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

MODELAMIENTO GEOESTADÍSTICO PARA EL MAPEO GEOLÓGICO PREDICTIVO A PARTIR DE INFORMACIÓN GEOQUÍMICA

TESIS PARA OPTAR AL GRADO DE DOCTOR EN INGENIERÍA DE MINAS

JOSÉ ARTURO GUARTÁN MEDINA

PROFESOR GUÍA:
XAVIER EMERY

MIEMBROS DE LA COMISIÓN:
BRIAN TOWNLEY CALLEJAS
JOSÉ MUNIZAGA ROSAS
EXEQUIEL SEPÚLVEDA ESCOBEDO

SANTIAGO DE CHILE
2021

RESUMEN DE LA MEMORIA PARA OPTAR
AL TITULO DE: Doctor en Ingenieria de Minas
POR: José Arturo Guartán Medina
FECHA: enero 2021
PROFESOR GUÍA: Xavier Emery

El modelamiento geológico predictivo es un elemento esencial en etapas de exploración de los recursos minerales, para lo cual se suele clasificar información geoquímica georreferenciada obtenida de una toma de muestras para poder predecir categorías geológicas, por ejemplo, tipos de roca, alteración y/o mineral, utilizando métodos estadísticos multivariables o de aprendizaje automático. Sin embargo, dichos métodos basan sus predicciones en la relación de dependencia colocalizada entre variables geoquímicas y geológicas, ignorando parte de la información espacial contenida en los datos de muestreo. Estos problemas de clasificación regionalizada son abordados en este trabajo desde un punto de vista geoestadístico, de modo de aprovechar la correlación espacial de los datos, así como las correlaciones entre las variables cuantitativas (concentraciones geoquímicas) y las categorías geológicas. Para ello, se proponen dos enfoques. El primero consiste en aplicar un clasificador, ajustado sobre datos de entrenamiento, a los valores de las variables cuantitativas interpolados en el espacio, introduciendo dos novedades: la primera es el uso de simulaciones geoestadísticas para la interpolación, lo cual permite generar numerosos escenarios que emulan la variabilidad espacial real y obtener tantas clasificaciones como escenarios hayan; la segunda novedad radica en el filtrado del componente de pequeña escala (efecto pepita), asociado a ruidos y errores de medición, al momento de construir los escenarios simulados. El segundo enfoque consiste en simular directamente las categorías geológicas en base a un modelo plurigaussiano, en el cual las concentraciones geoquímicas pueden ser incorporadas como covariables.

Los enfoques propuestos son aplicados a dos casos de estudio, consistentes en un depósito de tipo pórfido cuprífero reconocido por sondajes de exploración y en una zona de prospección minera reconocida por muestras de superficie. En ambos casos, se considera como predicción final la clasificación más probable en cada sitio del espacio (la que más se repite en los diferentes escenarios simulados). La calidad de las predicciones se evalúa sobre un conjunto de datos de prueba diferente del conjunto de datos de entrenamiento, con un desempeño significativamente mayor con respecto a métodos más tradicionales que no filtran la variabilidad de pequeña escala o que no consideran escenarios simulados para la clasificación. Los mapas predictivos y mapas de probabilidad de ocurrencia de categorías geológicas permiten interpretar los procesos geológicos y geoquímicos y ayudan a vectorizar la prospección mineral en las etapas iniciales de la exploración del recurso mineral.

ABSTRACT OF MEMORY TO QUALIFY FOR
THE TITLE OF: Doctor en Ingenieria de Minas
BY: Jose Arturo Guartan Medina
DATE: enero 2021
GUIDE PROFFESOR: Xavier Emery

Predictive geological mapping is an essential element in exploration stages of mineral resources, for which georeferenced geochemical information obtained from a set of sampling data is usually classified in order to predict geological categories, for example, rock types, alterations and/or mineral types, using multivariate statistical or machine learning approaches. However, these approaches base their predictions on the colocated dependence relationship between geochemical and geological variables, ignoring part of the spatial information conveyed by the sampling data.

These regionalized classification problems are addressed in this work from a geostatistical point of view, in order to take advantage of the spatial correlation of the data and the correlations between quantitative variables (geochemical concentrations) and geological categories. To this end, two approaches are proposed. The first approach consists in applying a classifier, fitted on training data, to the values of the quantitative variables interpolated in space, introducing two novel features: the first is the use of geostatistical simulations for interpolation, which allows generating numerous scenarios that mimic the real spatial variability and obtaining as many classifications as there are scenarios; the second novelty lies in the filtering of the small-scale component (nugget effect), associated with noise and measurement errors, when constructing the simulated scenarios. The second approach consists in directly simulating the geological categories based on a plurigaussian model, in which the geochemical concentrations can be incorporated as covariates.

The proposed approaches are applied to two case studies, consisting of a porphyry-type copper deposit recognized by exploration drilling and in a mining prospect zone recognized by surface sampling. In both cases, the most probable classification at each location in space (the one that is most repeated among the different simulated scenarios) is considered as the final prediction. The quality of the predictions is evaluated on a testing data set different from the training data set, showing a significantly higher performance compared to more traditional methods that do not filter the small-scale variability or that do not consider simulated scenarios for classification.

The predictions and maps of the probability of occurrence of geological categories are helpful to interpret geological and geochemical processes and to vector ore in mineral prospection and in the initial stages of mineral resource exploration.

Agradecimientos

Primero quiero agradecer a mis Dios Todopoderoso por haberme dado la oportunidad de seguir con salud, vida y culminar este objetivo trazado en mi vida profesional y académica.

Quiero expresar un agradecimiento muy especial al profesor Xavier Emery, asesor de este trabajo, con su apoyo incondicional, supe salir adelante en las diferentes etapas de mi estudio doctoral e investigación; gracias profe por su paciencia, y su alto conocimiento que me guió en todo el tiempo de la investigación y la escritura de esta tesis.

A los miembros de la comisión de calificación y evaluación, profesores Brian Townley, José Munizaga y Exequiel Sepúlveda, por sus comentarios muy acertados sobre el borrador de esta tesis. A mi querida y linda familia, a mi amada esposa Luz María; mis hijos Bryan, Steven, Fernanda y Sofhia Camila, quienes fueron el motor y motivo de salir adelante; gracias por todo ese apoyo de estar siempre pendiente de mí cuando estaba lejos de mi país.

Quiero agradecer al Departamento de Ingenieria de Minas y al Centro Avanzado de Tecnología para la Mineria (AMTC) de la Universidad de Chile, por el apoyo en el desarrollo de mi estudio doctoral.

Mi agradecimiento a la Secrectaria Nacional del Ecuador de Ciencia y Tecnología (SENESCYT) por el financiamiento otorgado para que pueda desarrollar mis estudios. A la institución donde trabajo en mis actividades de docencia e investigación, la Universidad Técnica Particular de Loja (UTPL) de Ecuador por el auspicio dado y apoyo para que pueda culminar con éxito esta etapa de mi vida académica. A la Compañía Minera Cornerstone de Ecuador, en especial a Osman Poma, por las facilidades que brindó de otorgar una de las bases de datos que fue parte del desarrollo de la investigación de esta tesis.

Finalmente, agradecer a todas las personas, compañeros de trabajo, compañeros de estudio y personal administrativo de la Universidad de Chile que me guiaron en todos los tramites desde cuando llegué a Santiago de Chile, gracias por todo.

Tabla de Contenido

Capítulo 1: Introducción

1. Motivación y planteamiento del problema

Un mapa geológico es el resultado de la caracterización de tipos de rocas, minerales o alteraciones. En prospección geológica minera, los diferentes tipos de litología se pueden estudiar en función de su mineralogía o su geoquímica, que analiza la composición de los sedimentos superficiales o del suelo (Jenny, 1941). La investigación de la química de la superficie a escala regional a menudo se dirige a determinar el fondo geoquímico o la línea de base de elementos simples (Stanley y Sinclair, 1989). Las concentraciones geoquímicas superficiales varían en el espacio de un lugar a otro con una continuidad más o menos pronunciada y pueden usarse para describir condiciones naturales locales como la geología o las actividades antropogénicas. Por lo tanto, la geoquímica de elementos múltiples proporciona información útil sobre la litología como parámetro esencial para la geología de áreas prospectivas (Grunsky et al. 2012).

Los mapas geológicos también son un elemento clave en exploración minera temprana o avanzada, ya sea para identificar blancos de exploración y planificar campañas de muestreo, o para delimitar la mineralización económica y evaluar los recursos minerales. Por ejemplo, la presencia de rocas volcánicas intermedias o félsicas que son alteradas por la presencia de un cuerpo ígneo intrusivo proporciona información relevante para la exploración de depósitos de tipo pórfido y de vetas hidrotermales (Sillitoe, 2003).

La aplicación de herramientas estadísticas multivariables y de aprendizaje automático (*machine learning*) proporciona un marco sistemático a través del cual se identifican procesos geoquímicos y geológicos, que serían, entre otros, el uso de técnicas de clasificación para el modelamiento de exploración mineral y mapeo litológico (Carranza, 2009; Grunsky, 2010; Grunsky et al., 2014; Rodriguez-Galiano et al., 2015; McKay and Harris, 2016; Zuo and Xiong, 2018; Kuhn et al., 2019; Liu et al., 2019; Sun et al., 2019; Xiang et al., 2020), por ejemplo, análisis discriminante lineal, redes neuronales, máquinas de vectores de soporte, árboles de decisión o bosques aleatorios (Fisher, 1936; Quinlan, 1993; Breiman, 2001; Hastie et al., 2008). Estos métodos funcionan en las variables cuantitativas y categóricas observadas en las ubicaciones de muestreo, pero generalmente no explican explícitamente la correlación espacial conjunta de las observaciones al clasificar, por lo tanto, es probable que pierdan parte de la información transmitida por los datos de muestreo.

Ahora bien, las variables regionalizadas como las concentraciones geoquímicas o los tipos de roca a menudo exhiben dependencias espaciales que pueden modelarse interpretando estas variables como realizaciones de funciones aleatorias hipotéticas, interpretación que sustenta el desarrollo de las técnicas geoestadísticas (Matheron, 1962; Chilès and Delfiner, 2012). En esta configuración, las dependencias espaciales entre datos regionalizados se caracterizan, entre otras herramientas, por funciones de covarianza o por variogramas que miden la similitud (para las covarianzas) o la disimilitud (para los variogramas) promedio existente entre dos datos en función de su separación espacial y que se pueden inferir a partir de la información de muestreo. Un enfoque de modelamiento común es representar las covarianzas o variogramas como la suma de diferentes componentes, llamados estructuras anidadas, que actúan a diferentes escalas espaciales y se asocian con diferentes alcances de correlación (Wackernagel, 1988, 2003). Los componentes de muy corto alcance (en particular, el efecto pepita) a menudo representan ruido o errores de medición, que

pueden perjudicar la clasificación de variables cuantitativas, como las concentraciones geoquímicas, para predecir categorías geológicas, como la litología.

Aprovechar la correlación espacial que exhibe la información geoquímica o geológica, así como sus diferentes escalas de variación, hace cada vez más imperativo el uso de herramientas y técnicas geoestadísticas para construir modelos predictivos, evaluar incertidumbre y hacer clasificaciones regionalizadas sobre la base de datos de muestreo (Olea, 1999; Gringarten and Deutsch, 2001; Darsow et al., 2009; Hassan et al., 2009; Adeli et al., 2018)

En este contexto, esta tesis aborda el problema de la clasificación que surge en la prospección y la exploración geológica-minera, desde una perspectiva geoestadística. La investigación propuesta consiste en predecir una variable categórica (litologías, rocas, alteraciones, zonas minerales) en el espacio, en base a un conjunto de muestras con información de la misma variable categórica y de concentraciones de elementos químicos (variables cuantitativas). Este problema se conoce en la literatura como *mapeo geológico predictivo* o como *clasificación regionalizada*, siendo tanto el aspecto regionalizado como el aspecto multivariable relevantes para poder aprovechar la información contenida en los datos de muestreo. Una de las principales novedades de la tesis radicará en el diseño de métodos aplicables a bases de datos altamente multivariables, luego capaces de incorporar varias decenas de concentraciones geoquímicas como covariables en la predicción de una variable categórica.

2. Estado del arte

Se han documentado numerosos ejemplos y aplicaciones de mapeo geológico predictivo (Olea, 1999; Barbosa et al., 2010; Grunsky, 2010; Grunsky et al., 2012; Grunsky et al., 2014; Adeli et al., 2018; Talebi et. al., 2019). Varias de las propuestas desarrolladas para abordar este problema se basan en la combinación de técnicas estadísticas y geoestadísticas. Así, Harff and Davis (1990), Bohling (1997) proponen utilizar el análisis discriminante, luego interpolar por kriging las métricas de distancia o las probabilidades de pertenencia a cada clase, desde la información de muestreo hacia una grilla que cubre la región de interés. Otros autores abundan en este sentido, tales como Olea (1999), Pacheco and Barbosa (2005), Leal-Pacheco and Barbosa-Landim (2005) y Barbosa et al. (2010), quienes combinan técnicas estadísticas tales como el análisis de componentes principales, análisis de correspondencia, análisis de conglomerados y análisis discriminante, con técnicas de análisis espacial para identificar las clases relevantes y para predecir estas clases en una región de interés. Los trabajos anteriores muestran diversas aplicaciones con datos de geoquímica de superficie o de subsuelo. Siguiendo esta línea, Grunsky et al. (2012, 2014) mapean la probabilidad de encontrar diferentes tipos de litología en base a la información geoquímica de muestras de sedimentos lacustres de la península Melville en Canadá, combinando técnicas estadísticas y geoestadísticas (análisis de componentes principales, análisis de varianza, análisis discriminante, análisis variográfico y kriging ordinario). Mueller and Grunsky (2016) proponen el uso de MAF (minimum/maximum autocorrelation factors) en lugar del análisis de componentes principales.

Por su parte, Adeli et al. (2018) y Talebi et al. (2018) siguen un orden inverso, puesto que primero aplican técnicas geoestadísticas para simular las variables cuantitativas (geoquímica) en el espacio, luego técnicas de clasificación (tales como árboles de decisión o bosques aleatorios) para obtener simulaciones categóricas, a partir de las cuales se puede determinar la probabilidad de ocurrencia de cada categoría en cualquier sitio de interés.

