Dist. espacial 5 clústers

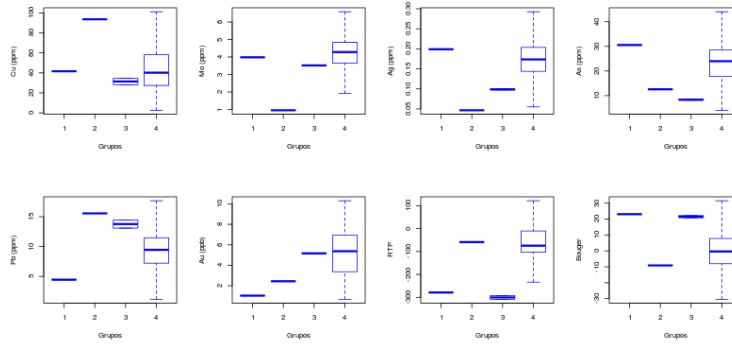


(d) Dist. espacial 10 clústers

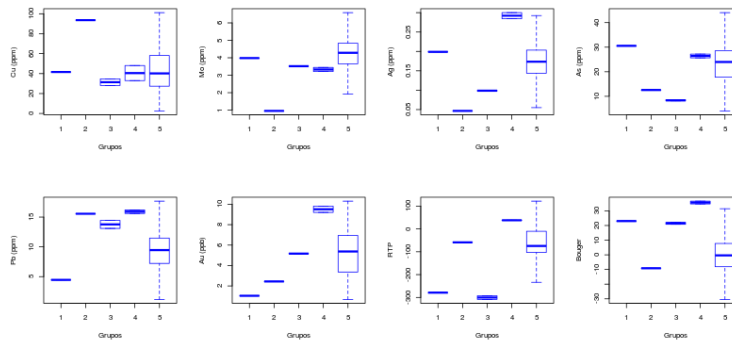
Figura A.5: Distribución espacial para los clústers formados por el método de la distancia mínima (Single linkage)



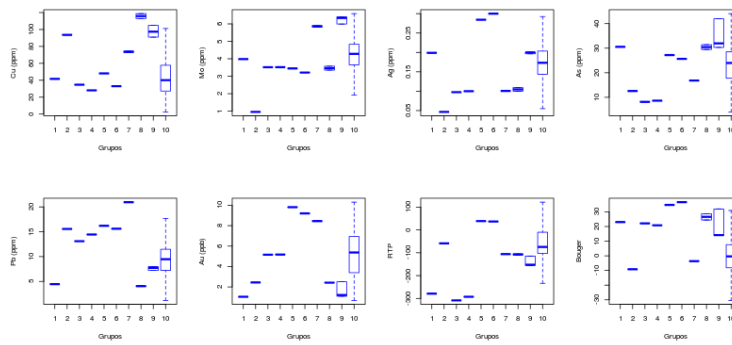
(a) Dist. de las variables para los 3 clústers



(b) Dist. de las variables para los 4 clústers



(c) Dist. de las variables para los 5 clústers



(d) Dist. de las variables para los 10 clústers

Figura A.6: Distribución de las variables de estudio en función de los clústers formados por el método de la distancia mínima (Single linkage)

Las Figuras A.7 y A.8 muestran los resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de Ward sin coordenadas espaciales, las Figuras A.9 y A.10 muestran los resultados del caso de estudio con el método de la distancia máxima y las Figuras A.11 y A.12 muestran los resultados del caso de estudio con el método de la distancia mínima.

A.4. Resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de Ward sin coordenadas espaciales

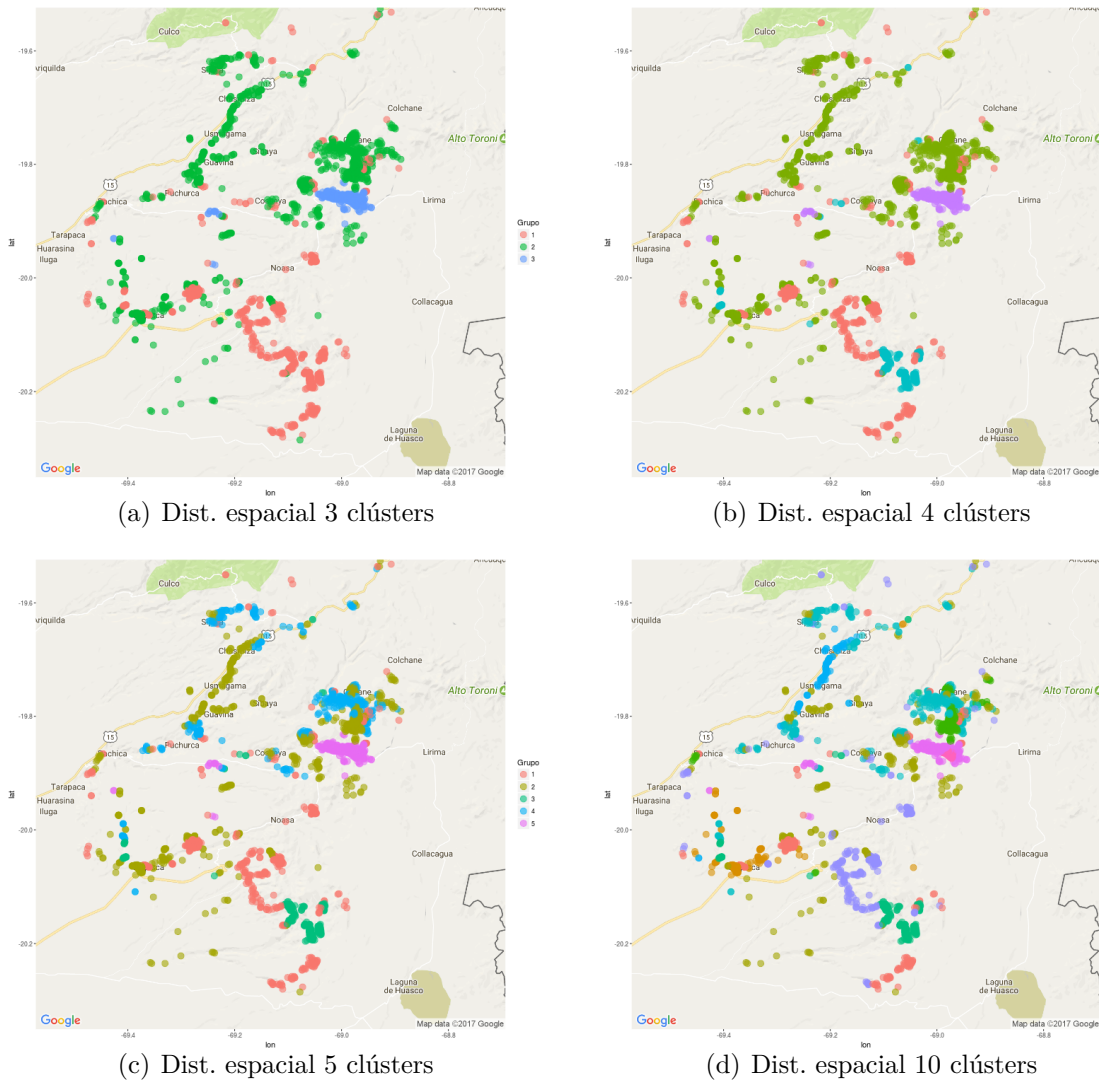
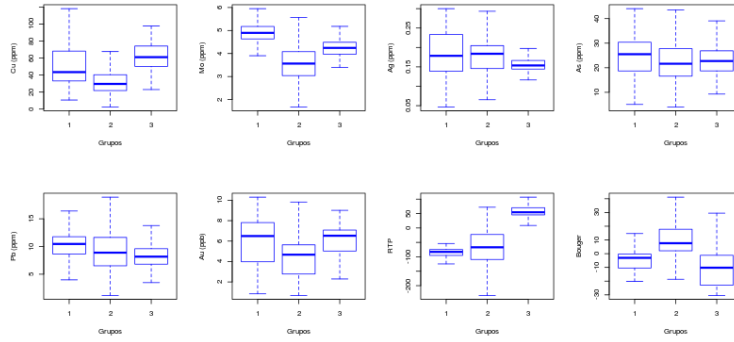
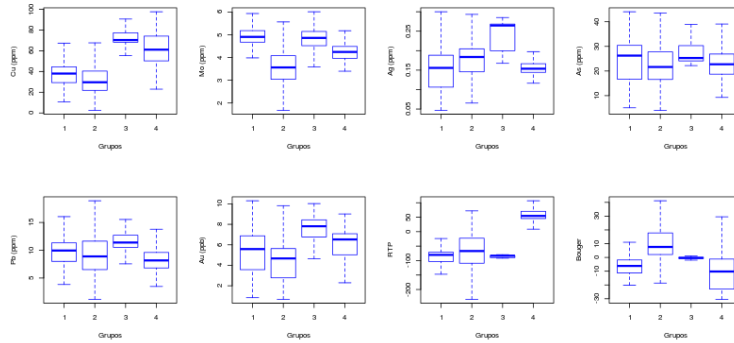


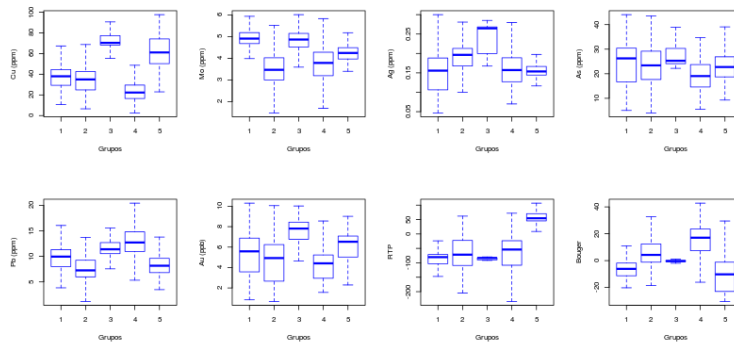
Figura A.7: Distribución espacial para los clústers formados por el Método de Ward