Los enfoques anteriormente mencionados hacen un uso combinado de técnicas geoestadísticas (por ejemplo, kriging) para interpolar las variables cuantitativas en el espacio, y de técnicas estadísticas o de aprendizaje automático para convertir la información cuantitativa en categórica. Ahora bien, estas últimas técnicas no aprovechan plenamente la información espacial para lograr la clasificación, puesto que la continuidad espacial de la variable categórica no se modela directamente, sino que a través de la continuidad espacial de las variables cuantitativas y de la dependencia estadística entre estas últimas variables y la variable categórica. En este sentido, un elemento novedoso de la propuesta de tesis consistirá en simular directamente las categorías en el espacio, condicionalmente a los datos de las variables cuantitativas y de la misma variable categórica conocidos en sitios de muestreo, siendo el principal desafío la alta dimensionalidad del problema (reflejada tanto por el número de categorías a simular, como por el número de variables cuantitativas auxiliares que incorporar).

3. Objetivos y contribución de la tesis

La propuesta pretende desarrollar herramientas y modelos geoestadísticos para la clasificación regionalizada, basada en la cosimulación de variables cuantitativas (concentraciones geoquímicas) y categóricas (geología) a partir de datos de muestreo. El análisis de las realizaciones permitirá, por ejemplo, disponer de un modelo realista, con la menor incertidumbre posible, que indique la pertenencia de minerales a sus litologías respectivas.

En base a la información recopilada, a la fecha (2020) existe poca información que especifique la simulación de numerosos elementos medidos en escalas continuas y litologías medidas en escalas categóricas. En particular, las investigaciones realizadas sobre simulación conjunta multigaussiana (para el modelamiento de variables cuantitativas) y plurigaussiana (para variables categóricas) (Freulon et al., 1990; Dowd, 1997; Bahar and Kelkar, 2000; Emery, 2007; Emery and Silva, 2009; Cáceres and Emery, 2010; Maleki and Emery, 2015, 2017; Emery and Maleki, 2019) se han enfocado a modelos de yacimientos tridimensionales con un número reducido de variables y de categorías, donde la base de datos es resultado de la exploración en base a sondajes de perforación. Esta tesis se centrará entonces en el problema de la cosimulación condicional de un gran conjunto de variables regionalizadas cuantitativas (concentraciones geoquímicas) y categóricas (tipos de roca o litología), extendiendo las herramientas, los modelos y los algoritmos de simulación geoestadística a casos altamente multivariables, en donde además todas las variables podrían no estar conocidas en los mismos sitios de muestreo (caso heterotópico).

Las principales contribuciones con respecto a la literatura existente son: (1) el uso de información espacial altamente multivariable (datos correspondientes a varias decenas de elementos geoquímicos) para resolver problemas de clasificación regionalizada y modelamiento geológico predictivo, aplicable a sitios del espacio que no han sido muestreados, (2) el diseño de métodos mejorados de clasificación en presencia de ruido o errores de medición en las variables cuantitativas, combinando metodologías de simulación geoestadística multivariable con filtrado de componentes espaciales de corto alcance (efecto pepita) y técnicas de aprendizaje automático, y (3) el diseño de modelos y algoritmos de simulación conjunta de variables cuantitativas y categóricas, basados en combinaciones de los modelos multigaussianos y plurigaussianos, de modo que todas las variables (cuantitativas y categórica) queden representadas por funciones aleatorias Gaussianas cuya estructura de correlación espacial se puede inferir a partir de los datos de las variables cuantitativas e indicadores de la variable categórica.

4. Hipótesis

La investigación propuesta se basa en las siguientes hipótesis:

- Las variables geoquímicas son variables corregionalizadas y se relacionan con las estructuras y composición de diversos tipos de litologías que dependen de la génesis de su formación y su estrecha relación con los minerales. Un análisis de la geoquímica permite identificar a priori la roca parental.
- El muestreo es fragmentario, en el cual el investigador realiza observaciones directas y asume que son representativas de la zona a estudiar, es decir, el muestreo no es preferencial.
- La geoestadística proporciona herramientas para modelar la estructura de correlación espacial de datos corregionalizados, e identificar diferentes escalas de variación en dicha estructura espacial.
- La simulación geoestadística permite construir escenarios que reproducen la variabilidad espacial y las dependencias de variables corregionalizadas (cuantitativas y/o categóricas) y evaluar la incertidumbre en los valores reales en lugares no muestreados. La simulación se basa en la construcción de realizaciones de una o varias funciones aleatorias de interés para determinar la interrelación geoquímica-litología (mineral-roca).

5. Organización de la tesis

Los capítulos siguientes de esta tesis están organizados de la siguiente manera. El capítulo 2 presenta los principales antecedentes y conceptos teóricos que serán utilizados en las diferentes propuestas, así como una revisión bibliográfica del estado del arte. El capítulo 3 plantea las dos propuestas metodológicas desarrolladas en esta investigación. En el capítulo 4 se muestra el análisis y discusión relacionada a la primera propuesta para un estudio de caso de un depósito de pórfido de cobre. El capítulo 5 plantea un estudio de caso para datos de geoquímica de superficie en el cual se desarrolla la segunda propuesta. Finalmente, en el capítulo 6 presenta las discusiones generales, conclusiones y perspectivas que deja la investigación realizada.

Capítulo 2: Antecedentes

1. Litología y geoquímica

La corteza terrestre continental está integrada mayoritariamente por 11 elementos químicos y el resto participa cada uno en cantidades menores al 0.1% en peso del total. El análisis químico rutinario de rocas determina primero la participación de los denominados elementos mayoritarios que se expresan en peso de los óxidos de elementos. En base a compilaciones de análisis de rocas publicados y estimaciones de las proporciones relativas de las rocas representadas en la corteza continental, se establecen los datos expresados en la Tabla 2.1. de los promedios de los elementos mayores (Poldevaart, 1955; Ronov and Yaroshevsky, 1976).

Tabla 2.1. Elementos mayoritarios y sus óxidos.

Elemento	% en peso	Óxidos
Oxigeno	46.40	---
Silicio	28.15	SiO_2; sílice
Aluminio	8.23	Al_2O_3; alúmina
Hierro	5.63	Fe_2O_3; óxido férrico
Calcio	4.15	CaO; cal
Sodio	2.36	Na_2O; soda
Magnesio	2.33	MgO; oxido de magnesio
Potasio	2.09	K_2O; potasa
Titanio	0.57	Ti_2O; óxido de titanio
Fosforo	0.105	P_2O_5; pentóxido de fosforo
Manganeso	0.095	MnO; óxido de manganeso

Las rocas se componen en su mayoría de minerales de tipo silicatos, los mismos que son la combinación de los elementos químicos mayoritarios. Las rocas y sus diversos tipos de litologías pueden ser estudiados a partir de la mineralogía o la prospección geoquímica que analiza la composición del sedimento superficial o del suelo (s), la cual varía espacialmente como resultado de los cambios del clima (cl), intervención de organismos (o), la forma de relieve (r), el material de origen o parental (p) y el tiempo de formación de los suelos (t). Jenny (1941) sostiene que la formación de los suelos lo expresa como: $s = f(cl, o, r, p, t)$. La predicción cuantitativa de la relación entre los elementos químicos y la litología parental es posible por el desarrollo de grandes bases de datos digitales, por avances de capacidades analíticas y aplicaciones de métodos de análisis estadísticos multivariables. La documentación o mapeo de suelos y rocas con una buena capacidad predictiva se beneficia en el descubrimiento de conocimientos en pedología (McBratney et al., 2003). Grunsky et al. (2012, 2014) concluyeron que los análisis estadísticos multivariables por análisis de componentes principales de los datos geoquímicos que describen combinaciones lineales de elementos, son controlados mediante la estequiometría de los minerales y los procesos geoquímicos/geológicos asociados; esto provee una base objetiva para la validación y para mejorar potencialmente la existencia de mapas geológicos.

Es importante conocer la litología del área o superficie de la tierra, puesto que el desconocimiento del tipo de roca puede traer complicaciones en cuanto a la errada asignación de algún uso antrópico que se le dé a una zona en estudio. En base a investigación se puede obtener conocimientos sobre el material parental, litologías, estructuras que llevarán a realzar una buena cartografía geológica (US Geological Survey, 2006), convirtiéndose en una herramienta importante para la exploración en la búsqueda de depósitos minerales.

En la fase de campo, se desarrollan campañas de muestreo ya sea superficial o subterráneamente, con la finalidad de determinar el tipo de formaciones geológicas en base al análisis de su litología. Tener un conjunto grande de datos puede ser un desafío para reconocer el valor y el potencial que tienen los datos para ofrecer información sobre los procesos geológicos (Grunsky, 2010). Investigaciones de la geoquímica de muestras de suelo en Sumatra, Indonesia (Grunsky and Smee, 1999) y de sedimentos lacustres de la Península de Melville, Canadá (Grunsky et al., 2012), indican que la geoquímica tiene un potencial de proporcionar información útil sobre la geología de las zonas muestreadas, por lo cual los valores de los datos geoquímicos de múltiples elementos son una ayuda importante para la exploración geología-minera.

La mayoría de la mineralización de metales preciosos y de elementos cerca de la superficie probablemente ya se han descubierto. La exploración continua de recursos metálicos requiere extender la búsqueda de mineralización no descubierta hacia la tercera dimensión, especialmente en áreas de cobertura gruesa (Cameron et al., 2004). La utilidad de la geoquímica para explorar la mineralización profundamente enterrada se ve obstaculizada por la falta de comprensión de los procesos por los cuales las especies asociadas a los minerales migran a través de una sobrecarga en escalas de tiempo significativas. Se han postulado muchos mecanismos para explicar las anomalías geoquímicas de la superficie, que incluyen difusión ascendente, advección, células electroquímicas, transporte de vapor, bombeo barométrico y bombeo sísmico (Hamilton, 1998; Cameron et al., 2004; Cohen et al., 2010).

Los grandes conjuntos de datos geoquímicos de elementos múltiples pueden ser interpretados con mayor efectividad cuando se aplican procedimientos multivariables de reducción de dimensión. La aplicación de métodos estadísticos a menudo revela patrones y relaciones dentro de los datos atribuidos a procesos geológicos/geoquímicos. Grunsky (2010) describe un enfoque sistemático para evaluar los datos geoquímicos que implica el examen de los datos geoquímicos como elementos individuales y asociaciones multivariables. Técnicas como el análisis de componentes principales, el análisis discriminante y otros procedimientos de clasificación proporcionan un marco sistemático mediante el cual se identifican los procesos geoquímicos/geológicos. Grunsky (2014) establece un enfoque objetivo para descubrir y clasificar los procesos geoquímicos a partir de los cuales se pueden probar y validar los mapas geológicos existentes y se pueden hacer nuevos mapas geológicos en áreas donde no existe suficiente información geológica.

Se puede cuantificar los elementos químicos que tiene una muestra de roca o suelo mediante técnicas modernas como es la Fluorescencia de rayos X (RFX), una técnica rápida, rentable y que tiene relativamente bajas incertidumbres y límites de detección (en el orden de pocas decenas de ppm). Lozano-Santacruz et al. (1995) utilizaron técnicas para la construcción de la mineralogía normativa en base a los elementos químicos, lo que ayuda a poder predecir el tipo de roca. Se debe indicar que estas normas no son aplicables a los sedimentos porque estos están raramente en equilibrio y constituyen casi siempre una mezcla de derivados de varias rocas de origen diferente, a menos que se tenga descrita con certeza la roca parental. Tolosana-Delgado (2011) indica que la

composición mineralógica modal es conocida por llevar mayor información de elementos geoquímicos principales, estos elementos geoquímicos son el resultado de ensayos de laboratorio. Se puede reconstruir la composición mineral de muestras, admitiendo estequiometrias de minerales (Tabla 2.2) razonables con respecto a las rocas de origen y teniendo en cuenta la información disponible (Debón and Lemmet, 1999; Deer et al., 1996).

Tabla 2.2. Estequiometria fija de minerales, enfoque geométrico, el peso de los óxidos en % para cada mineral.

Mineral	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MgO	MnO	CaO	Na_2O	K_2O	P_2O_5
Apatito	0	0	0	0	0	0	56.84	0	0	43.16
Biotita	38	3	16	19	13	1	0	0	10	0
Epidota	42.86	0	24.26	18.97	0	0	13.91	0	0	0
Clorita-Fe	25.59	0	20.45	53.96	0	0	0	0	0	0
Fluorita	0	0	0	0	0	0	100	0	0	0
Granate	38	0	20	25	4	7	6	0	0	0
Ortoclasa	64	0	19	0	0	0	0	1	16	0
Clorita-Mg	34.92	0	27.90	0	37.18	0	0	0	0	0
Moscovita	47.41	0	40.22	0	0	0	0	0.04	12.33	0
Albita-Anortita	67	0	21	0	0	0	1	11	0	0
Cuarzo	100	0	0	0	0	0	0	0	0	0
Titanita	30.65	40.74	0	0	0	0	28.61	0	0	0

2. Análisis multivariable de datos

2.1. Análisis de componentes principales

El análisis de componentes principales (PCA, por su sigla en inglés) es una herramienta de análisis multivariable usada para un amplio conjunto de propósitos: descubrimiento de posibles procesos de control (análisis factorial), compresión de datos (reducción de dimensión) y análisis exploratorio de datos multidimensionales.

El método transforma un conjunto de variables originales, en general correlacionadas entre ellas, en un nuevo conjunto de variables creadas por combinaciones lineales de las variables originales (llamadas *componentes principales* o *factores*), no correlacionadas y jerarquizadas en cuanto a la cantidad de información que capturan. Específicamente, el primer factor es el componente que explica la mayor parte de la varianza total (suma de las varianzas de las variables originales), el segundo factor es el factor que, estando no correlacionado con el primer factor, explica la mayor parte de la varianza residual, y así sucesivamente. De este modo es posible obtener tantos componentes o factores como variables originales. La construcción de los factores se basa en la diagonalización de la matriz de varianza-covarianza de las variables originales.

Camacho y Ferrer (2014) propusieron una taxonomía para las aplicaciones de PCA, dependiendo de si el objetivo es (1) la aproximación precisa de las variables observadas, como la compresión de datos o la reducción de la dimensionalidad, o (2) la comprensión y la interpretación de las variables

originales. El primer objetivo consiste en considerar solamente los primeros factores, que explican una fracción importante (mayoritaria) de la varianza total; de este modo, el análisis de componentes principales aparece como un método que permite reducir el número de variables que inicialmente se han considerado. En cuanto al segundo objetivo y que es parte del análisis del primer enfoque de esta tesis, el PCA permite proyectar los datos multivariables en sub-espacios de pequeña dimensión (típicamente, planos), identificar los datos más atípicos, o visualizar la matriz de correlación de las variables originales en un gráfico llamado "círculo de correlaciones", donde las variables están representadas por puntos dentro de un círculo unitario y la proximidad o el alejamiento de los puntos indican su mayor correlación o antagonismo (Wackernagel, 2003).

El análisis de componentes principales tiene aplicación en la geoquímica en la búsqueda de asociaciones de variables (elementos químicos), que puede proporcionar una información muy valiosa de los distintos procesos geoquímicos que se están produciendo (definición de elementos guías, yacimientos minerales, contaminantes, procedencia de aguas, etc.) (Wackernagel, 1998, 2003). Por ejemplo, Caritat y Grunsky (2013) aplicaron el análisis de componentes principales a cuatro tipos de muestras luego de haber aplicado la llamada transformación logarítmica centrada (clr). El resultado permitió la descripción e interpretación de una serie de procesos geológicos y geoquímicos.

2.2. Clasificación

Se considera una serie de observaciones con M variables cuantitativas, separadas en varias clases o categorías. El problema de la clasificación consiste en modelar las categorías a partir de las variables cuantitativas. En general, la relación entre las clases observadas y las clases modeladas se informa en una "matriz de confusión". En esta matriz, el número de casos correctamente clasificados se encuentra en la diagonal principal, mientras que el número de casos mal clasificados se encuentra en fuera de esta diagonal.

La clasificación construida permite predecir a qué categoría es más probable que pertenezca una nueva observación, es decir, permite asignar nuevas observaciones cuantitativas a la categoría que mejor corresponde en la clasificación establecida a partir de observaciones distintas.

La clasificación se relaciona con el análisis de conglomerados o agrupamiento (*cluster analysis*), la diferencia principal siendo que, en la clasificación, se conoce a priori las clases o categorías a modelar, mientras que el análisis de conglomerados sirve para formar grupos (conglomerados, clusters, racimos, etc.), inicialmente desconocidos, que sean lo más homogéneos posibles (Jain et al., 1999; Hardle and Simar, 2007; Barbosa et al, 2010). Por tanto, el objetivo esencial de la clasificación es utilizar los valores conocidos de las variables cuantitativas (variables independientes) para predecir a qué categoría de la variable dependiente (variable categórica) corresponde.

Varias técnicas han sido desarrolladas para resolver el problema de la clasificación, entre las cuales cabe mencionar el análisis discriminante, los bosques aleatorios, los árboles de decisión, las máquinas de vectores de soporte y las redes neuronales. A continuación, se entrega detalles de las dos primeras técnicas, que son parte del desarrollo de la tesis.

2.2.1. Análisis discriminante

El análisis discriminante (AD) (Hardle and Simar, 2007) busca encontrar relaciones lineales entre las variables cuantitativas que mejor discriminen las categorías previamente definidas. Esta técnica, introducida por R.A. Fisher en 1936, se base en la construcción de un conjunto de funciones lineales, conocidas como funciones discriminantes DF_i, tales como:

$$DF_i = a_1x_1 + a_2x_2 + \cdots + a_nx_n + c \qquad (2.1)$$

donde los $a's$ son coeficientes discriminantes, los $x's$ son las variables de entrada o predictores, y c es una constante. Estas funciones discriminantes son utilizadas para predecir la clase de una nueva observación D_i: esta observación se asigna a la clase cuya función discriminante tiene el valor más alto.

2.2.2. Bosques aleatorios

Los bosques aleatorios (RF, por su sigla en inglés) es un algoritmo de clasificación de conjunto supervisado y una extensión del método del árbol de decisión (Breiman, 2001). A partir de un conjunto de datos etiquetados, este clasificador construye un "bosque" que comprende muchos árboles de decisión (Fig. 2.1.), lo que permite un rendimiento superior y una menor sensibilidad al ajuste excesivo en comparación con los clasificadores individuales (Hastie et al., 2009).

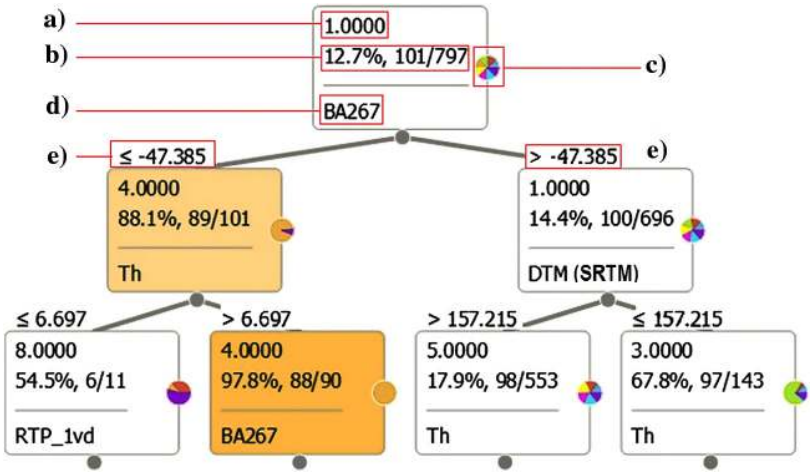

Figura 2.1. El ejemplo muestra tres niveles de un árbol de clasificación, mostrando en cada nodo: (a) la clase más numerosa, (b) la proporción de muestras de la clase más numerosa, relativo a todas las muestras en el nodo (muestra un porcentaje y cuenta del total), (c) la distribución gráfica circular de todas las clases presentes, (d) la variable utilizada para dividir el nodo padre en nodos hijos, y (e) el umbral en el cual la división fue ejecutada (Kuhn, 2018).

Este algoritmo de clasificación se ha aplicado en problemas de clasificación litológica combinando con otros algoritmos como el MLA (algoritmo de aprendizaje autónomo) (Waske et al. 2009) y el SVM (máquinas de vectores de soporte) (Vapnik 1995, 1998), en el cual se utilizan imágenes hiperespectrales, dando como resultado mejores interpretaciones que un clasificador estándar.

3. Geoestadística

La geoestadística puede ser considerada como una colección de herramientas y técnicas numéricas que se ocupan de la caracterización de variables regionalizadas, es decir, distribuidas en el espacio, empleando principalmente modelos probabilísticos (Matheron, 1963; Olea, 1999) que permiten comprender y modelar la variabilidad espacial (Deutsch and Journel, 1998). Las variables regionalizadas exhiben cierta "estructura", ya que generalmente muestran alguna forma de continuidad basada en el hecho que las ubicaciones cercanas en el espacio tienden a asumir valores cercanos (Chilès and Delfiner, 2012).

3.1. Variable regionalizada y función aleatoria

Con el supuesto de que un atributo o una variable *z* ha sido medido en diferentes objetos espaciales, se obtiene *n* observaciones tomadas en las ubicaciones x_α, con $\alpha = 1, \dots, n$, y simbolizadas por $z(x_\alpha)$. Los sitios u objetos muestreados en una región *D* pueden ser considerados como una parte de una colección más grande de objetos, en todas las posiciones posibles en *D*. Limitaciones como el costo y el esfuerzo pueden ser causas de no tomar más muestras que las *n* observaciones recopiladas. Por ejemplo, si los objetos son puntos, es posible obtener infinitas observaciones en la región. Esta posibilidad de tomar infinitas observaciones del mismo tipo es introducida eliminando el índice α y definiendo la variable regionalizada como $z(x)$ para todas las posiciones *x* en la región *D*. El conjunto de datos de muestreo es ahora considerado como una colección finita de valores de la variable regionalizada (Wackernagel, 2003).

Una variable regionalizada es una función determinista que representa, en cada ubicación de una región *D* del espacio geográfico y/o del tiempo, el valor de una variable o atributo asociado con un fenómeno natural (llamado *fenómeno regionalizado*). Una variable regionalizada generalmente involucra al menos dos aspectos geométricos: la región o dominio *D* en el cual es definida, y el objeto o soporte geométrico en el cual se mide (Wackernagel, 2003, Chilès and Delfiner, 2012).

Según Chilès and Delfiner (2012), existen tres familias de variables regionalizadas diferentes: variables cuantitativas, categóricas y de objetos. En el primer grupo se ubican las variables que se miden en una escala continua cuantitativa, mientras que en el segundo grupo se presentan las variables cuya escala es una jerarquización, indicador o clasificación desorganizada. En nuestra investigación se trabajará con valores de elementos químicos (concentraciones geoquímicas) que vendrían a ser variables cuantitativas, y tipos de rocas identificadas en la zona de estudio que serían variables categóricas.

3.1.1. Valor regionalizado, variable aleatoria y función aleatoria

Cada valor medido en el dominio *D* es llamado un valor regionalizado. Un nuevo punto de vista es considerar un valor regionalizado como la salida de algún mecanismo aleatorio subyacente, o sea, una variable aleatoria, de modo que un valor muestreado $z(x_\alpha)$ (*z* minúscula) sea el resultado de una variable aleatoria real $Z(x_\alpha)$ (*Z* mayúscula). Lo anterior se formaliza con la definición de un espacio de probabilidad (Ω,*A*,*P*) conformado por un conjunto Ω (*universo* o *espacio muestral*), una σ-álgebra *A* (conjunto de *sucesos* o *eventos*) y una medida o distribución de probabilidad *P*, siendo la variable aleatoria *Z* una aplicación medible desde Ω hacia el espacio de los números reales provisto de la σ-álgebra boreliana. La distribución de probabilidad responsable de hacer un valor

$z(x_\alpha)$ puede ser diferente en cada punto x_α de la región, por lo tanto, $Z(x_\alpha)$ podría tener una disbtribución diferente en cada punto.

La colección de todas las variables aleatorias cuando *x* pertenece a la región de interés se denomina *campo aleatorio* o *función aleatoria*. En tal configuración, los valores de los datos constituyen una realización específica de esta función aleatoria en ubicaciones determinadas dispersas en el espacio (Deutsch and Journel, 1998; Leuangthong et al., 2008).

3.1.2. Distribución a priori y a posteriori de una variable aleatoria

Como se mencionó, una variable aleatoria *Z*(*x*) se caracteriza por una distribución de probabilidad, la cual puede depender de la ubicación espacial *x* en la región *D*. Dicha distribución se describe mediante una función de distribución acumulada (*cdf*, por su sigla en inglés):

$$F(x;z) = Prob\{Z(x) \leq z\}, z \in \text{R}. \quad (2.2)$$

Esta distribución representa la incertidumbre en los valores que pueda tomar la variable aleatoria previo al conocimiento de cualquier información sobre ésta (incertidumbre "a priori"). Ahora bien, el conocimiento de *n* valores $Z(x_\alpha) = z(x_\alpha), con\ \alpha = 1, \dots, n$, medidos en posiciones vecinas de *x*, reduce la incertidumbre. Para modelar esta incertidumbre "a posteriori", se utiliza el concepto de distribución condicional. La función de la distribución acumulativa condicional (*ccdf*, por su sigla en inglés) es definida como:

$$F(x;z|(n)) = Prob\{Z(x) \leq z|Z(x_1) = z(x_1) \dots Z(x_n) = z(x_n)\}. \quad (2.3)$$

Las expresiones (2.1) y (2.2) modelan la incertidumbre sobre el valor no muestreado *z*(*x*), la primera antes de usar el conjunto de información (*n*), la segunda una vez que se ha contabilizado el conjunto de información (*n*). El objetivo de cualquier algoritmo predictivo es ir desde modelos de incertidumbre "a priori" hacia modelos "a posteriori" o "condicionales", que incorporan la información de muestreo disponible en la variable regionalizada. La *ccdf* es una función de la ubicación *x*, el tamaño de la muestra (número de datos), la configuración geométrica y los valores de esta muestra (Deutsch and Journel, 1998).

3.1.3. Distribución espacial de una función aleatoria

Una función aleatoria, como colección de variables aleatorias, se caracteriza por su *distribución espacial*, consistente en todas las distribuciones de probabilidad conjuntas asociadas a cualquier número "*p*" de posiciones en *D*:

$$Prob\{Z(x_1) < z_1 \dots Z(x_p) < z_p\}, z_1 \dots z_p \in \text{R}. \quad (2.4)$$

Al igual que en el caso de una variable aleatoria, esta distribución espacial "a priori" (antes de la toma de datos) se ve modificada cuando se dispone de mediciones sobre la función aleatoria, lo que permite definir una distribución espacial "a posteriori" que describe la incertidumbre conjunta en varias posiciones del espacio condicionalmente a la muestra observada. En la práctica, la

determinación analítica de esta distribución a posteriori es sumamente compleja y se realiza en forma numérica mediante simulaciones condicionales (ver sección 3.5 más abajo).

3.2. Análisis exploratorio de datos

Las herramientas geoestadísticas no reemplazan, si no que complementan a las herramientas estadísticas estándar. Antes de calcular variogramas y otras estadísticas espaciales, se debe realizar un análisis exploratorio de datos (EDA) (Tukey, 1977) para maximizar la visión en la comprensión y estructura de la base de datos, así como visualizar las relaciones potenciales entre variables. Para ello se debe examinar la distribución estadística de los datos; detectar valores atípicos y anomalías para dirigir la prueba específica de su validez; determinar si todas las variables han sido medidas en cada sitio de muestreo (muestreo *isotópico*) o si algunas variables están sub-muestreadas (muestreo *heterotópico*); analizar la malla de muestreo para determinar si es representativa de la región de estudio o, al contrario, es *preferencial*; elegir las variables a estudiar y, si fuese necesario, realizar un cambio de variables (por ejemplo, cuando se está en presencia de restricciones composicionales o estequiométricas); definir si la región de estudio es homogénea o si debe ser particionada en varias subregiones o "unidades geológicas" en las cuales las variables en estudio tienen un comportamiento espacial homogéneo (Rossi and Deutsch, 2014; Emery and Séguret, 2020).

Los análisis se complementan usualmente a través de gráficos (Natrella, 2010), tablas y estadísticas univariables, multivariables y/o espaciales. Las principales herramientas que se emplean para un EDA son, entre otras (Deutsch and Journel, 1998; Wackernagel, 2003; Grunsky, 2010; Chilès and Delfiner, 2012; Rossi and Deutsch, 2014; Emery and Séguret, 2020): mapas de ubicación espacial de los datos, estadísticas descriptivas de las variables, histogramas para determinar las distribuciones experimentales, gráficos de probabilidad y cuantiles contra cuantiles para la comparación de distribuciones, nubes de correlación para visualizar variables en función de otras y determinar así su relación, nubes de correlación diferida para analizar el comportamiento espacial a pequeña escala. También pueden utilizarse las técnicas de análisis multivariable de datos mencionadas en la sección 2 de este capítulo.

3.3. Análisis variográfico

Para analizar la continuidad o estructura espacial de una o varias variables regionalizadas, se definen herramientas variográficas como covarianzas y variogramas. Se asume que las funciones aleatorias parientes son *estacionarias*, es decir, su distribución espacial no varía cuando uno se traslada en el espacio, de modo que las herramientas variográficas solo dependen del vector de separación entre datos no de su posición absoluta (Wackernagel, 2003; Chilès and Delfiner, 2012). En adelante, consideramos el caso general de N variables regionalizadas $\{z_1,\dots z_N\}$ asociadas a igual número de funciones aleatorias $\{Z_1,\dots Z_N\}$.