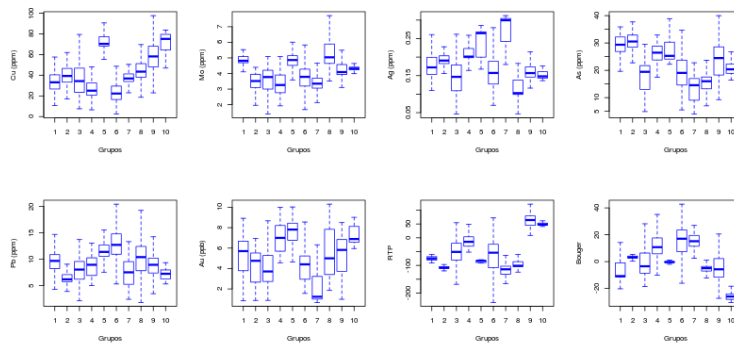
(a) Dist. de las variables para los 3 clústers



(b) Dist. de las variables para los 4 clústers



(c) Dist. de las variables para los 5 clústers



(d) Dist. de las variables para los 10 clústers

Figura A.8: Distribución de las variables de estudio en función de los clústers formados por el Método de Ward

A.5. Resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de la distancia máxima (Complete linkage) sin coordenadas espaciales

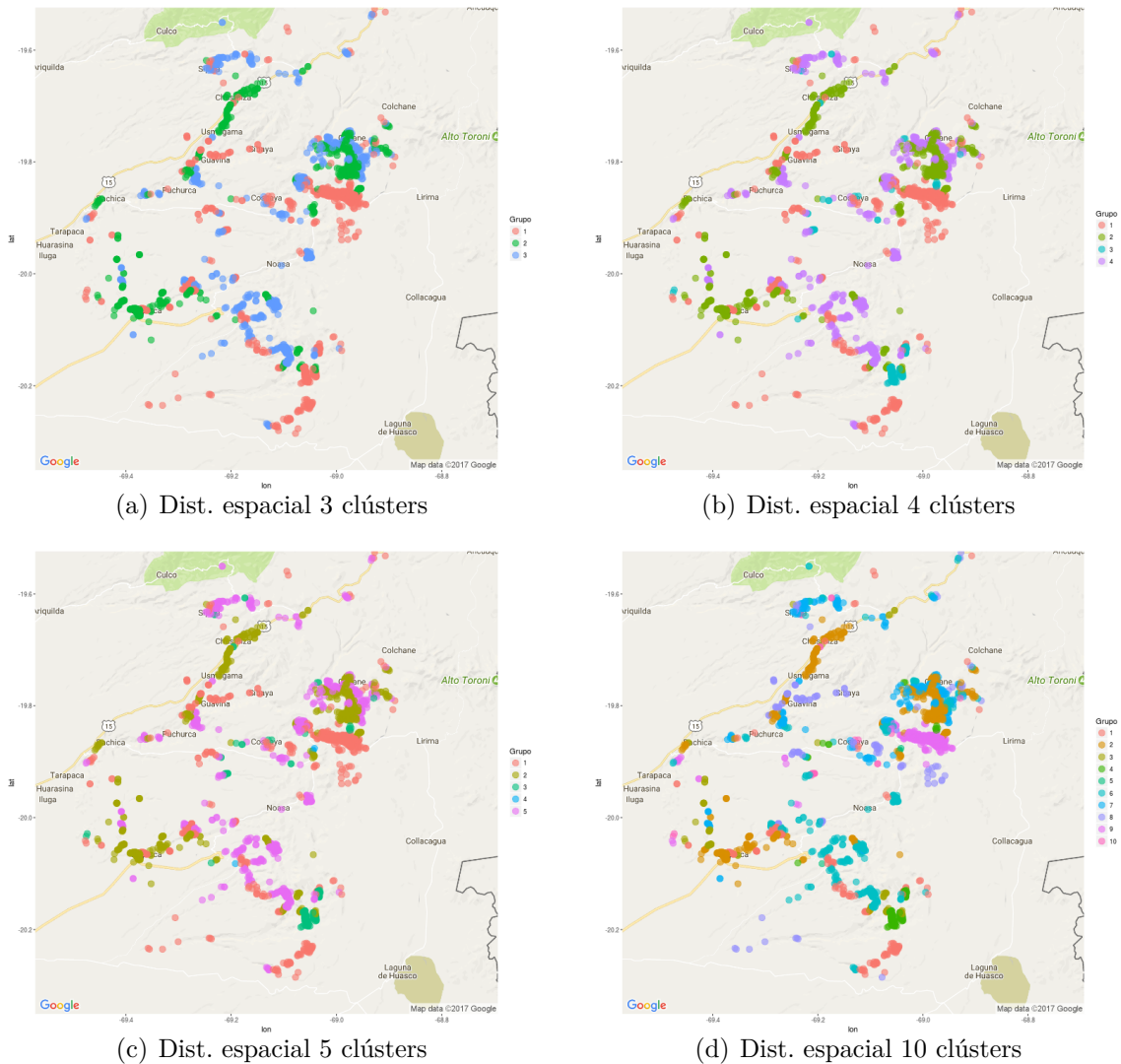
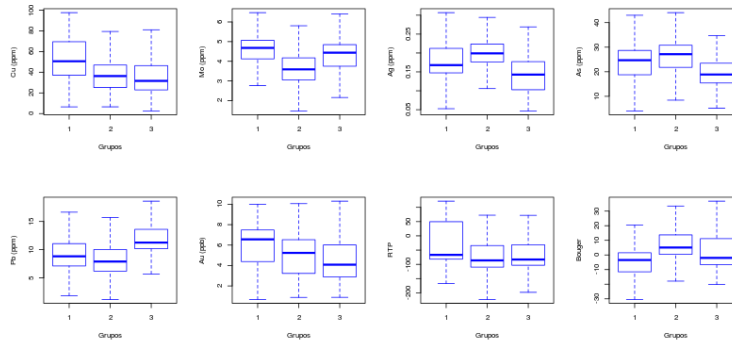
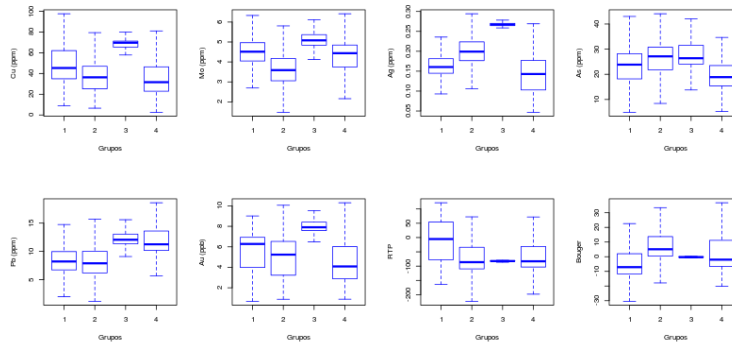


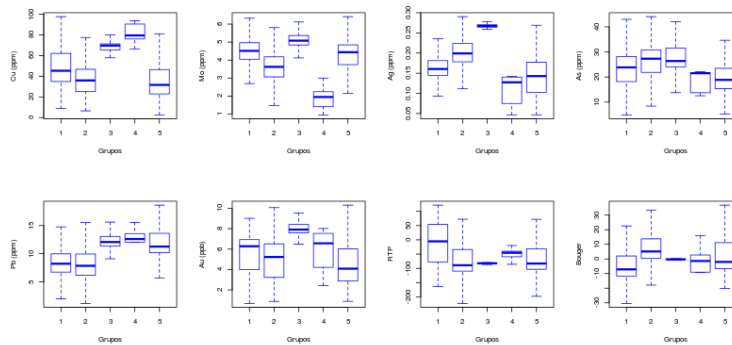
Figura A.9: Distribución espacial para los clústers formados por el método de la distancia máxima (Complete linkage)



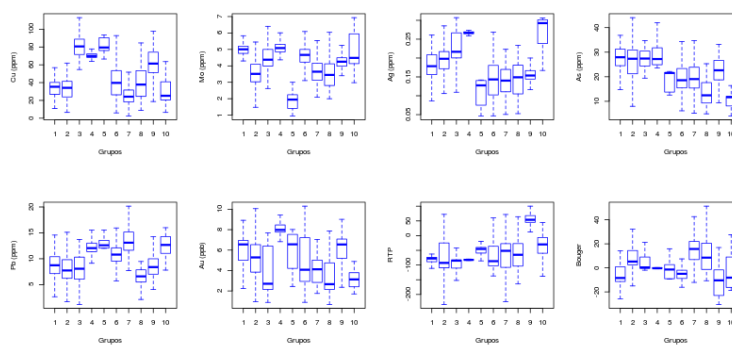
(a) Dist. de las variables para los 3 clústers



(b) Dist. de las variables para los 4 clústers



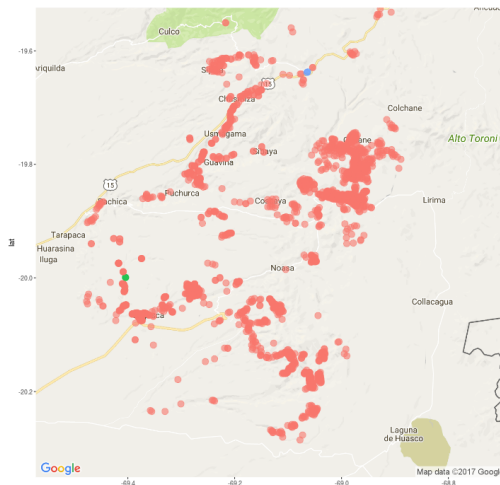
(c) Dist. de las variables para los 5 clústers



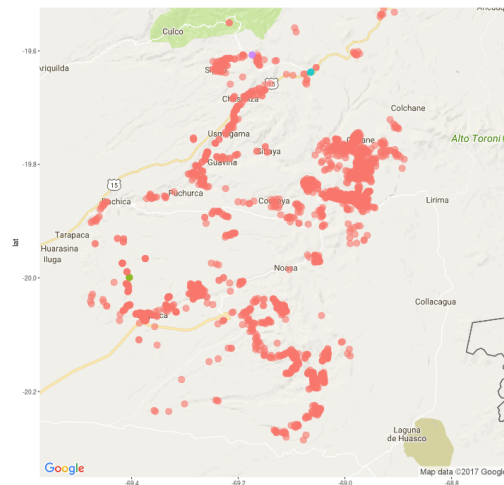
(d) Dist. de las variables para los 10 clústers