3.3.1. Covarianzas directas y cruzadas

La covarianza cruzada entre dos funciones aleatorias estacionarias (Z_i y Z_j) para un vector de separación h se define como:

$$C_{ij}(h) = cov\{Z_i(x+h), Z_j(x)\}$$

$$= E\{Z_i(x+h)Z_j(x)\} - E\{Z_i(x+h)\} \times E\{Z_j(x)\} \quad (2.5)$$

Cuando $i = j$, se encuentra la función de autocovarianza de Z_i, también llamada covarianza simple o covarianza directa. También se puede adoptar una presentación matricial, definiendo una matriz $\mathbf{C}(h)$ de funciones de covarianzas directas y cruzadas:

$$\mathbf{C}(h) = \left[C_{ij}(h)\right]_{i,j=1,\dots,N} \quad (2.6)$$

Las covarianzas directas y cruzadas presentan las siguientes propiedades (Wackernagel, 2003; Chilès and Delfiner, 2012):

- *Simetrías y asimetrías*: $C_{ij}(h) \neq C_{ji}(h)$, $C_{ij}(h) \neq C_{ij}(-h)$, pero $C_{ij}(h) = C_{ji}(-h)$.
- *Signo y extremos*: Las covarianzas directas toman su valor máximo (positivo) en el origen ($C_{ii}(0) = \sigma_i^2 > 0$), pero las covarianzas cruzadas pueden ser funciones negativas o pueden tomar su máximo (o mínimo) en vectores *h* diferentes de 0.
- *Desigualdad de Cauchy-Schwarz*: $\forall i, j \in \{1, \dots, N\}, C_{ii}(0)C_{jj}(0) \geq \left|C_{ij}(h)\right|^2$
- *Carácter de tipo positivo*: Para todo conjunto de sitios $\{x_\alpha, \alpha = 1 \dots p\}$ y ponderadores $\{\lambda_\alpha^i, i = 1 \dots N, \alpha = 1 \dots p\}$, se tiene

  $$\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{\alpha=1}^{p}\sum_{\beta=1}^{p}\lambda_\alpha^i\, C_{ij}\left(x_\alpha - x_\beta\right)\lambda_\beta^i \geq 0 \quad (2.7)$$

- *Representación espectral*: la matriz de covarianzas directas y cruzadas $\mathbf{C}(h)$ es la transformada de Fourier de una matriz de medidas espectrales $\mathbf{F}(du)$. En caso de que estas medidas sean absolutamente continuas, se tiene $\mathbf{F}(du) = \mathbf{f}(u)du$, donde $\mathbf{f}(u)$ es la matriz de densidades espectrales. Se trata de una matriz compleja, hermitiana y de tipo semi-definido positivo (es decir, con valores propios positivos o nulos) para toda frecuencia u. Esta propiedad caracteriza el conjunto de covarianzas directas y cruzadas permisibles para representar un conjunto de funciones aleatorias estacionarias en el espacio Euclidiano. La representación espectral es de particular interés para desarrollar algoritmos eficientes de simulación multivariable (Emery et al., 2016), aplicables en particular en presencia de numerosas variables (varias decenas) y sitios a simular (miles a millones), un problema recurrente a lo largo de esta tesis.

3.3.2. Variogramas directos y cruzados

El variograma cruzado (Matheron, 1965) entre las funciones aleatorias Z_i y Z_j para un vector *h* se define como:

$$\gamma_{ij}(h) = \frac{1}{2}cov\{Z_i(x+h) - Z_i(x), Z_j(x+h) - Z_j(x)\}$$
$$= \tfrac{1}{2}E\{[Z_i(x+h) - Z_i(x)][Z_j(x+h) - Z_j(x)\}]. \quad (2.8)$$

El variograma simple o variograma directo corresponde al caso *i* = *j*. Se puede definir la matriz de variogramas de la siguiente manera:

$$\mathbf{\Gamma}(h) = [\gamma_{ij}(h)]_{i,j=1\dots N} \quad (2.9)$$

Los variogramas directos y cruzados tienen las siguientes propiedades:

- *Simetría*: $\gamma_{ij}(h) = \gamma_{ji}(h)\ y\ \gamma_{ij}(h) = \gamma_{ij}(-h)$
- *Nulidad en el origen*: $\gamma_{ij}(0) = 0$
- *Signo*: los variogramas directos son funciones positivas pero los variogramas cruzados no lo son necesariamente.
- *Desigualdad de Cauchy-Schwarz*: $\forall i, j \in \{1, \dots, N\}, \gamma_{ii}(h)\gamma_{jj}(h) \geq \left|\gamma_{ij}(h)\right|^2$
- *Carácter de tipo positivo*: para un vector h dado, la matriz $\Gamma(h)$ es simétrica y de tipo positivo, es decir que:

  $$\forall \lambda_1 \dots \lambda_N \in \mathbb{R}, \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j \gamma_{ij}(h) \geq 0 \qquad (2.10)$$

- *Relación con la función de covarianza cruzada*:

  $$\gamma_{ij}(h) = C_{ij}(0) - \frac{1}{2}[C_{ij}(h) + C_{ij}(-h)] \qquad (2.11)$$

  Matricialmente:

  $$\mathbf{\Gamma}(h) = \mathbf{C}(0) - \frac{1}{2}[\mathbf{C}(h) + \mathbf{C}(-h)] \qquad (2.12)$$

El variograma cruzado omite el termino impar de la covarianza cruzada, por lo que contiene menos información sobre la correlación espacial conjunta de ambas funciones aleatorias Z_i y Z_j que la covarianza cruzada. En particular, existen fenómenos donde las variables tienen un *efecto de retardo* (desfase máximo de la función covarianza cruzada respecto al origen), que no se pueden modelar utilizando los variogramas cruzados (Wackernagel, 2003).

3.3.3. Variogramas experimentales y modelados

El variograma cruzado experimental entre dos variables regionalizadas z_i y z_j se define como:

$$\widehat{\gamma_{ij}}(h) = \frac{1}{2|N_{ij}(h)|} \sum_{N_{ij}(h)} \left[z_i(x_\alpha) - z_i(x_\beta)\right]\left[z_j(x_\alpha) - z_j(x_\beta)\right] \qquad (2.13)$$

donde $N_{ij}(h) = \{(\alpha, \beta)$ tal que $x_\alpha - x_\beta = h$, siendo a la vez z_i y z_j medidas en x_α y $x_\beta\}$.

El variograma indicado en la ecuación anterior es un estimador del variograma teórico γ_{ij} definido en la ecuación (2.8). Se obtiene de manera experimental, porque depende de los datos disponibles y considera solamente determinadas distancias y orientaciones de acuerdo al vector h. Para poder ser calculado, el variograma cruzado experimental necesita tener datos de las variables en los mismos sitios. Por lo tanto, no se puede calcular en el caso de heterotopía total (caso en el cual el conjunto $N_{ij}(h)$ es vacío).

Los variogramas experimentales, directos y cruzados, deben modelarse porque se necesita tener su expresión en forma continua para los propósitos de predicción o simulación (ver las secciones 3.4 y 3.5 más adelante). Para el modelamiento, se suele utilizar el llamado modelo lineal de corregionalización, que consiste en utilizar combinaciones de variogramas elementales llamados

estructuras anidadas (Journel and Huijbregts, 1978; Wackernagel, 2003). Entre los modelos elementales más comunes están el efecto pepita, así como los variogramas esférico, exponencial y Gaussiano (Fig. 2.2.).

La ecuación general del modelo en forma matricial se define como sigue:

$$\boldsymbol{\Gamma}(h) = \sum_{s=1}^{S} \mathbf{B}_s g_s(h) \qquad (2.14)$$

donde $\boldsymbol{\Gamma}(h)$ es la matriz de variogramas directos y cruzados para un determinado vector h, S es el número de estructuras anidadas, $\mathbf{B}_s = \left[b_{ij}^s\right]$ (con $i, j = 1, \ldots, N$, y $s = 1, \ldots, S$) se define como una matriz de corregionalización (simétrica y de tipo semi-definida positiva) y $g_s(h)$ es un modelo elemental de variograma, elegido por el usuario entre los típicamente usados (exponencial, esférico, Gaussiano, etc). En la práctica, existen algoritmos que permiten determinar las matrices de corregionalización en forma automática, de modo de minimizar el error cuadrático total entre los variogramas experimentales y modelados (Goulard and Voltz, 1992; Emery, 2010; Desassis and Renard, 2013).

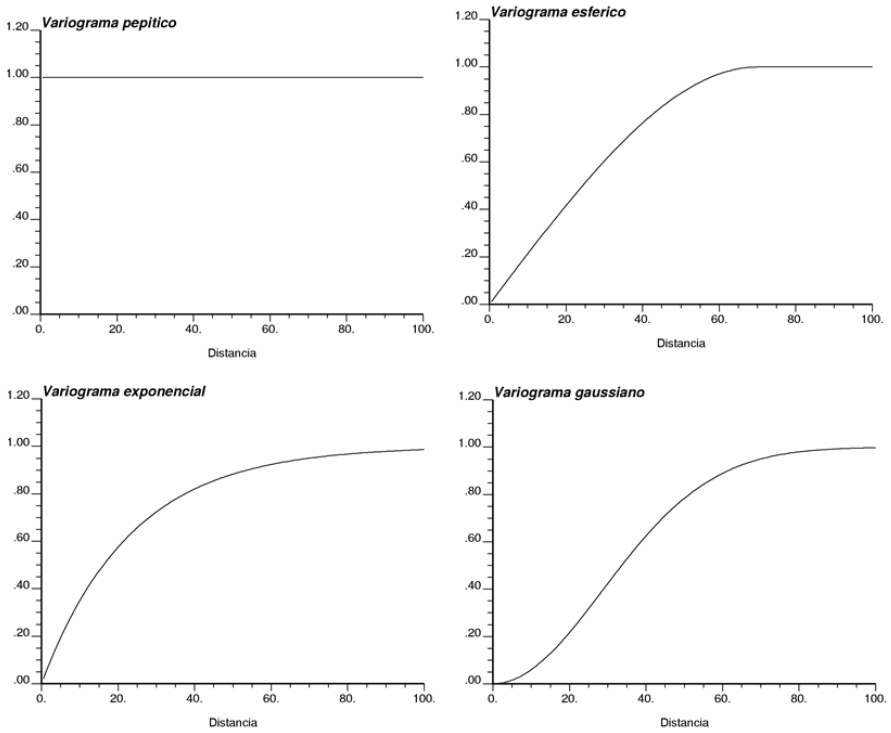

Figura 2.2. Modelos elementales para la construcción de variogramas anidados

3.4. Predicción espacial: cokriging

El cokriging corresponde a una técnica de predicción que permite interpolar el valor de una o varias variables regionalizadas, que usualmente presentan correlación entre ellas (Wackernagel, 2003; Chilès and Delfiner, 2012). En caso de una sola variable, se habla de kriging. En la investigación de tesis realizada se trabajó con la formulación multivariable general (cokriging), ya que se tiene varias variables.

3.4.1. Cokriging simple

El cokriging simple corresponde al caso donde las medias (valores esperados) de las funciones aleatorias parientes son conocidas. Se puede formalizar como sigue:

$$\boldsymbol{Z}^*(x_0) = \boldsymbol{a} + \sum_{\alpha=1}^{n} \wedge_\alpha^T \boldsymbol{Z}(x_\alpha) \tag{2.15}$$

donde N es el número de funciones aleatorias o variables, n es el número de datos en una vecindad del punto x_0 a predecir, $\{x_\alpha, \alpha = 1, \dots, n\}$ son las posiciones con datos, $\boldsymbol{Z}(x_\alpha)$ es el vector con las funciones aleatorias en el punto x_α, $\boldsymbol{a}$ es un vector de tamaño $N \times 1$, $\{\wedge_\alpha, \alpha = 1, \dots, n\}$ son matrices de tamaño $N \times N$ (ponderadores de cokriging) y $\boldsymbol{Z}^*(x_0)$ es el vector con la predicción de todas las funciones aleatorias en la posición x_0. (Se utilizan aquí negritas para denotar vectores y diferenciar de escalares usados en el caso univariable; la T en exponente indica el operador de trasposición matricial).
Se tiene:

$$\boldsymbol{a} = \boldsymbol{m} - \sum_{\alpha=1}^{n} \wedge_\alpha^T \boldsymbol{m} \tag{2.16}$$

donde $\boldsymbol{m}$ es un vector de tamaño $N \times 1$ con los valores de las medias conocidas.

Los ponderadores $\wedge_\alpha$ se determinan por medio del sistema de ecuaciones lineales que se presenta a continuación:

$$\begin{pmatrix} \mathbf{C}(x_1 - x_1) & \dots & \mathbf{C}(x_1 - x_n) \\ \vdots & \ddots & \vdots \\ \mathbf{C}(x_n - x_1) & \dots & \mathbf{C}(x_n - x_n) \end{pmatrix} \begin{pmatrix} \wedge_1 \\ \vdots \\ \wedge_n \end{pmatrix} = \begin{pmatrix} \mathbf{C}(x_1 - x_0) \\ \vdots \\ \mathbf{C}(x_n - x_0) \end{pmatrix} \tag{2.17}$$

donde $\mathbf{C}(x_\alpha - x_\beta)$ es una matriz de tamaño $N \times N$ cuyo término genérico $C_{ij}(x_\alpha - x_\beta)$ es la covarianza cruzada de las variables *i, j* entre los puntos x_α y x_β.

En el caso de datos faltantes, se remueven las filas y columnas asociadas a los índices de aquellos datos en el sistema matricial anterior.

Esta aplicación se considera adecuada cuando se conoce perfectamente las medias de todas las variables en el espacio. Para los casos contrarios, asumir que las medias son conocidas es una hipótesis fuerte que limita el resultado de la predicción.

3.4.2. Cokriging ordinario

Plantea que, debido al desconocimiento en las medias de las variables a predecir, se agrega una restricción para evitar que el predictor tenga sesgo (i.e., para que el error de predicción tenga una esperanza nula), la que se traduce en que la suma de los ponderadores de la variable a predecir sea igual a uno, mientras que los de cada covariable suman cero.
La expresión del predictor es la siguiente.

$$\boldsymbol{Z}^*(x_0) = \sum_{\alpha=1}^{n} \wedge_\alpha^T \boldsymbol{Z}(x_\alpha) \tag{2.18}$$

con las mismas notaciones empleadas en el cokriging simple.

Las matrices de ponderadores $\wedge_\alpha$ se determinan por medio del sistema de ecuaciones lineales que se presenta a continuación:

$$\begin{pmatrix} \boldsymbol{\Gamma}(x_1 - x_1) & \dots & \boldsymbol{\Gamma}(x_1 - x_n) & \mathbf{I} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{\Gamma}(x_n - x_1) & \dots & \boldsymbol{\Gamma}(x_n - x_n) & \mathbf{I} \\ \mathbf{I} & \dots & \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \wedge_1 \\ \vdots \\ \wedge_n \\ -\boldsymbol{M} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Gamma}(x_1 - x_0) \\ \vdots \\ \boldsymbol{\Gamma}(x_n - x_0) \\ \mathbf{I} \end{pmatrix} \tag{2.19}$$

donde $\boldsymbol{\Gamma}(x_\alpha - x_\beta)$ es una matriz de tamaño $N \times N$ cuyo término genérico $\gamma_{ij}(x_\alpha - x_\beta)$ es el variograma cruzado de las variables *i, j* entre los puntos x_α y x_β, $\boldsymbol{M}$ es una matriz de tamaño $N \times N$ (multiplicadores de Lagrange), $\mathbf{0}$ corresponde a una matriz de ceros de tamaño $N \times N$ e $\mathbf{I}$ es la matriz identidad de tamaño $N \times N$.

Nuevamente, en el caso de datos faltantes, se remueven las filas y columnas asociadas a los índices de aquellos datos en el sistema matricial anterior.

3.4.3. Propiedades del cokriging

La aplicación de cokriging posee las siguientes propiedades:

- Interpolación exacta, es decir se restituye el valor de un dato en un sitio con dato medido.
- Aditividad, que implica que el valor predicho en un bloque es el promedio de los valores puntuales predichos dentro de éste.
- Suavizamiento, que se refleja en que las predicciones son menos variables (estadística y espacialmente) que los datos originales.

3.5. Simulación

Las técnicas de determinación de los tipos de rocas se dividen en modelos deterministas y modelos estocásticos. Mientras que los primeros entregan un resultado único basado en la experiencia operacional, métodos manuales, técnicas gráficas, interpretaciones de vista y secciones o en predicciones, los segundos entregan múltiples escenarios equiprobables, llamados *simulaciones* o *realizaciones*, que reproducen la variabilidad espacial del fenómeno regionalizado.

Las técnicas de simulación se basan en la interpretación de la variable regionalizada como una realización de una función aleatoria y en el modelamiento de su distribución espacial (Lantuéjoul, 2002; Chilès and Delfiner, 2012). A través de la construcción de numerosos escenarios, las simulaciones permiten cuantificar la incertidumbre asociada al desconocimiento de los valores reales de una variable regionalizada. Según se consideren o no los datos de muestreo al momento de simular se hablará de simulación condicional (incertidumbre "a posteriori", representada teóricamente por la *ccdf* introducida en la sección 3.1.2) o no condicional (incertidumbre "a priori", representada teóricamente por la *cdf*). Al igual que en el caso de las predicciones, la simulación puede ser realizada considerando múltiples variables, en cuyo caso se denomina cosimulación o simulación conjunta.

El modelo de simulación empleado depende en gran medida del tipo de variable a simular, diferenciando los casos de variables cuantitativas (concentraciones geoquímicas), categóricas (tipo de roca) u objetos (fracturas, fallas). De acuerdo a lo expuesto por Chilès and Delfiner (2012), los principales modelos empleados según el tipo de variable se resumen a continuación:

- variables cuantitativas: modelo multigaussiano;
- variables categóricas: modelo gaussiano truncado, modelo plurigaussiano, simulación secuencial de indicadores, simulación con estadísticas multipuntos;
- objetos: modelo Booleano, procesos de puntos marcados.

Las simulaciones reproducen la variabilidad espacial original observada en los datos y permiten una evaluación de la incertidumbre del fenómeno que se estudia en sitios no muestreados. En particular, los valores extremos de la distribución original son conservados y no son eliminados como el caso de la predicción (cokriging) por su efecto de suavizamiento. Los escenarios simulados pueden ser utilizados para: (i) análisis de riesgo, examinando la simulación más optimista y la simulación más pesimista o calculando la frecuencia de ocurrencia de un evento para estimar la probabilidad de este evento, (ii) predicción, promediando las simulaciones, (iii) evaluación de la incertidumbre, verificando cuán diferentes son las simulaciones, (iv) análisis de sensibilidad y riesgo frente a escenarios base, pesimistas y optimistas (Goovaerts, 1997; Deutsch, 1998; Rossi and Deutsch, 2014; Emery and Séguret, 2020).

3.5.1. Simulación multigaussiana para variables cuantitativas

3.5.1.1. *Modelo*

Las técnicas de simulación tienen como objetivo reproducir la distribución espacial de la función aleatoria estudiada. Sin embargo, esta restricción es muy fuerte, pues en general no es posible inferir una distribución espacial a partir de una cantidad limitada de datos. Una excepción lo constituye el caso de las funciones aleatorias multigaussianas (llamadas también "funciones aleatorias gaussianas"), para las cuales la distribución espacial queda enteramente caracterizada por sus dos primeros momentos, luego el modelo se reduce a especificar una media y una función de covarianza o un variograma.

Debido a lo poco frecuente que una variable regionalizada tenga una distribución gaussiana, se requiere una transformación previa, llamada *anamorfosis*, que consiste en deformar la distribución de los datos en una distribución gaussiana estándar (Fig. 2.3.), es decir, de media 0 y varianza 1 (Chilès and Delfiner, 2012). En otras palabras, como en general la distribución de la función aleatoria pariente $\{Z(x): x \in \mathbb{R}^d\}$ que se desea simular no tiene distribución gaussiana, no se trabaja directamente sobre ella, sino que sobre su transformada gaussiana $\{Y(x): x \in \mathbb{R}^d\}$, definida por:

$$Z(x) = \Phi[Y(x)] \qquad (2.20)$$

donde $\Phi = \mathrm{F}^{-1} \circ G$ (función de anamorfosis gaussiana), siendo F y G las funciones de distribución de $Z(x)$ y de la gaussiana estándar, respectivamente.

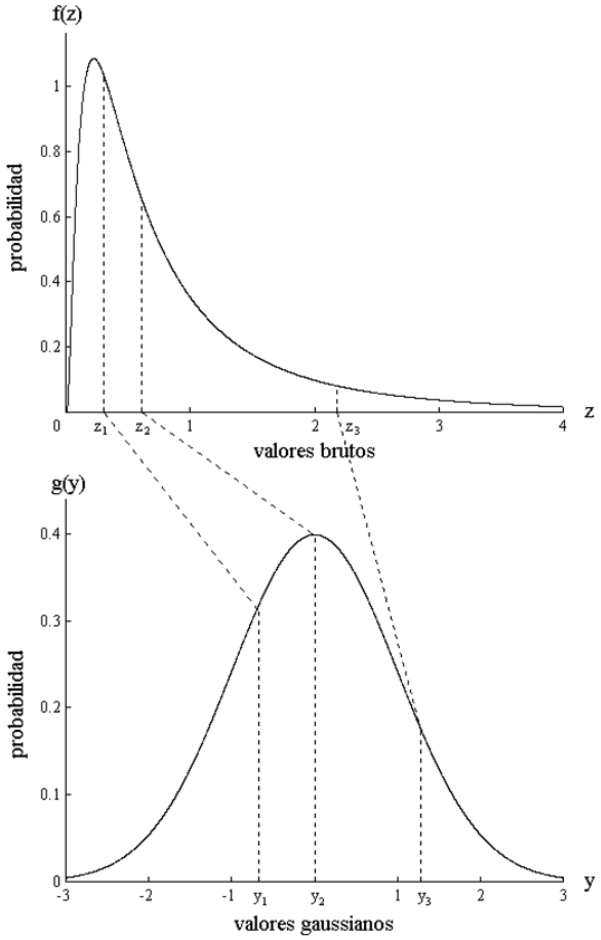

Figura 2.3. Construcción gráfica de la anamorfosis Gaussiana.

La hipótesis que $\{Y(x): x \in \mathbb{R}^d\}$ tiene una distribución multigaussiana implica que la densidad de probabilidad de un conjunto de valores ubicados en los sitios $\{x_1, \dots, x_n\}$ es de la forma:

$$g(y_1, \dots, y_n) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det(\mathbf{C})}} \exp\left\{-\frac{1}{1} y^t \mathbf{C}^{-1} y\right\} \tag{2.21}$$

donde $y = (y_1, \dots, y_n)^t$ y $\mathbf{C}$ es la matriz de varianza-covarianza del vector $\big(Y(x_1), \dots, Y(x_n)\big)$. Las distribuciones de probabilidad solo dependen de las varianzas y covarianzas, por lo que el modelo queda enteramente determinado una vez que se ha ajustado la función de covarianza o, equivalentemente, el variograma de los datos transformados.

En caso de varias variables regionalizadas, se requiere transformar cada variable a una gaussiana y determinar las covarianzas directas y cruzadas de las variables transformadas. Una alternativa a la transformación gaussiana de cada variable separadamente, es una transformación multivariable (co-anamorfosis) que permite que las variables transformadas tengan, conjuntamente, una distribución mas cercana a una distribución multigaussiana. Entre las propuestas de anamorfosis multivariable, destacan:

- la transformación condicional paso a paso (Leuangthong and Deutsch, 2003), basada en un esquema jerárquico donde la transformación de una variable queda supeditada a aquella de las variables previas;
- la factorización de las variables en componentes independientes (Tercan and Sohrabian, 2013), que se trasnforman y simulan en forma separada;
- la transformación basada en búsqueda de proyección (*projection pursuit multivariate transform*, PPMT), donde se aplica una serie de transformaciones gaussianas univariables, cada una a lo largo de un vector (combinación de las variables a transformar) identificado como aquel cuya distribución marginal es la más alejada de una distribución gaussiana (Barnett et al., 2013);
- la transformación basada en deformación de flujo (*flow anamorphosis*), que deforma continuamente una distribución multivariable experimental en una distribución multigaussiana (van den Boogaart et al., 2017).

3.5.1.1. *Un algoritmo de simulación: bandas rotantes*

Este algoritmo, utilizado primero por Chentsov (1957) y extendido por Matheron (1973), reduce el problema de la simulación en un espacio de varias dimensiones a un problema de simulación unidimensional.

a) **Caso univariable**

Consiste en simular la función aleatoria de interés a lo largo de rectas que discretizan el espacio, para posteriormente esparcir estas rectas al espacio total y considerar la simulación como la suma de ellas mediante la siguiente expresión:

$$Y(x) = \frac{1}{\sqrt{L}} \sum_{i=1}^{L} Y_i^1 (< X | u_i >) \qquad (2.22)$$

donde $\{u_i : i = 1, \dots, L\}$ son vectores distribuidos en la esfera unitaria, $\{Y_i^1 : i = 1, \dots, L\}$ son simulaciones unidimensionales independientes y $< | >$ representa el producto escalar usual. Los vectores se pueden elegir con direcciones uniformes o, preferentemente, casi regularmente distribuidas (secuencias equidistribuidas). Dentro de sus principales ventajas se encuentra la rapidez del algoritmo y que, además, se reproduce la covarianza o el variograma de forma exacta. Su principal debilidad radica en el hecho de que la función aleatoria simulada no es exactamente gaussiana, lo que se resuelve mediante la elección del número L (usualmente en varios miles) de las direcciones (Lantuéjoul, 2002; Emery and Lantuéjoul, 2006).

b) **Caso multivariable**

La simulación multivariable puede realizarse de varias maneras. Por ejemplo, consideremos un conjunto de N funciones aleatorias gaussianas $Y_1,\dots, Y_N$ cuya estructura de correlación corresponde a un modelo lineal de corregionalización con S estructuras básicas:

$$\boldsymbol{\Gamma}(h) = \sum_{s=1}^{S} \mathbf{B}_s g_s(h) \qquad (2.23)$$

Entonces el vector de funciones aleatorias $\boldsymbol{Y} = (Y_1,\dots, Y_N)$ se puede descomponer como sigue:

$$\boldsymbol{Y}(x) = \sum_{s=1}^{S} \boldsymbol{Z}_s(x) \tag{2.24}$$

donde $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_S$ son vectores de N funciones aleatorias gaussianas independientes, cada uno de ellos asociado a una estructura anidada, es decir, los variogramas directos y cruzados de $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_S$ son $\mathbf{B}_1 g_1, \ldots, \mathbf{B}_S g_S$, respectivamente.

A su vez, cada matriz de corregionalización puede descomponerse según la siguiente expresión:

$$\mathbf{B}_s = \mathbf{Q}_s \Delta_s \mathbf{Q}_s^T = \mathbf{A}_s \mathbf{A}_s^T \tag{2.25}$$

donde $\mathbf{Q}_s$ corresponde a la matriz $N \times N$ de vectores propios y Δ_s es la matriz diagonal de valores propios, también de tamaño $N \times N$. Se tiene entonces:

$$\boldsymbol{Z}_s(x) = \mathbf{A}_s \boldsymbol{W}_s(x) \tag{2.26}$$

donde $\boldsymbol{W}_s$ es un vector de N funciones aleatorias gaussianas independientes, cada una de ellas con el variograma g_s. En palabras simples, la simulación de N funciones aleatorias correlacionadas (componentes de $\boldsymbol{Y}$) se reduce a la de $N \times S$ funciones aleatorias independientes (componentes de $\boldsymbol{W}_s$, para $s = 1 \ldots S$), cuyos variogramas corresponden a las estructuras anidadas básicas utilizadas en el modelo lineal de corregionalización (Emery, 2008).

Una alternativa que no requiere la descomposición del modelo de corregionalización consiste en simular las funciones aleatorias con una sumatoria ponderada de funciones coseno, cuyas frecuencias y fases son elegidas aleatoriamente (método espectral-bandas rotantes) (Emery et al., 2016). Esta variante se basa en el cálculo de la matriz de densidades espectrales asociadas a la matriz de covarianzas directas y cruzadas.

c) **Condicionamiento**

Con el objetivo de restituir los valores originales en las simulaciones se debe aplicar una etapa posterior de condicionamiento al método de bandas rotantes (Journel and Huijbregts, 1978; Chilès and Delfiner, 2012). Este consiste en la aplicación de cokriging, vía el método de "sustitución de los residuos", donde el valor simulado queda representado por la siguiente expresión:

$$\boldsymbol{Y}_{SC}(x) = \boldsymbol{Y}^{CKS}(x) + [\boldsymbol{Y}_{SNC}(x) - \boldsymbol{Y}_{SNC}^{CKS}(x)] \tag{2.27}$$

donde $\boldsymbol{Y}^{CKS}(x)$ es el cokriging simple de las funciones aleatorias a simular a partir de los datos condicionantes, $\boldsymbol{Y}_{SNC}(x)$ es una simulación no condicional en el sitio x y $\boldsymbol{Y}_{SNC}^{CKS}(x)$ es el cokriging simple de la simulación no condicional a partir de los valores de dicha simulación en los sitios con datos. En la práctica, basta con un solo cokriging para condicionar varias simulaciones en un sitio determinado, ya que los ponderadores no varían de una simulación a otra. Es posible sustituir el cokriging simple por uno ordinario, en caso de que se suponga que las medias son desconocidas.