Figura A.10: Distribución de las variables de estudio en función de los clústers formados por el método de la distancia máxima (Complete linkage)

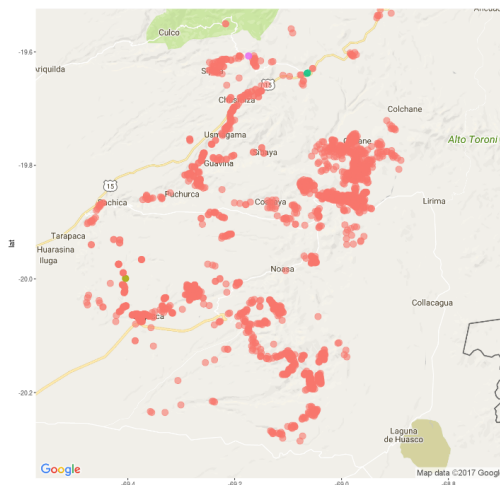
A.6. Resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de la distancia mínima (Single linkeage) sin coordenadas espaciales



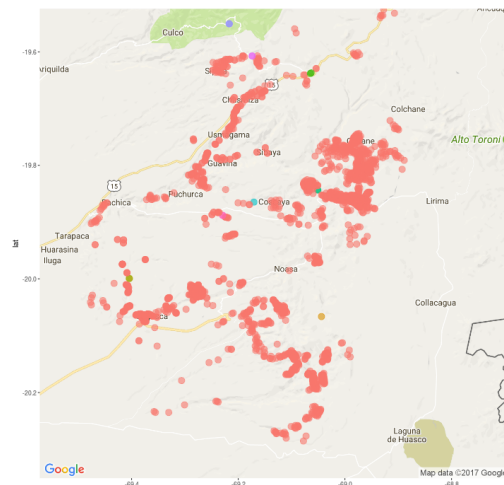
(a) Dist. espacial 3 clústers



(b) Dist. espacial 4 clústers

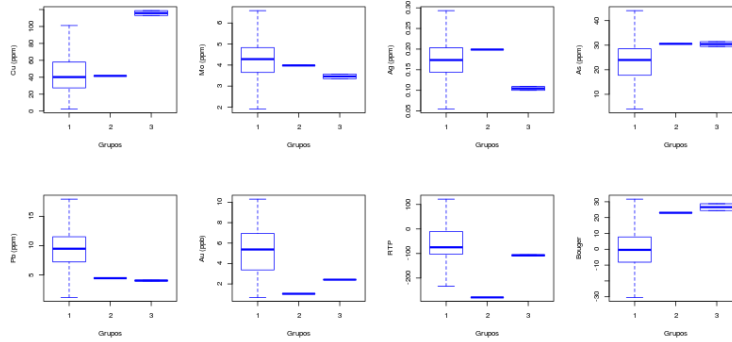


(c) Dist. espacial 5 clústers

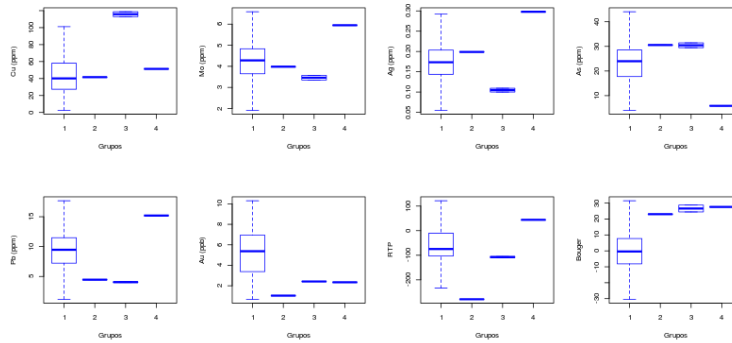


(d) Dist. espacial 10 clústers

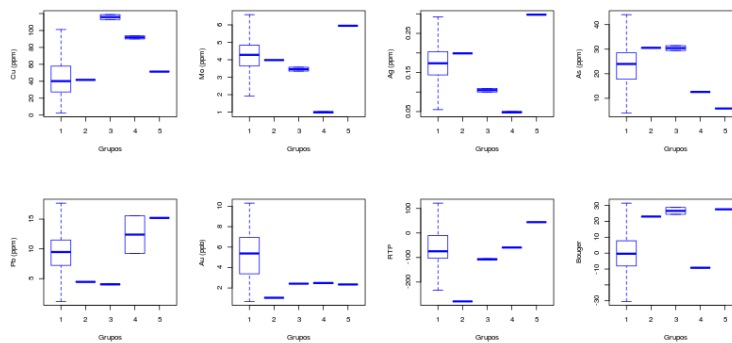
Figura A.11: Distribución espacial para los clústers formados por el método de la distancia mínima (Single linkeage)



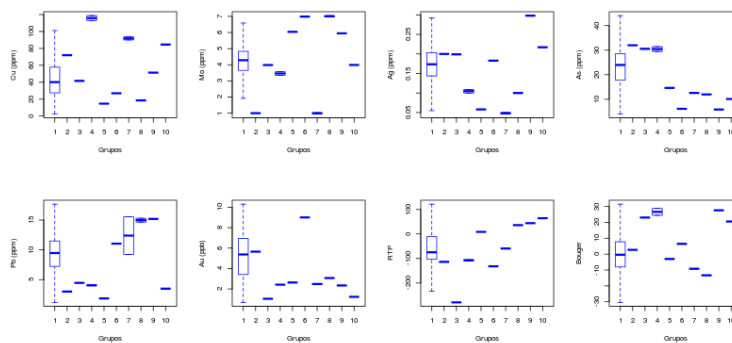
(a) Dist. de las variables para los 3 clústers



(b) Dist. de las variables para los 4 clústers



(c) Dist. de las variables para los 5 clústers



(d) Dist. de las variables para los 10 clústers

Figura A.12: Distribución de las variables de estudio en función de los clústers formados por el método de la distancia mínima (Single linkage)

Las Figuras A.13 y A.14 muestran los resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de Ward con coordenadas espaciales, las Figuras A.15 y A.16 muestran los resultados del caso de estudio con el método de la distancia máxima y las Figuras A.17 y A.18 muestran los resultados del caso de estudio con el método de la distancia mínima.

A.7. Resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de Ward con coordenadas espaciales

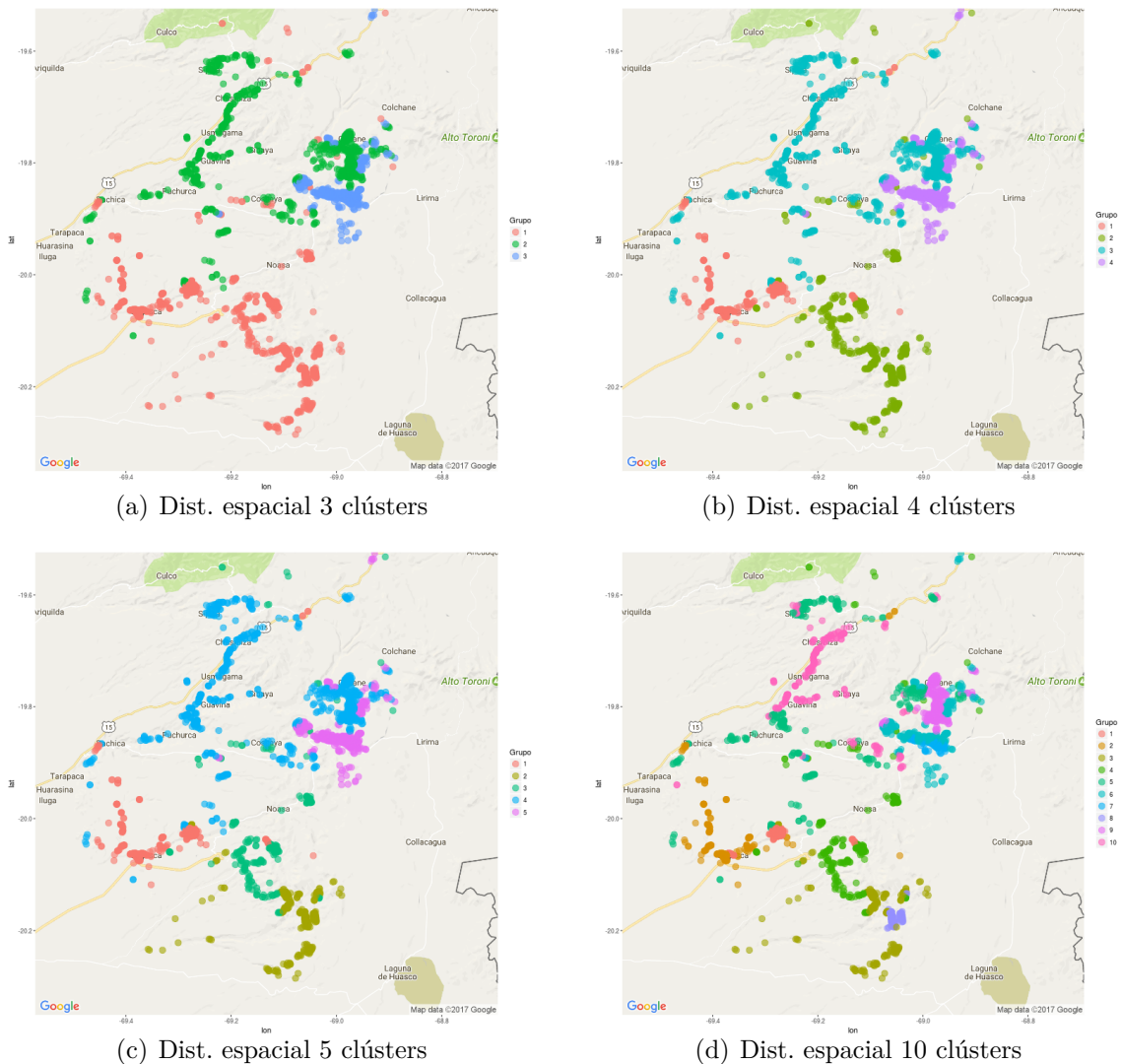
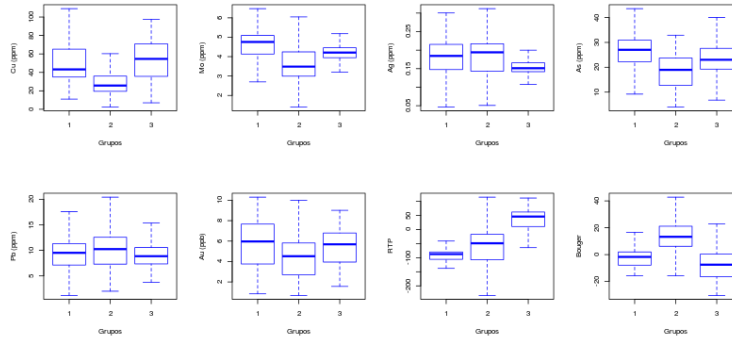
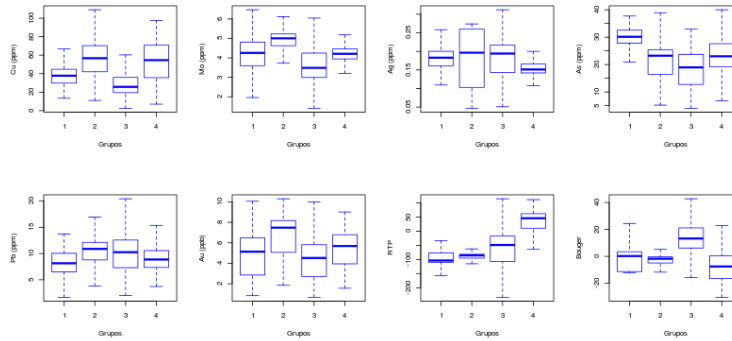


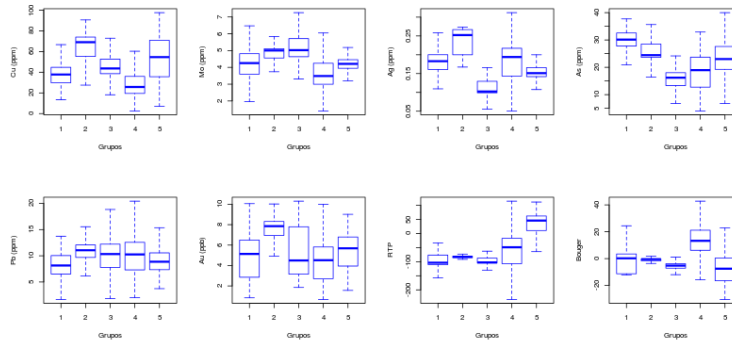
Figura A.13: Distribución espacial para los clústers formados por el Método de Ward



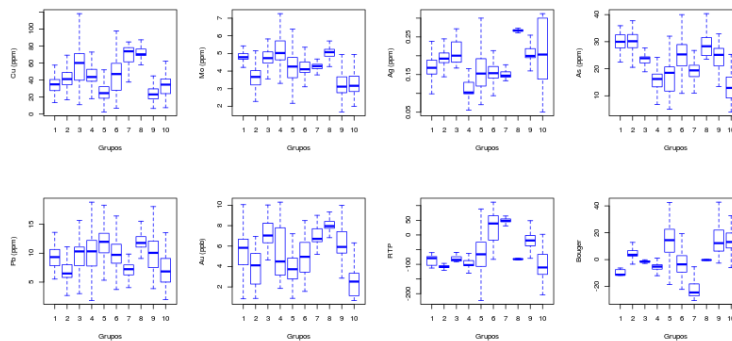
(a) Dist. de las variables para los 3 clústers



(b) Dist. de las variables para los 4 clústers



(c) Dist. de las variables para los 5 clústers



(d) Dist. de las variables para los 10 clústers

Figura A.14: Distribución de las variables de estudio en función de los clústers formados por el Método de Ward

A.8. Resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de la distancia máxima (Complete linkeage) con coordenadas espaciales

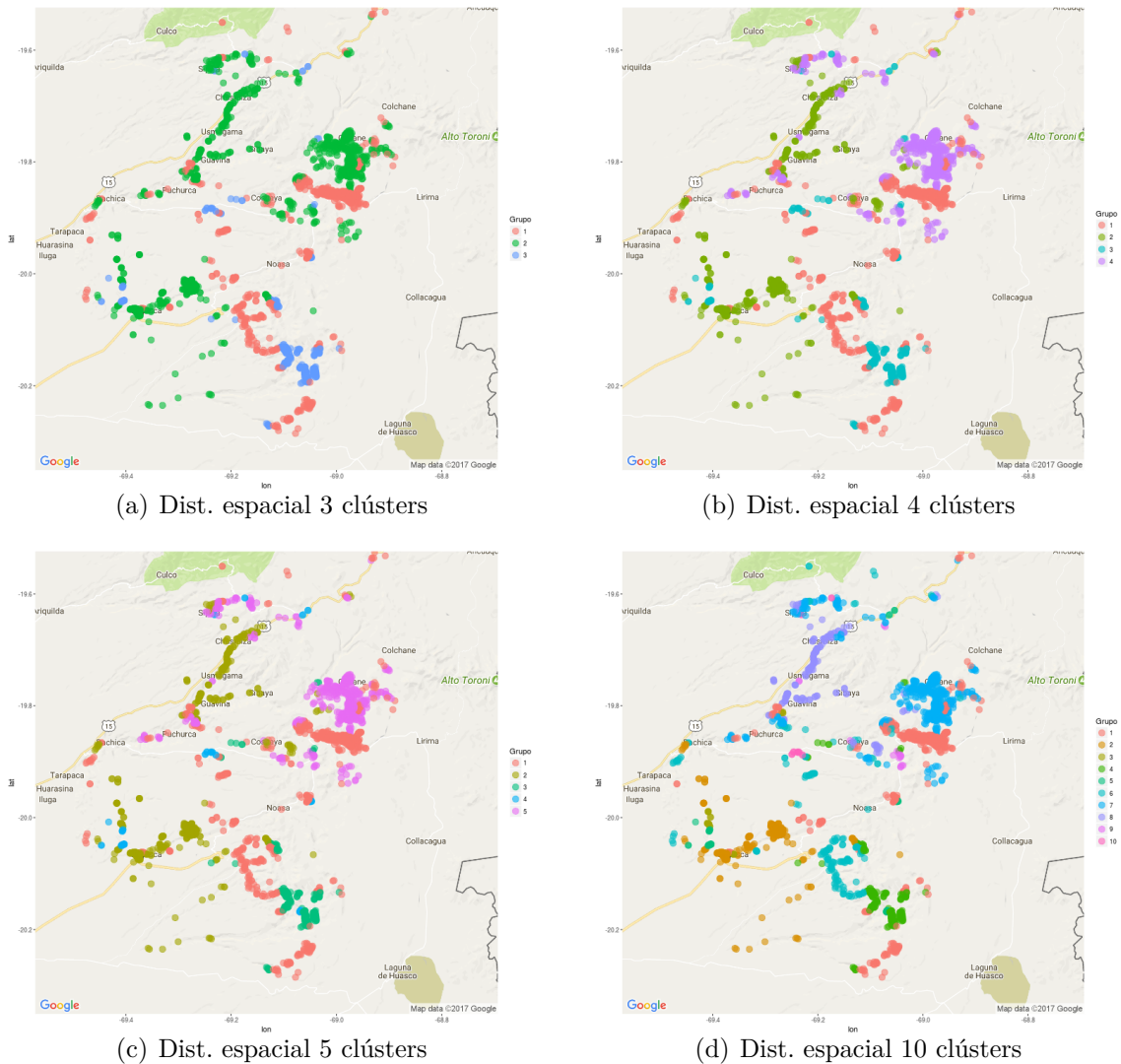
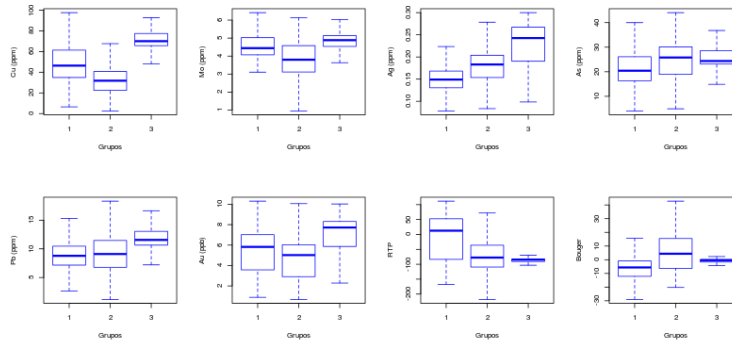
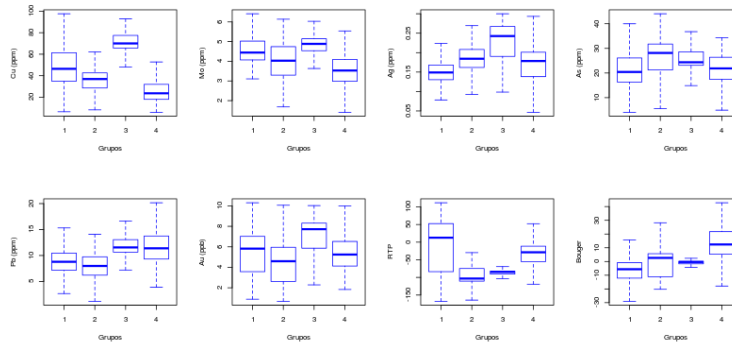