3.5.2. Simulación plurigaussiana para variables categóricas

La simulación plurigaussiana fue introducida por Galli et al. (1994) para simular variables categóricas, representando tipos de roca o dominios geológicos. Está diseñada para adaptarse a un rango amplio de fenómenos, permitiendo numerosos tipos de contactos entre los dominios geológicos bajo consideración, y es actualmente usado tanto en la industria petrolera como minera (Armstrong et al., 2011). También ha habido aplicaciones en otros campos de las geociencias, como la hidrogeología (Mariethoz et al., 2009).

La idea de la simulación plurigaussiana es simular funciones aleatorias gaussianas $Y_1, Y_2, \ldots, Y_N$ en cada punto de la región de estudio; a menudo, $N = 2$, pero ha habido aplicaciones con un mayor número de funciones aleatorias (Xu et al., 2006; Emery, 2007; Madani and Emery, 2015, 2017; Maleki et al., 2016). Luego se usa una regla de truncamiento para convertir los valores gaussianos en valores categóricos (dominios geológicos).

3.5.2.1. *Inferencia de los parámetros del modelo*

- **Regla de truncamiento**

Para delinear el contacto entre dominios geológicos, se requiere definir una partición del espacio $\mathbb{R}^N$ (regla de truncamiento), en la que cada clase de la partición corresponde a un dominio geológico. El diseño de la partición tiene implicaciones en las relaciones espaciales entre los dominios geológicos, en particular en los contactos permisibles y prohibidos y en la secuencia o el orden cronológico de los dominios (Le Loc'h et al., 1994; Lantuéjoul, 2002; Armstrong et al., 2011; Madani and Emery, 2015). La Fig. 2.4. muestra que los dominios geológicos que se tocan entre sí en la regla de truncamiento están también en contacto en el espacio.

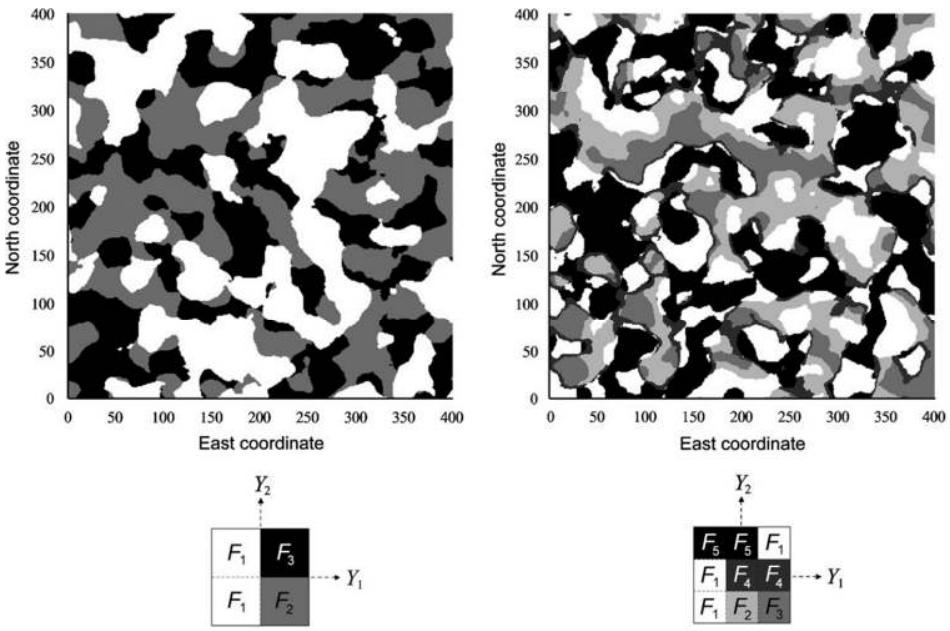

Figura 2.4. Ejemplos de simulaciones plurigaussianas obtenidas al truncar dos funciones aleatorias gaussianas independientes. La regla de truncamiento es representada por una bandera debajo de la simulación. Cada dominio geológico (F1-F3 o F1-F5) está asociado con un tono gris especifico (Emery, 2007)

- **Umbrales**

Para N funciones aleatorias gaussianas y M dominios geológicos, usualmente nos encontramos con $M - 1$ umbrales que deben ser definidos según la proporción de cada dominio. Denotemos como $g(.)$ la densidad de probabilidad multigaussiana. Para calcular la probabilidad del dominio i (F_i) en la ubicación x, se necesita resolver la siguiente ecuación:

$$P_{F_i}(x) = \int_{D_i} g(y_1, y_2, \dots, y_N) dy_1 dy_2 \dots dy_N \qquad (2.28)$$

donde D_i es la porción del espacio $\mathbb{R}^N$ asociada a F_i en la regla de truncamiento.

Incluso con dos funciones aleatorias gaussianas ($N = 2$), este sistema de ecuaciones a menudo no puede ser resuelto analíticamente, pero métodos iterativos de prueba y error pueden dar la solución. La Fig. 2.5. muestra un ejemplo donde el método de prueba y error es rápido, donde la partición deseada se muestra en la línea superior; la segunda línea muestra el orden en que se evalúan los umbrales, comenzando con el bloque superior.

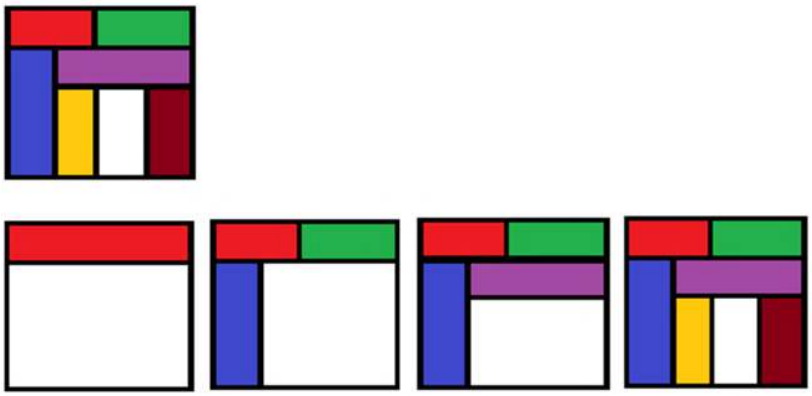

Figura 2.5. Agrupamientos sucesivos para obtener los umbrales de truncamiento (modificado de Armstrong et al., 2011)

- **Covarianzas o variogramas de las funciones aleatorias gaussianas**

Para cualquier vector de separación h, el variograma cruzado entre indicadores de dos dominios geológicos (con índices i y j) es derivado de la correspondiente covarianza no centrada:

$$\gamma_{ij}(\mathrm{h}) = C_{ij}(0) - \frac{1}{2}[C_{ij}(\mathrm{h}) + C_{ij}(-\mathrm{h})] \qquad (2.29)$$

Si D_i y D_j son paralelepípedos rectangulares de $\mathbb{R}^N$ y las funciones aleatorias gaussianas son independientes, el segundo miembro de la ecuación 2.29 es una función de los variogramas o de las covarianzas directas de las funciones aleatorias gaussianas y puede ser calculado por integración numérica (Dowd et al., 2003) o utilizando desarrollos en polinomios de Hermite (Emery, 2007). Esto establece un vínculo entre los variogramas de las funciones aleatorias gaussianas subyacentes y los variogramas directos y cruzados de indicadores. Estos últimos son experimentalmente accesibles a partir de la información de los dominios geológicos observados en sitios de muestreo.

Los variogramas de las funciones aleatorias gaussianas pueden, por lo tanto, ser determinados de manera de ajustar los variogramas de indicadores, por ejemplo, a través de un procedimiento de prueba y error (Le Loc'h and Galli, 1997, Emery, 2007, Armstrong et al., 2011), o a través de una inversión de la relación entre variogramas de indicadores y variogramas de las funciones gaussianas (Emery and Cornejo, 2010).

3.5.2.1. *Simulación condicional*

Una vez especificados los parámetros del modelo plurigaussiano, la simulación puede ser realizada en tres pasos principales (Lantuéjoul, 2002; Dowd, 2003; Emery, 2007; Armstrong et al., 2011):

1. Simular las N funciones aleatorias gaussianas subyacentes $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_N)$ en las ubicaciones de datos, condicionando a los datos categóricos.
2. Simular las funciones aleatorias gaussianas en las ubicaciones objetivo, condicionando a los valores obtenidos en el paso anterior.
3. Aplicar la regla de truncamiento para obtener los dominios simulados.

El tercer paso es sencillo, mientras que el segundo paso puede realizarse por cualquier algoritmo para simular funciones aleatorias gaussianas estacionarias, por ejemplo, el algoritmo de bandas rotantes. En cuanto al primer paso, puede ser realizado por un algoritmo iterativo conocido como el *muestreador de Gibbs* (Geman and Geman, 1984; Casella and George, 1992; Armstrong et al., 2011), que pertenece a la familia de los enfoques de la Cadena de Markov de Monte Carlo (MCMC) (Lantuéjoul, 2002; Chilès and Delfiner, 2012). El procedimiento general es el siguiente:

a) **Inicialización**

Para cada ubicación con dato x_α, generar un vector con N componentes $\boldsymbol{z}_\alpha$ en $D_{i(\alpha)}$, donde $i(\alpha)$ es el índice del dominio geológico presente en x_α.

b) **Iteración**

a. Seleccionar una ubicación de dato x_α, regularmente o aleatoriamente.
b. Calcular la distribución de $\boldsymbol{Y}(x_\alpha)$ condicional a los otros datos $\{\boldsymbol{Y}(x_\beta): \beta \neq \alpha\}$. En el caso estacionario, ésta es una distribución gaussiana, con media igual a la predicción de $\boldsymbol{Y}(x_\alpha)$ por cokriging simple y una matriz de varianza-covarianza igual a la matriz de varianza-covarianza de los errores de cokriging simple.
c. Simular un vector $\boldsymbol{y}_\alpha$ de acuerdo a la distribución condicional anterior.
d. Si $\boldsymbol{y}_\alpha$ es compatible con el dominio geológico presente en x_α (es decir, $\boldsymbol{y}_\alpha \in D_{i(\alpha)}$), reemplazar el valor actual de $\boldsymbol{Y}(x_\alpha)$ por $\boldsymbol{y}_\alpha$.
e. Regresar al paso (a) y repetir numerosas veces.

El muestreador de Gibbs así presentado es una cadena de Markov irreducible, aperiódica y reversible, con la distribución Gaussiana objetivo como su límite ergódico. En otras palabras, si el número de iteraciones aumenta infinitamente, la distribución de los vectores simulados en las ubicaciones con datos converge a la distribución condicional de las funciones aleatorias gaussianas deseadas (Lantuéjoul, 2002).

3.5.3. Simulación conjunta de variables cuantitativas y categóricas

3.5.3.1. *Enfoque jerárquico*

Un enfoque simple y directo para simular una o varias variables cuantitativas junto con una variable categórica es el enfoque jerárquico. El área de cada categoría (dominio geológico o tipo de roca) existente debe primero estar delimitada en la región de estudio. Para este fin, se puede utilizar modelos deterministas o técnicas de geomodelamiento (Royer et al., 2015) o modelos estocásticos (gaussiano truncado, plurigaussiano u otro) basados en simular la ocurrencia de cada dominio geológico (Armstrong et al., 2011; Mariethoz and Caers, 2014). Después de la delimitación de todos los dominios geológicos, las variables cuantitativas son simuladas en cada dominio por separado, por ejemplo, con el modelo multigaussiano, utilizando solo los datos de muestreo que pertenecen a este dominio para el condicionamiento de la simulación (Alabert and Massonnat, 1990; Haldorsen and Damsleth, 1990; Dubrule, 1993; Talebi et al., 2016; Paithankar and Chatterjee, 2018; Maleki and Emery, 2020).

Aunque este es un método simple, presenta algunos inconvenientes. En particular, produce transiciones abruptas (discontinuidades) en el valor de las variables cuantitativas al atravesar los límites entre dominios geológicos. En otras palabras, se producen límites "duros". Además, como las variables continuas se simulan en cada dominio geológico por separado, existe una mayor incertidumbre en los dominios con menos datos (los datos de otros dominios se ignoran al momento de condicionar la simulación).

3.5.3.1. *Simulación conjunta*

En este método de simulación, las variables cuantitativas se modelan (después de anamorfosis) por un conjunto de funciones aleatorias gaussianas $Y_1, Y_2, \dots, Y_P$ (modelo multigaussiano), mientras que la variable categórica se modela por el truncamiento de otro conjunto de funciones aleatorias gaussianas $Y_{P+1}, Y_{P+2}, \dots, Y_{P+N}$ (modelo plurigaussiano). La idea subyacente es que todas estas funciones aleatorias estén correlacionadas entre ellas, lo que permite correlacionar la variable categórica con las variables cuantitativas.

La inferencia de las covarianzas directas y cruzadas de $Y_1, Y_2, \dots, Y_{P+N}$ se basa en el ajuste de las covarianzas o variogramas directos y cruzados de los datos gaussianos (caso de las variables cuantitativas) y de los datos de indicadores (caso de la variable categórica). La simulación se lleva a cabo de forma similar al modelo plurigaussiano: primero, el muestreador de Gibbs se usa para simular las funciones aleatorias $Y_{P+1}, Y_{P+2}, \dots, Y_{P+N}$ en los sitios con datos, luego se cosimulan las funciones aleatorias $Y_1, Y_2, \dots, Y_{P+N}$ en la región de interés, finalmente se transforma de vuelta las primeras P funciones aleatorias $Y_1, Y_2, \dots, Y_P$ para obtener las variables cuantitativas y se trunca las últimas N funciones aleatorias $Y_{P+1}, Y_{P+2}, \dots, Y_{P+N}$ para obtener la variable categórica. La Fig. 2.6. muestra los diferentes pasos de la simulación conjunta.

Este enfoque, con algunas variantes o simplificaciones, fue aplicado por varios autores a casos con una sola variable cuantitativa (Dowd, 1994, 1997; Bahar and Kelkar, 2000; Emery and Silva, 2009; Cáceres and Emery, 2010; Maleki and Emery, 2015, 2017). Uno de los aspectos de esta propuesta de tesis es extender el modelo y su aplicación a varias decenas de variables cuantitativas (concentraciones geoquímicas), posiblemente no conocidas en los mismos sitios del espacio.