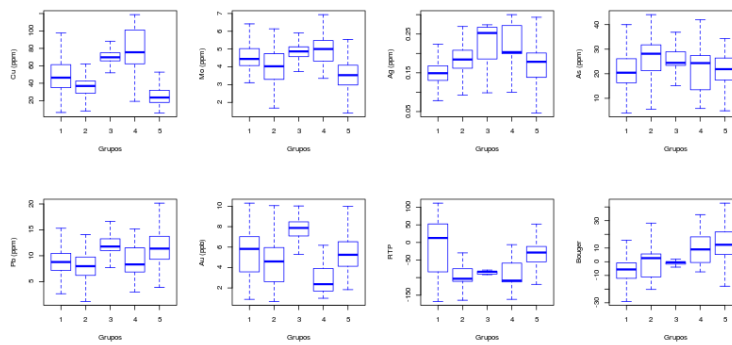
Figura A.15: Distribución espacial para los clústers formados por el método de la distancia máxima (Complete linkeage)



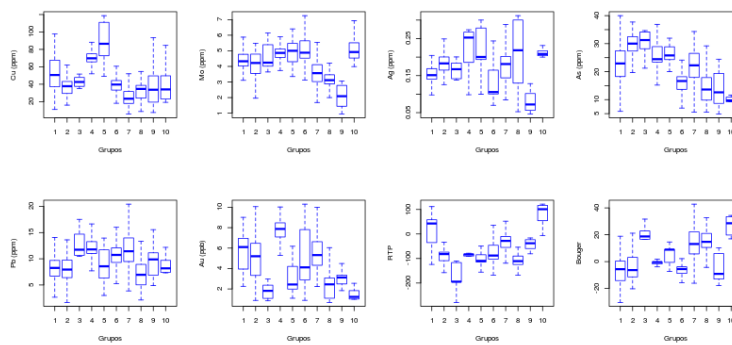
(a) Dist. de las variables para los 3 clústers



(b) Dist. de las variables para los 4 clústers



(c) Dist. de las variables para los 5 clústers



(d) Dist. de las variables para los 10 clústers

Figura A.16: Distribución de las variables de estudio en función de los clústers formados por el método de la distancia máxima (Complete linkage)

A.9. Resultados del caso de estudio aplicando Conglomerados Jerárquicos con el método de la distancia mínima (Single linkage) con coordenadas espaciales

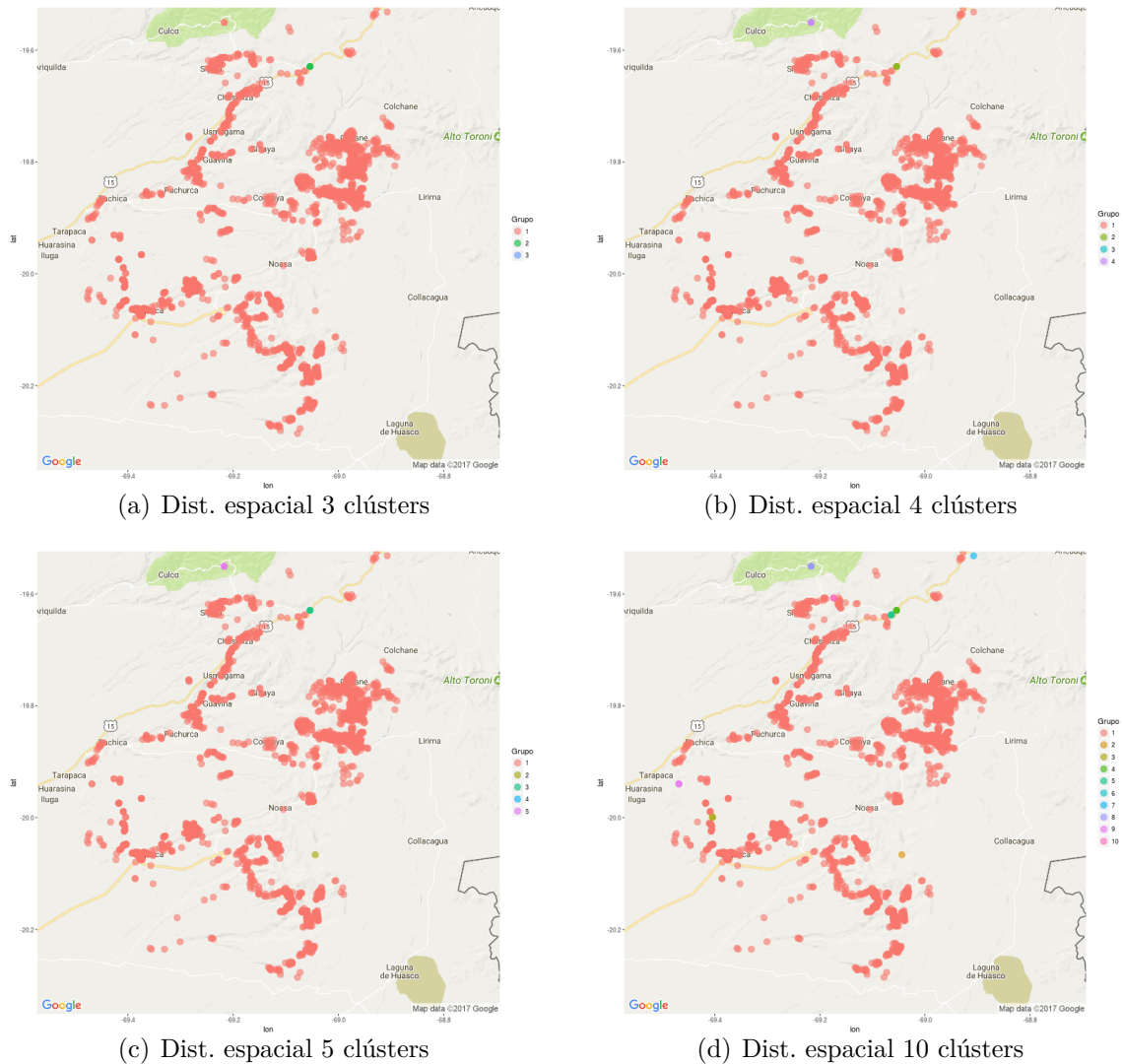
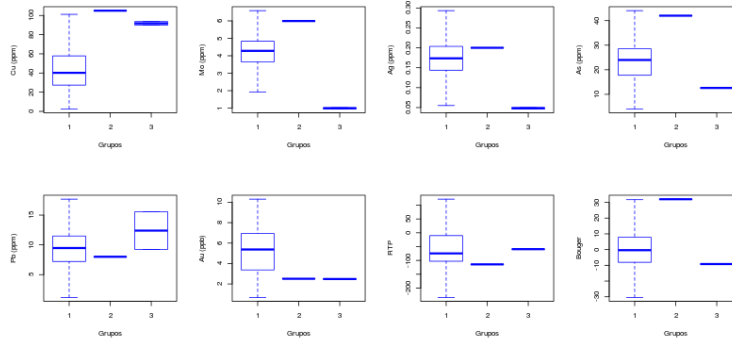
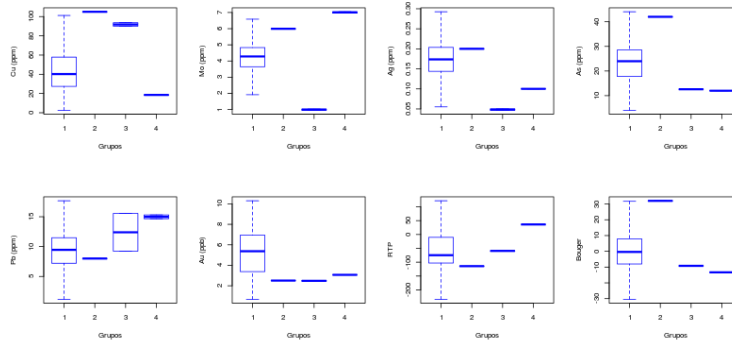


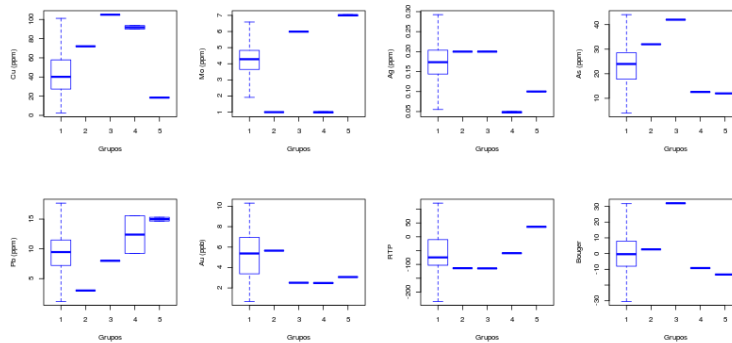
Figura A.17: Distribución espacial para los clústers formados por el método de la distancia mínima (Single linkage)



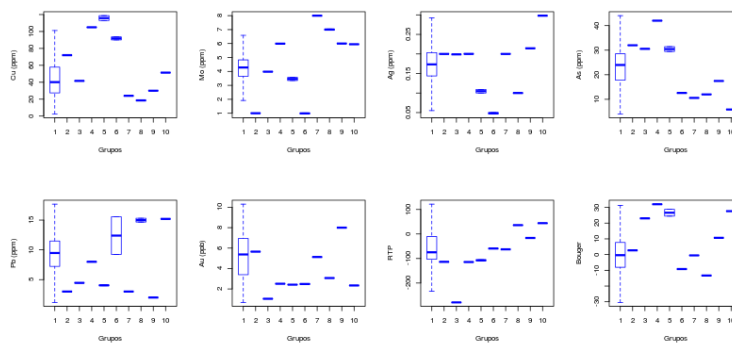
(a) Dist. de las variables para los 3 clústers



(b) Dist. de las variables para los 4 clústers



(c) Dist. de las variables para los 5 clústers

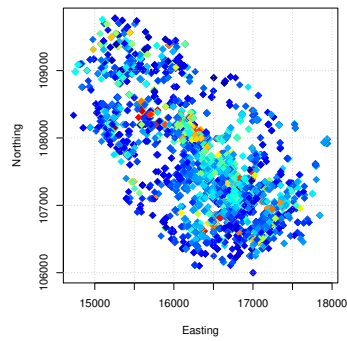


(d) Dist. de las variables para los 10 clústers

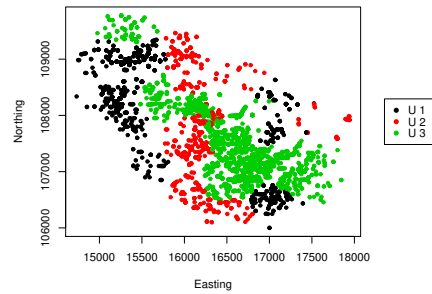
Figura A.18: Distribución de las variables de estudio en función de los clústers formados por el método de la distancia mínima (Single linkage)

Anexo B

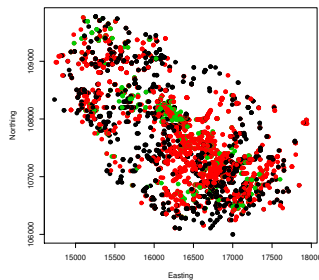
Caso de estudio yacimiento Escondida, II Región de Antofagasta



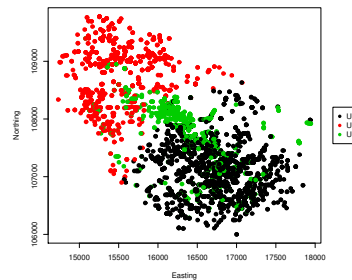
(a) Distribución en planta de la ley de cobre (Cu) total



(b) Distribución en planta de las tres mezclas para la ley de cobre (Cu) total con Mezclas de Distribuciones Geoes-tadísticas

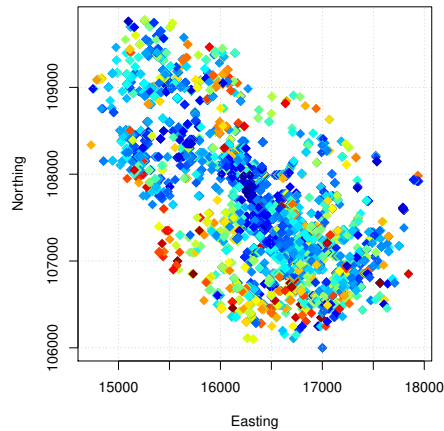


(c) Distribución en planta de las tres mezclas para la ley de cobre (Cu) total con Mezclas de Distri-buciones Gaussianas sin coorde-nadas espaciales

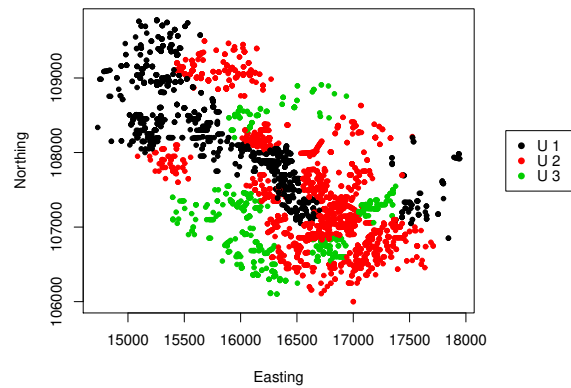


(d) Distribución en planta de las tres mezclas para la ley de cobre (Cu) total con Mezclas de Distri-buciones Gaussianas con coorde-nadas espaciales

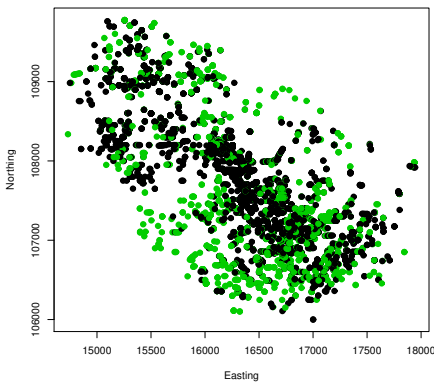
Figura B.1: Comparación entre resultados de las tres mezclas encontradas para la ley de cobre (Cu) total



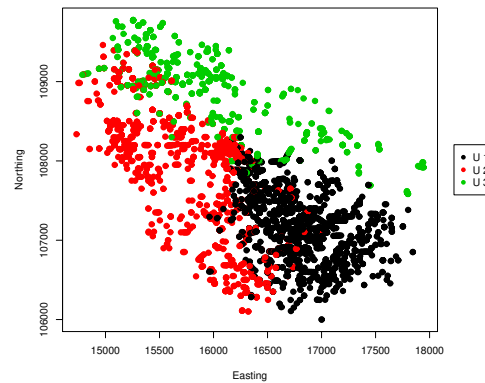
(a) Distribución en planta de la ley de fierro (Fe) total



(b) Distribución en planta de las tres mezclas para la ley de fierro (Fe) total con Mezclas de Distribuciones Geoestadísticas



(c) Distribución en planta de las tres mezclas para la ley de fierro (Fe) total con Mezclas de Distribuciones Gaussianas sin coordenadas espaciales

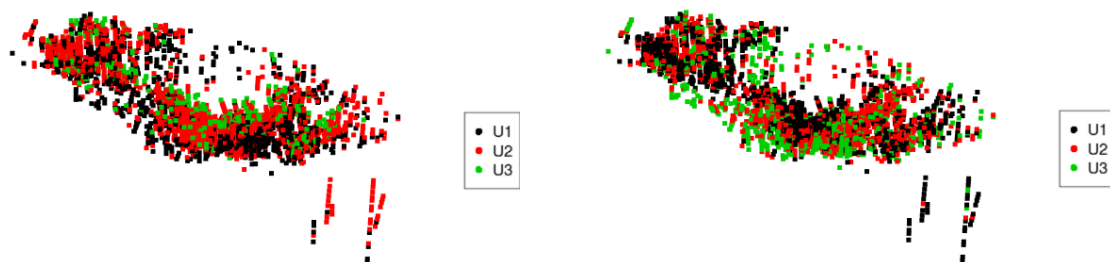


(d) Distribución en planta de las tres mezclas para la ley de fierro (Fe) total con Mezclas de Distribuciones Gaussianas con coordenadas espaciales

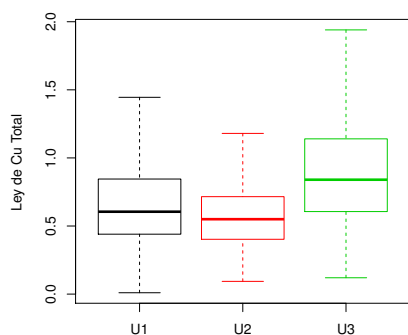
Figura B.2: Comparación entre resultados de las tres mezclas encontradas para la ley de fierro (Fe) total

Las Figuras B.1 y B.2 muestran una comparación entre resultados de las tres mezclas encontradas para la ley de cobre (Cu) total y para la ley de fierro (Fe) total respectivamente. Las Figuras B.3 y B.4 muestran los resultados del caso de estudio aplicando Mezclas de Distribuciones Geoestadísticas para la ley de cobre (Cu) total y ley de fierro (Fe) total para tres y cuatro unidades descubiertas respectivamente. Las Figuras B.5 y B.6 muestran los resultados del caso de estudio aplicando Mezclas de Distribuciones Gaussianas sin coordenadas espaciales para la ley de cobre (Cu) total y ley de fierro (Fe) total para tres y cuatro unidades descubiertas respectivamente. Las Figuras B.7 y B.8 muestran los resultados del caso de estudio aplicando Mezclas de Distribuciones Gaussianas con coordenadas espaciales para la ley de cobre (Cu) total y ley de fierro (Fe) total para tres y cuatro unidades descubiertas respectivamente.

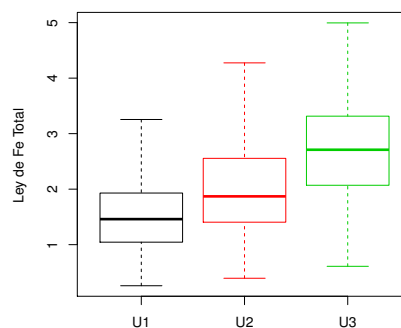
B.1. Resultados del caso de estudio aplicando Mezclas de Distribuciones Geoestadísticas para la ley de cobre (Cu) total y ley de fierro (Fe) total



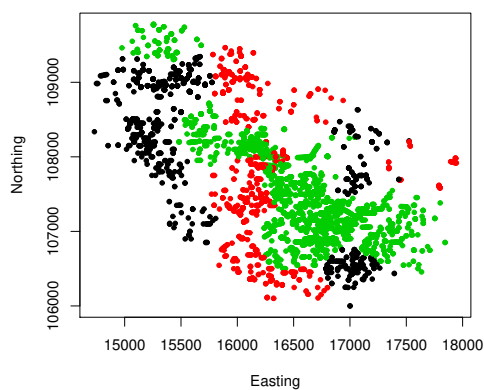
(a) Distribución espacial de las tres mezclas para la ley de cobre (Cu) total (b) Distribución espacial de las tres mezclas para la ley de fierro (Fe) total



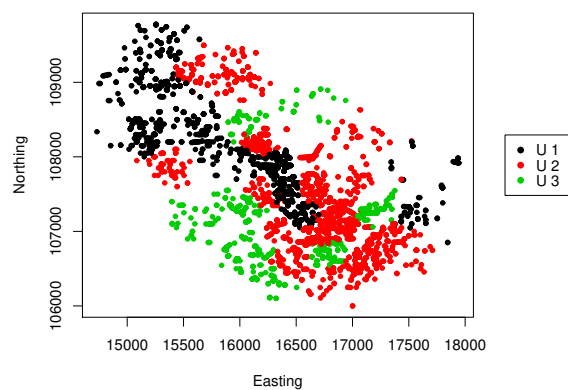
(c) Dist. de la ley de cobre (Cu) total para las tres mezclas