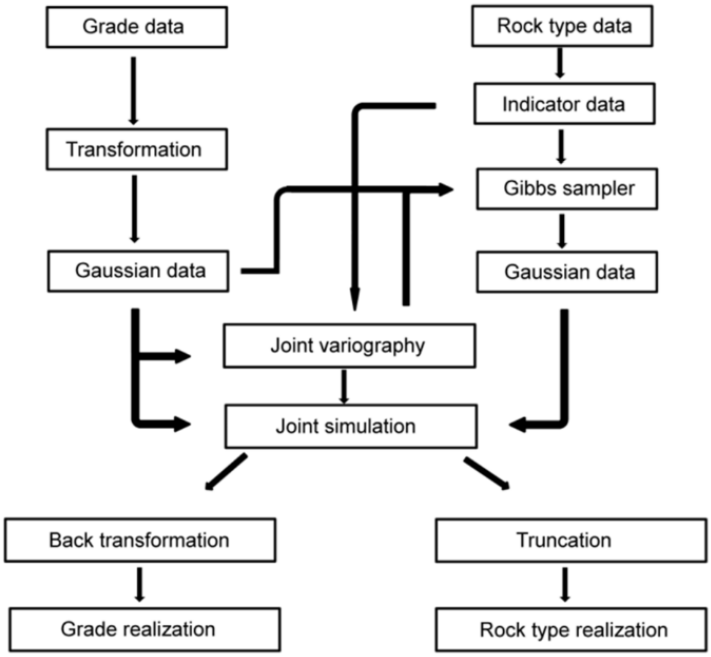

Figura 2.6. El proceso de simulación conjunta (Maleki and Emery, 2015)

La simulación conjunta toma en cuenta la incertidumbre existente en los límites de los dominios geológicos, que es ignorada cuando se utiliza un modelo geológico determinista. También considera la correlación espacial de las variables cuantitativas a través de los límites geológicos, que es ignorada cuando se realiza una simulación jerárquica de la variable categórica y de las variables cuantitativas, produciendo límites "blandos" (Maleki and Emery, 2015).

Capitulo 3. Metodología propuesta

Se proponen dos enfoques principales para la predicción de una variable categórica a partir de esa información categórica y de varias covariables cuantitativas. Estos enfoques se aplicarán a casos de estudio diferentes en los capítulos 4 y 5, respectivamente. La validación de ambos enfoques se realiza mediante la técnica de jack-knife, consistente en subdividir la base de datos originales en dos subconjuntos:

- un *subconjunto de entrenamiento*, con el cual se ajustan los diferentes modelos y se condicionan las simulaciones de las variables cuantitativas y/o categórica;
- un *subconjunto de prueba* o de validación, sobre el cual se construyen las simulaciones y se predice la variable categórica, lo que permite comparar la predicción con los valores reales y evaluar la precisión de cada enfoque.

1. Propuesta 1: clasificación mediante simulación con filtraje de ruido

En este primer enfoque, siguiendo lo expuesto por Adeli et al. (2018), se propone simular las variables cuantitativas, condicionalmente a los datos de estas mismas variables, luego clasificar las realizaciones obtenidas de modo de convertirlas en realizaciones de la variable categórica. La clasificación se realiza por un método de aprendizaje automático, el que se ajusta a partir de la información de muestreo disponible. Una novedad con respecto a la propuesta de Adeli et al. (2018) es la eliminación de la variabilidad de pequeña escala, asociada a errores de medición o ruido (efecto pepita) en la simulación de las variables cuantitativas. La validación de la propuesta se realiza con la técnica de la subdivisión de la muestra (*jack-knife*), de modo de poder evaluar la capacidad predictiva del modelo (porcentaje de acierto en la predicción de la variable categórica) en sitios con datos que no han sido utilizados en el ajuste del modelo o en el condicionamiento de las simulaciones.

A continuación, se detallan los pasos a seguir para la implementación práctica de la propuesta.

1.1. Transformación de las variables continuas a valores gaussianos

Las concentraciones geoquímicas son visualizadas como la realización de funciones aleatorias estacionarias. Para aplicar la simulación multigaussiana, es necesario transformar cada una de estas concentraciones en una distribución gaussiana, tal como se mencionó en el capítulo 2 en la sección 3.5.1.1. Esta transformación refleja un cambio de unidades de los datos con una media igual a 0 y la varianza igual a 1 (Chilès and Delfiner, 2012).

El tener varias decenas de concentraciones geoquímicas en cada uno de los casos de estudio impide utilizar técnicas de co-anamorfosis o anamorfosis multivariables (Leuangthong and Deutsch, 2003; Barnett et al., 2013; van den Boogaart et al., 2017), las cuales en la práctica son aplicables a un número reducido de variables. La razón principal es que, mientras mayor el número de variables, mayor es la cantidad de datos necesaria para poder inferir la distribución multivariable con un cierto grado de precisión. Por ejemplo, si unos 50 datos son suficientes para aproximar experimentalmente una distribución univariable, se necesitarían 2500 para aproximar una distribución bivariable, y 125000 para una distribución trivariable; en este contexto, la cantidad de

datos se vuelve prohibitiva cuando se busca modelar conjuntamente varias decenas de variables. De este modo, la transformación gaussiana se realizará considerando cada variable por separado.

1.2. Análisis variográfico

El siguiente paso es la identificación de la estructura de correlación espacial de las variables gaussianas (transformadas). Para conocer si estas variables tienen una geometría de isotropía o anisotropía y para identificar sus direcciones principales de continuidad, se puede elaborar mapas variográficos y/o calcular variogramas experimentales a lo largo de varias direcciones del espacio. Los variogramas experimentales en las direcciones principales (de mayor y de menor continuidad, respectivamente) son modelados por un modelo lineal de corregionalización, con ayuda de un algoritmo de ajuste semiautomático de las mesetas (Goulard y Voltz, 1992), resultante en una combinación de estructuras anidadas. Cada una de estas estructuras está asociada a un alcance de correlación representativo de una cierta escala de variación (Journel and Huijbregts, 1978). En particular, el efecto pepita corresponde a variaciones de muy pequeña escala, asimilables a un "ruido" que, en general, no tiene relación con los procesos físicos subyacentes en los entornos geológicos-mineros, sino que con errores de medición (Chilès and Delfiner, 2012), las cuales pueden perjudicar la clasificación de las variables cuantitativas (concentraciones geoquímicas) en categorías geológicas (litología u otra característica).

1.3. Filtrado de ruido y simulación condicional

Para minimizar el impacto de estos tipos de ruido en la clasificación regionalizada, se realiza un filtrado del efecto pepita al simular las funciones aleatorias gaussianas (transformaciones de las concentraciones geoquímicas) en base a los siguientes pasos:

a. El modelo de corregionalización permite descomponer las funciones aleatorias gaussianas en factores independientes, utilizando el llamado análisis de corregionalización o análisis de factores geoestadísticos (Goovaerts, 1992; Wackernagel, 2003). Tal factorización es uno de los enfoques utilizados para cosimular funciones aleatorias correlacionadas, ver ecuaciones 2.24 y 2.26 en el capítulo 2.
b. Cada uno de los factores es simulado en las ubicaciones de los datos de entrenamiento, así como en las ubicaciones de los datos de prueba, utilizando el algoritmo de bandas rotantes. Este algoritmo es preferido sobre otras alternativas como los métodos de descomposición matricial o secuencial, por su aplicabilidad a la simulación altamente multivariable, su baja complejidad numérica, su requisito mínimo de almacenamiento y su exactitud, pudiendo reproducir la estructura de correlación espacial de cada factor sin aproximación.
c. En cada ubicación de datos de entrenamiento, los factores simulados son combinados para reconstruir las funciones aleatorias gaussianas de interés, luego los "residuos" (es decir, las diferencias entre los valores simulados así reconstruidos y los valores gaussianos resultantes de la anamorfosis gaussiana) son calculados.
d. En las ubicaciones de los datos de entrenamiento y de prueba, los factores asociados con las estructuras no pepíticas son combinados para reconstruir variables filtradas (sin ruido).
e. Las variables filtradas son condicionadas a los datos de entrenamiento, sumando los valores simulados obtenidos en el paso *d* y el cokriging de los residuos calculados en el paso *c*. En comparación con la etapa de condicionamiento tradicional (Chilès y Delfiner, 2012), la

diferencia se encuentra en el miembro derecho del sistema de ecuaciones de cokriging, donde el efecto pepita es removido y las demás estructuras anidadas son mantenidas.

Este procedimiento lleva a la construcción de realizaciones de las funciones aleatorias gaussianas, filtradas del efecto pepita y condicionadas a los datos de entrenamiento. Estas realizaciones son construidas tanto en los subconjuntos de entrenamiento como en los de prueba. Se construye un gran número de realizaciones (300), destinado a producir análisis estadísticamente robustos de los resultados de la clasificación.

A modo de comparación, el procedimiento de simulación es también realizado sin filtrar el efecto pepita, es decir, se realiza una cosimulación tradicional de las funciones aleatorias gaussianas.

1.4. Clasificación supervisada

Se utilizan técnicas de aprendizaje automático (machine learning) para clasificar las realizaciones gaussianas (con y sin filtrado del efecto pepita) y así obtener realizaciones de la variable categórica a predecir. Se ponen a prueba varios algoritmos de clasificación (análisis discriminante lineal, máquinas de vectores de soporte y árboles de decisión) y se selecciona el algoritmo más eficaz (el que conlleva la mejor tasa de acierto) en base a los datos de entrenamiento disponibles.

Para la clasificación dos estrategias de clasificación son consideradas:

- utilizando como conjunto de entrenamiento el promedio de los 300 valores simulados en cada ubicación de los datos de entrenamiento, lo que representa un escenario promedio (esperado);
- repitiendo la clasificación 300 veces, obteniendo tantas clasificaciones como haya escenarios de las variables geoquímicas, utilizando en cada repetición los valores simulados de un solo escenario como conjunto de entrenamiento; luego, con las 300 clasificaciones se determina la "clasificación más probable" (es decir, la categoría que se repite más entre las 300 clasificaciones) para cada uno de los datos de entrenamiento.

Para cada estrategia, la precisión de la clasificación se cuantifica por el porcentaje de datos clasificados correctamente en una matriz de confusión. Esto es aplicado para las realizaciones con filtrado de ruido, así como para las realizaciones tradicionales sin filtrado.

Finalmente, para predecir la litología en los sectores en el cual se desconoce la información litológica y geoquímica, se aplican los clasificadores ajustados en el subconjunto de los datos de entrenamiento a los valores simulados (con o sin ruido) en ubicaciones sin datos que cubren la región de estudio.

En el capítulo 4, se indica el desarrollo de esta estrategia a un conjunto de datos de sondajes de exploración en un yacimiento de tipo pórfido cuprífero. Esta misma estrategia ha sido aplicada a una base de datos de prospección geoquímica y presentada en un artículo titulado "*Regionalized classification of geochemical data with filtering of measurement noises for predictive lithological mapping*" y aceptado para publicación en la revista *Natural Resources Research* (Anexo A).

2. Propuesta 2: modelamiento directo de las variables categóricas con modelo mixto pluri-multigaussiano

Para esta propuesta se compara dos enfoques geoestadísticos para la predicción espacial de clases litológicas, basado en información litológica y geoquímica disponible en datos de muestreo. El primer enfoque consiste en un cokriging de los indicadores de las clases litológicas, mientras que el segundo enfoque, que se constituye como la principal novedad de esta propuesta, se basa en una simulación plurigaussiana de las clases litológicas y su extensión para incluir las covariables geoquímicas, permitiendo el cálculo de las probabilidades de ocurrencia de las clases litológicas y la predicción de concentraciones geoquímicas en cualquier sitio no muestreado.

2.1. Cokriging de indicadores

Los indicadores predichos (generalmente, con valores entre 0 y 1) pueden ser interpretados como estimaciones de las probabilidades de ocurrencia de las clases litológicas (Solow, 1986; Goovaerts, 1997). La clasificación final puede ser tomada como aquella en el cual el indicador predicho es el mayor, equivalente a considerar como la litología más probable (Goovaerts, 1997; Kasmaee et al., 2019; Talebi et al., 2019b).

Para el desarrollo de este enfoque, los tipos de litologías son codificados en variables indicadoras de tipo binario. Se determina la estructura espacial de los indicadores mediante el cálculo de sus variogramas experimentales y el ajuste de un modelo lineal de corregionalización. Luego, en cada ubicación de los datos de prueba, se predice los indicadores mediante cokriging ordinario, utilizando una vecindad móvil para incorporar las muestras de entrenamiento adyacentes. Finalmente, se selecciona las clases litológicas para la cual el indicador predicho es el mayor. Se repite el proceso incorporando a las covariables geoquímicas en el modelo de corregionalización y en el cokriging, con el objetivo de mejorar las predicciones de los indicadores.

2.2. Simulación multigaussiana - plurigaussiana

En primer lugar, se ajusta un modelo plurigaussiano que permite simular la variable categórica en el espacio, condicionalmente a los datos del subconjunto de entrenamiento. Con las realizaciones así construidas, se determina las probabilidades de ocurrencia de los tipos de litologías en los sitios de datos de prueba. En segundo lugar, se enriquece el modelo anterior al incorporar las variables cuantitativas como covariables y al diseñar un modelo mixto plurigaussiano – multigaussiano que permita cosimular los tipos de litologías y las concentraciones geoquímicas.

2.2.1. Modelo plurigaussiano

Para conseguir los resultados esperados se realiza los siguientes pasos:

a. Se identifica una regla de truncamiento en base a la interpretación de los contactos entre las clases litológicas y sus cronologías, lo que permite asociar las clases litológicas a un conjunto de funciones aleatorias gaussianas latentes $Y_1, \ldots Y_N$.

b. Los umbrales de truncamiento son calculados en base a las proporciones de cada clase litológica en la zona de estudio, lo que permite determinar los valores numéricos utilizados en la regla de truncamiento.
c. Las clases litológicas son codificadas en indicadores de las funciones aleatorias gaussianas subyacentes. El indicador puede ser desconocido si la función aleatoria gaussiana no está usada en la definición de la clase; de lo contrario es 0 o 1.
d. Se calculan las covarianzas directas y cruzadas de los indicadores y, siguiendo el enfoque de Emery and Cornejo (2010) se convierte estas covarianzas experimentales de indicadores en covarianzas experimentales de las funciones aleatorias gaussianas, a las cuales se ajusta un modelo lineal de corregionalización.
e. Se simulan las funciones aleatorias gaussianas en los datos de prueba, condicionalmente a los datos de entrenamiento adyacentes, mediante el muestreador de Gibbs (Armstrong et al., 2011). Se generan trescientas realizaciones en cada ubicación de prueba.
f. Se realiza la predicción de la litología en la ubicación de los datos de prueba, considerando la clase más probable en cada sitio (clase que más se repite en las 300 realizaciones).