(d) Dist. de la ley de fierro (Fe) total para las tres mezclas

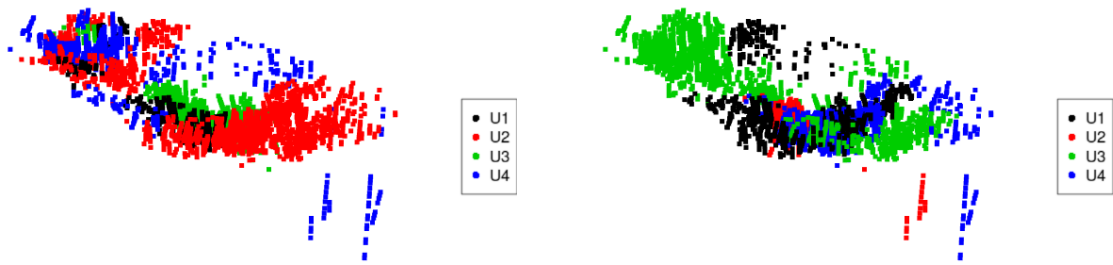


(e) Distribución en planta de las tres mezclas para la ley de cobre (Cu) total



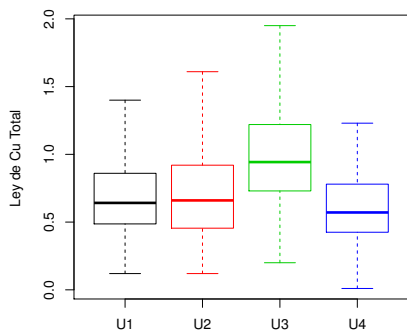
(f) Distribución en planta de las tres mezclas para la ley de fierro (Fe) total

Figura B.3: Distribución de las variables de estudio en función de las tres mezclas encontradas

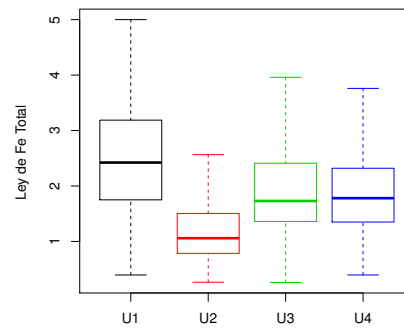


(a) Distribución espacial de las cuatro mezclas para la ley de cobre (Cu) total

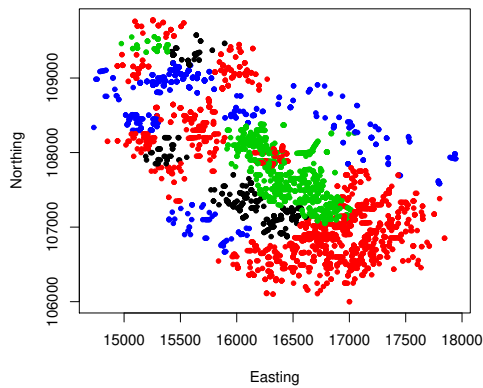
(b) Distribución espacial de las cuatro mezclas para la ley de hierro (Fe) total



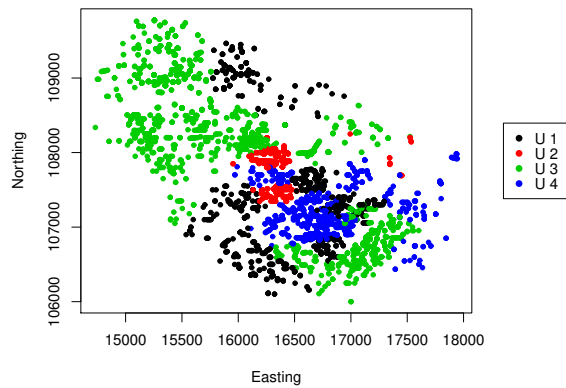
(c) Dist. de la ley de cobre (Cu) total para las cuatro mezclas



(d) Dist. de la ley de hierro (Fe) total para las cuatro mezclas



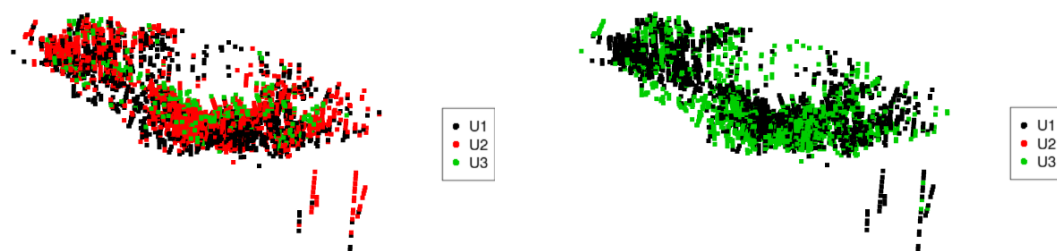
(e) Distribución en planta de las cuatro mezclas para la ley de cobre (Cu) total



(f) Distribución en planta de las cuatro mezclas para la ley de hierro (Fe) total

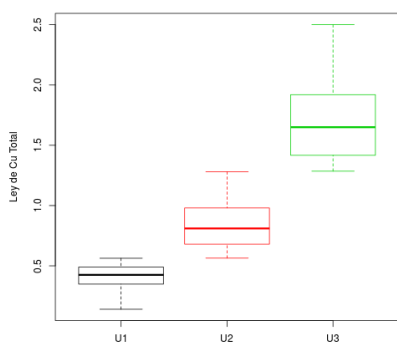
Figura B.4: Distribución de las variables de estudio en función de las cuatro mezclas encontradas

B.2. Resultados del caso de estudio aplicando Mezclas de Distribuciones Gaussianas sin coordenadas espaciales para la ley de cobre (Cu) total y ley de hierro (Fe) total

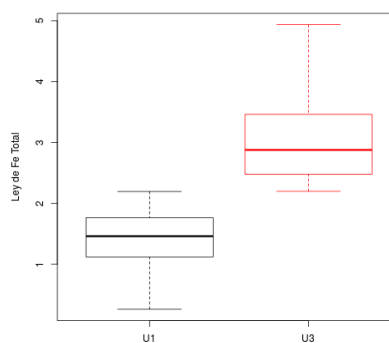


(a) Distribución espacial de las tres mezclas para la ley de cobre (Cu) total

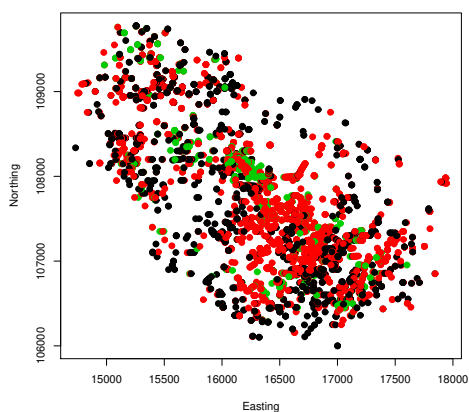
(b) Distribución espacial de las tres mezclas para la ley de hierro (Fe) total



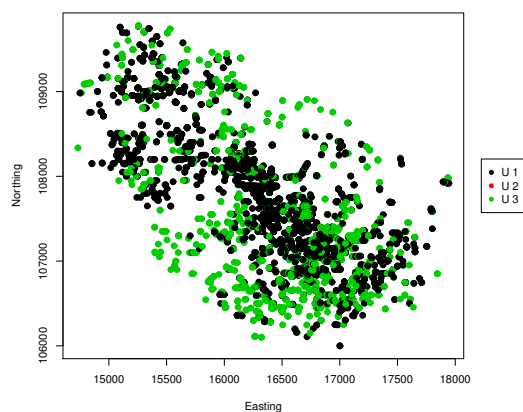
(c) Dist. de la ley de cobre (Cu) total para las tres mezclas



(d) Dist. de la ley de hierro (Fe) total para las tres mezclas

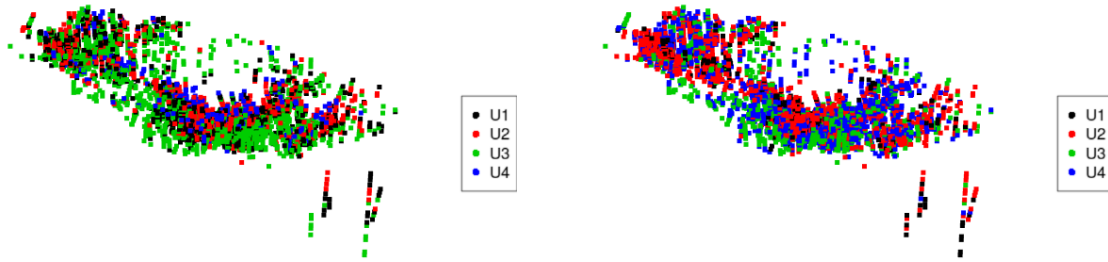


(e) Distribución en planta de las tres mezclas para la ley de cobre (Cu) total



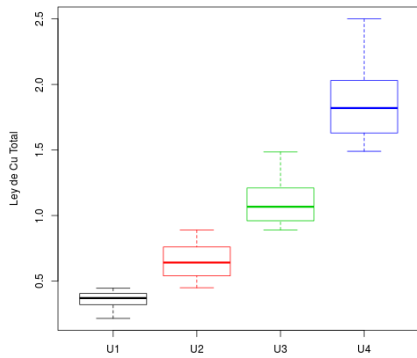
(f) Distribución en planta de las tres mezclas para la ley de hierro (Fe) total

Figura B.5: Distribución de las variables de estudio en función de las tres mezclas encontradas

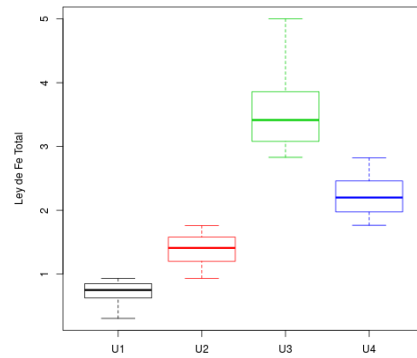


(a) Distribución espacial de las cuatro mezclas para la ley de cobre (Cu) total

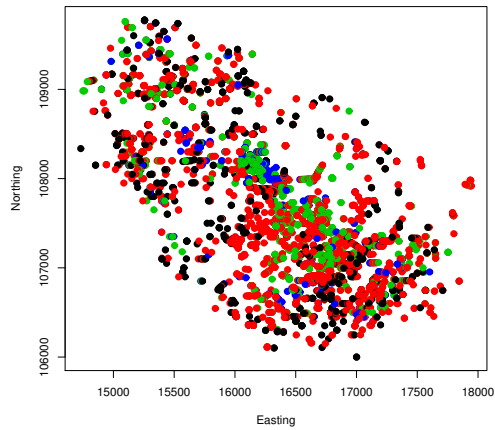
(b) Distribución espacial de las cuatro mezclas para la ley de hierro (Fe) total



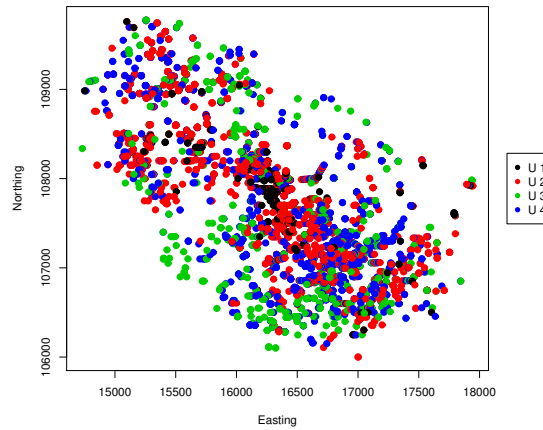
(c) Dist. de la ley de cobre (Cu) total para las cuatro mezclas



(d) Dist. de la ley de hierro (Fe) total para las cuatro mezclas



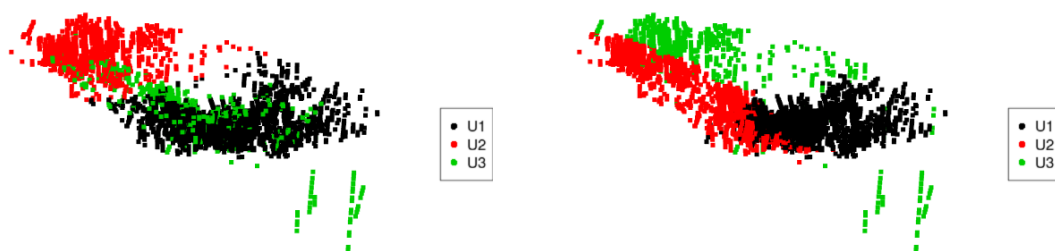
(e) Distribución en planta de las cuatro mezclas para la ley de cobre (Cu) total



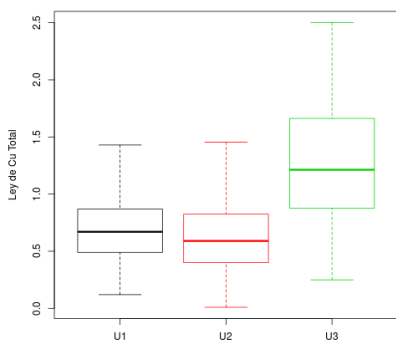
(f) Distribución en planta de las cuatro mezclas para la ley de hierro (Fe) total

Figura B.6: Distribución de las variables de estudio en función de las cuatro mezclas encontradas

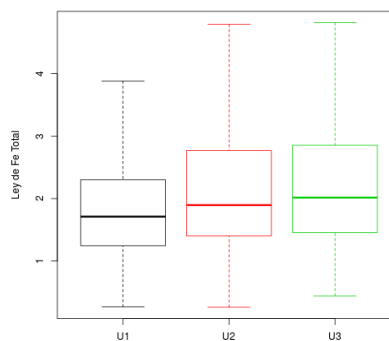
B.3. Resultados del caso de estudio aplicando Mezclas de Distribuciones Gaussianas con coordenadas espaciales para la ley de cobre (Cu) total y ley de fierro (Fe) total



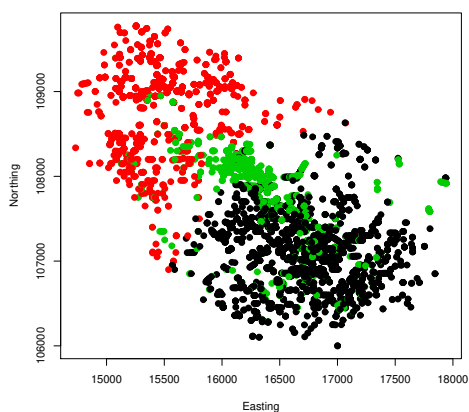
(a) Distribución espacial de las tres mezclas para la ley de cobre (Cu) total (b) Distribución espacial de las tres mezclas para la ley de fierro (Fe) total



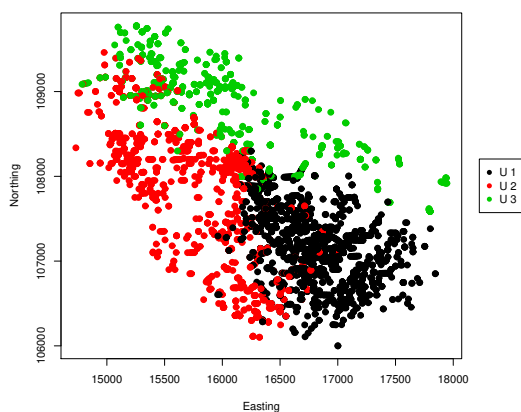
(c) Dist. de la ley de cobre (Cu) total para las tres mezclas



(d) Dist. de la ley de fierro (Fe) total para las tres mezclas

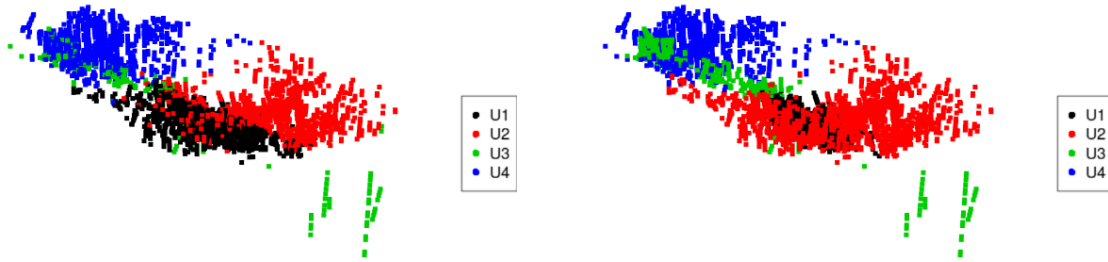


(e) Distribución en planta de las tres mezclas para la ley de cobre (Cu) total



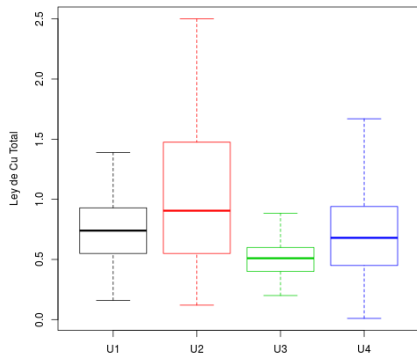
(f) Distribución en planta de las tres mezclas para la ley de fierro (Fe) total

Figura B.7: Distribución de las variables de estudio en función de las tres mezclas encontradas

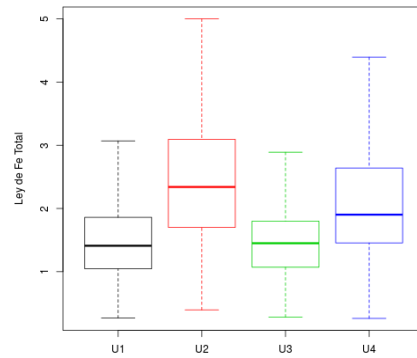


(a) Distribución espacial de las cuatro mezclas para la ley de cobre (Cu) total

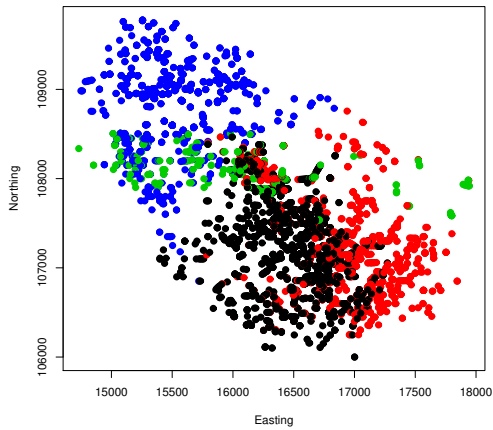
(b) Distribución espacial de las cuatro mezclas para la ley de hierro (Fe) total



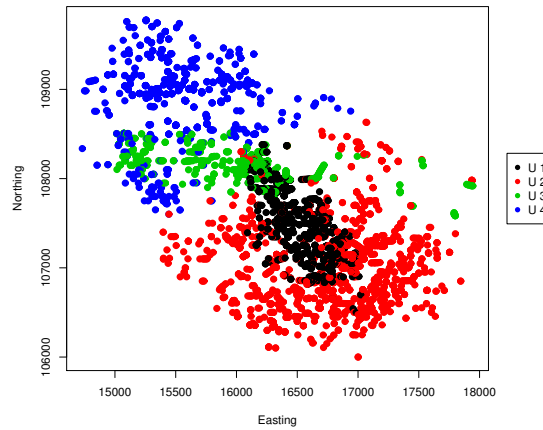
(c) Dist. de la ley de cobre (Cu) total para las cuatro mezclas



(d) Dist. de la ley de hierro (Fe) total para las cuatro mezclas



(e) Distribución en planta de las cuatro mezclas para la ley de cobre (Cu) total



(f) Distribución en planta de las cuatro mezclas para la ley de hierro (Fe) total

Figura B.8: Distribución de las variables de estudio en función de las cuatro mezclas encontradas