2.2.2. Incorporación de concentraciones geoquímicas como covariables

El desarrollo se realiza mediante los siguientes pasos:

a. Anamorfosis gaussiana: las concentraciones geoquímicas en los datos de entrenamiento son transformadas en datos gaussianos.
b. Análisis estructural: se ajusta un modelo lineal de corregionalización de las funciones aleatorias gaussianas asociadas a las clases litológicas y a las concentraciones geoquímicas. Dado el gran número de variables, el ajuste de un modelo de corregionalización completo es laborioso, incluso con el recurso a algoritmos de ajuste semiautomático, por lo cual se opta por un modelo parsimonioso, que involucra a las variables geoquímicas que presentan la mayor correlación con las variables gaussianas asociadas a la litología.
c. Predicción en las ubicaciones de datos de prueba: las funciones aleatorias gaussianas son simuladas en cada ubicación de los datos de prueba, siguiendo el mismo esquema que para modelo plurigaussiano, excepto que aquí también se consideran las funciones aleatorias gaussianas asociadas con las covariables geoquímicas. Se generan 300 realizaciones y la clase litológica más probable es utilizada para la predicción en los sitios de los datos de prueba.

Capitulo 4. Estudio de caso para un depósito de pórfido de cobre

1. Antecedentes

La construcción de modelos geológicos con la finalidad de comprender la constitución de los depósitos y mejorar la exploración y planificación en el desarrollo de la extracción de los recursos minerales ha llevado a desarrollar metodologías para predecir las características geológicas y las propiedades de las rocas del subsuelo (Adeli and Emery, 2020). Dichas metodologías se basan en la información geológica y geoquímica obtenida del muestreo y logueo de testigos de sondajes, así como en el uso de técnicas estadísticas, geoestadísticas y/o de aprendizaje automático, permitiendo representar los tipos de roca, alteración y mineralización en el depósito de interés (Duke and Hanna, 2001; Sinclair and Blackwell, 2002; Rossi and Deutsch, 2014; Chanderman et al., 2017; Fouedjio et al., 2018; Emery and Séguret, 2020).

Los errores en las mediciones de variables cuantitativas (tales como los ensayos de leyes) pueden ser filtrados en la construcción de modelos predictivos, pudiendo mejorar el uso de estos modelos en la definición de los tipos de rocas, alteraciones y zonas minerales mediante una clasificación supervisada. Es así como la primera propuesta metodológica indicada en el capítulo 3 es aplicada aquí a un depósito de pórfido de cobre, cuyo nombre y ubicación quedan en reserva por razones de confidencialidad de la información.

Los datos de muestreo disponibles consisten en el logueo y en ensayos de sondajes de diamantina (Fig. 4.1), obteniendo una base de 42562 datos con 25 concentraciones geoquímicas, además de tres variables categóricas (tipos de roca, tipos de alteraciones y zonas minerales) (Tablas 4.1 y 4.2).

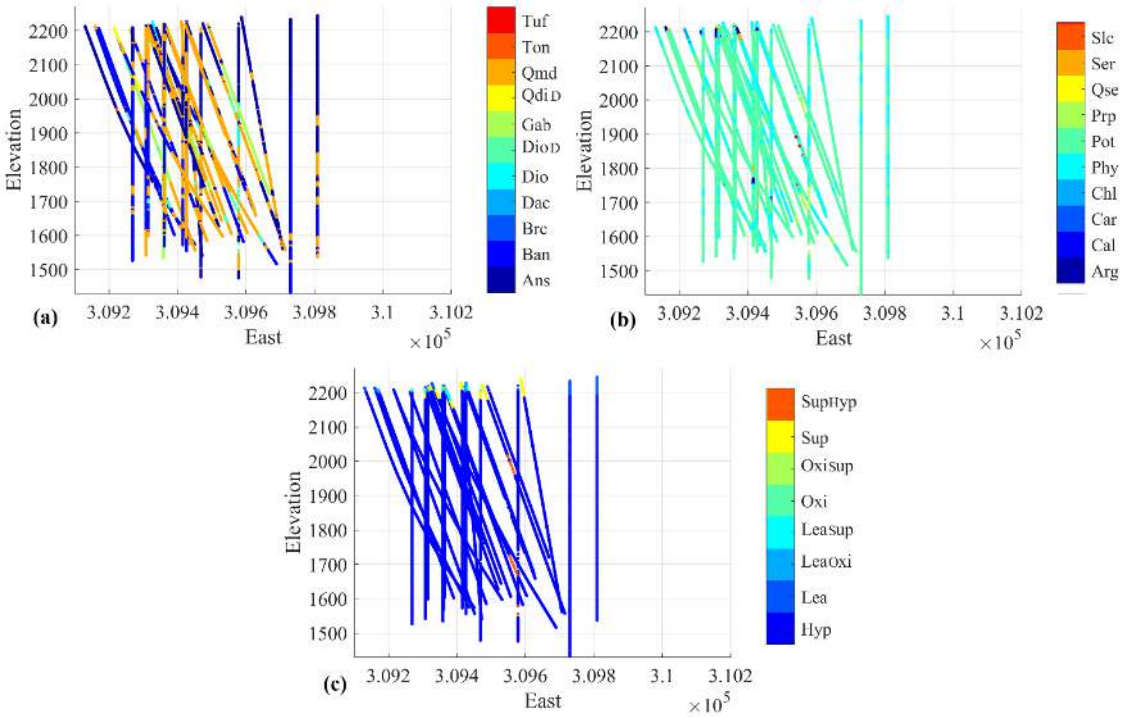

Figura 4.1. Representación espacial de las perforaciones en el depósito de pórfido de cobre con cada tipo de roca (a), alteración (b) y zona mineral (c). Por razones de confidencialidad, sólo se muestra una sección vertical del depósito. Las coordenadas locales (este, norte, vertical) están expresadas en metros

Tabla 4.1. Concentración de 25 elementos químicos (ppm) con su definición y codificación

Variable	Símbolo	Variable	Símbolo	Variable	Símbolo	Variable	Símbolo
Plata	Ag	Cromo	Cr	Manganeso	Mn	Antimonio	Sb
Aluminio	Al	Cobre	Cu	Molibdeno	Mo	Escandio	Sc
Arsénico	As	Hierro	Fe	Níquel	Ni	Torio	Th
Calcio	Ca	Lantano	La	Fosforo	P	Vanadio	V
Cadmio	Cd	Litio	Li	Plomo	Pb	Yterbio	Yb
Cerio	Ce	Magnesio	Mg	Azufre	S	Zinc	Zn
Cobalto	Co						

Tabla 4.2. Información de tipo de roca, tipo de alteración y zona mineral para el depósito de pórfido de cobre

Roca		Alteración		Zona mineral	
Tipo	**Número de datos**	**Tipo**	**Número de datos**	**Tipo**	**Número de datos**
Andesita (Ans)	11763	Argílica (Arg)	173	Hipógeno (Hyp)	39280
Basalto (Ban)	7419	Calcosilicatada (Cal)	32	Lixiviado (Lea)	651
Brecha (Brc)	3	Carbonatada (Car)	4	Mixto lixiviado/óxido (Lea_Oxi)	208
Dacita (Dac)	203	Clorítica (Chl)	76	Mixto lixiviado/supérgeno (Lea_Sup)	123
Diorita (Dio)	550	Fílica (Phy)	12425	Óxido (Oxi)	174
Dique dioritico (Dio_D)	1027	Potásica (Pot)	28537	Mixto óxido/supérgeno (Oxi_Sup)	201
Gabro (Gab)	2031	Propilítica (Prp)	813	Supérgeno (Sup)	1407
Dique Cuarzo Dioritico (Qdi_D)	212	Cuarzo sericita (Qse)	2	Mixto supérgeno/hipógeno (Sup_Hyp)	108
Cuarzo monzodiorítico (Qmd)	19339	Sericita (Ser)	380	Sin información	410
Tonalita (Ton)	12	Silicatada (Slc)	7		
Toba (Tuf)	3	Sin información	113		
Total	***42562***	***Total***	***42562***	***Total***	***42562***

2. Análisis estadístico de las concentraciones geoquímicas

Las concentraciones geoquímicas tienen una distribución asimétrica positiva, con valores altos en los elementos: Aluminio, Calcio, Cobre, Hierro, Magnesio, Manganeso, Fosforo, Azufre (Fig. 4.2, Tabla 4.3). Las correlaciones entre concentraciones indican cuatro grupos marcados que tienen correlación positiva, los mismos que son indicados a través de un circulo de correlación (Fig. 4.3); entre Cr, Al, Sc, V, Mg, Li; entre La, Ce, Th; entre Cu, Mo, asociado a Zn, Ag, Mn, S, As, Pb, Cd. Estos grupos tienen afinidad química entre los elementos que lo integran, ya sea por sus propiedades químicas (Robb, 2011) o transición de metales. El grupo Cu-Mo es parte de las mineralizaciones de los depósitos de pórfido de cobre con asociaciones de depósitos de Pb, Zn, Ag.

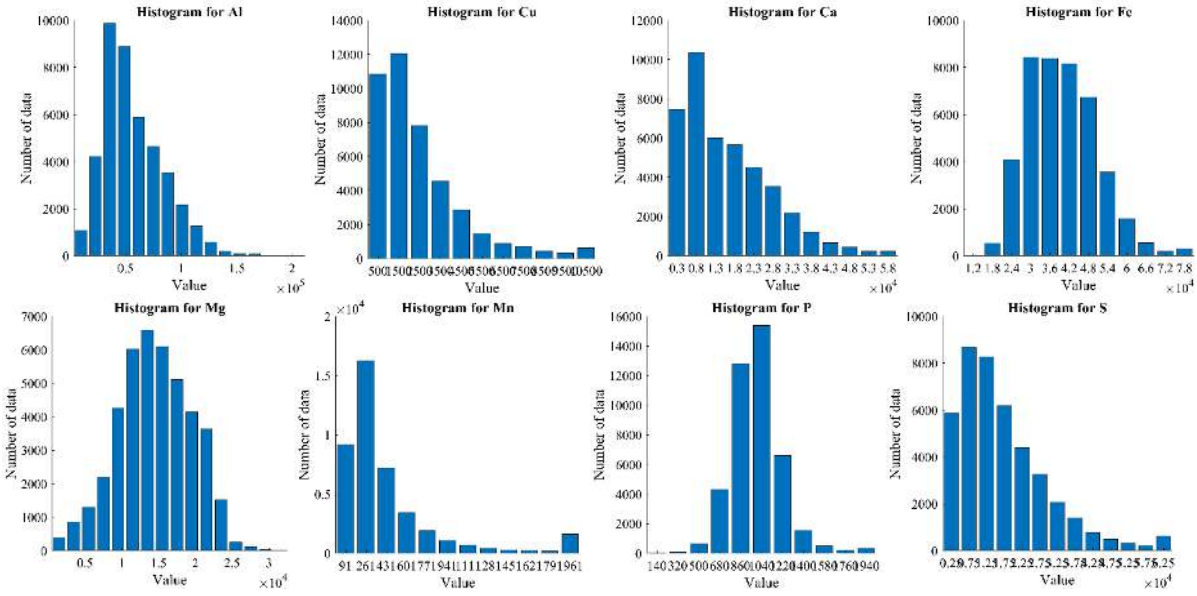

Figura 4.2. Grupo de elementos con distribución de frecuencia asimétrica positiva y altas concentraciones.

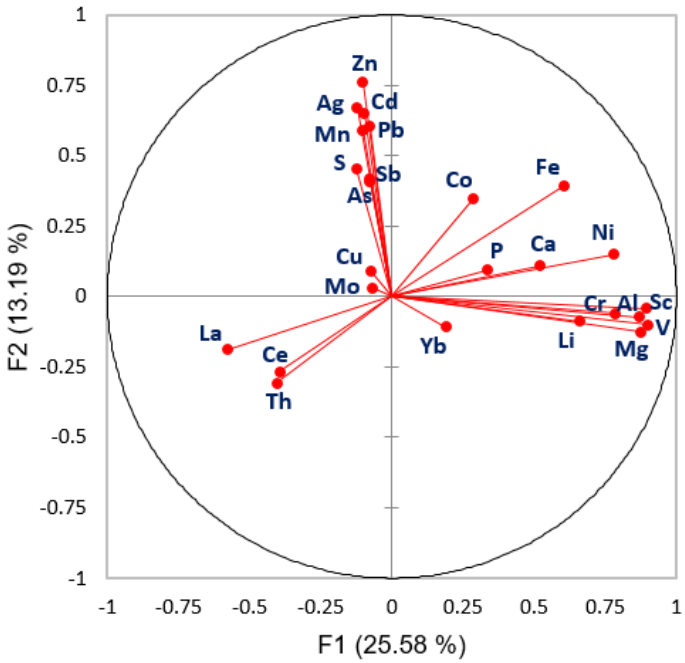

Figura 4.3. Circulo correlación geoquímica en el primer plano factorial de PCA; los ángulos entre vectores indican correlaciones positivas (ángulos agudos) y negativas (ángulos obtusos).

Tabla 4.3. Resumen estadístico para las 25 concentraciones geoquímicas

	Ag	Al	As	Ca	Cd	Ce	Co	Cr	Cu
Media	0.56	56750.17	8.09	16135.96	0.45	23.47	20.82	27.52	2493.90
Mediana	0.34	50575.00	2.40	13297.00	0.24	22.00	19.00	19.00	1842.00
Moda	0.27	38329.00	2.30	4038.00	0.24	19.00	16.00	8.00	728.00
Desv Stand	1.33	27399.65	47.13	11629.38	1.98	10.02	11.30	26.35	2608.80
Mínimo	0.11	3309.00	1.70	501.00	0.14	2.00	2.00	0.00	3.00
Máximo	49.30	229834.00	4777.50	145997.00	194.60	130.00	313.00	367.00	121135.00
Rango	49.19	226525.00	4775.80	145496.00	194.46	128.00	311.00	367.00	121132.00
Datos	42562	42562	42562	42562	42562	42562	42562	42562	42562

	Fe	La	Li	Mg	Mn	Mo	Ni	P	Pb
Media	39801.27	16.38	7.39	14475.38	523.96	15.75	21.27	999.62	47.77
Mediana	38867.00	15.00	7.00	14394.00	290.00	7.30	19.00	986.00	11.00
Moda	34041.00	11.00	6.00	14939.00	172.00	5.30	13.00	1000.00	7.00
Desv Stand	11226.91	6.55	4.22	4993.80	1014.83	37.34	9.82	243.82	488.80
Mínimo	9291.00	2.00	0.00	508.00	6.00	0.50	2.00	61.00	4.00
Máximo	150267.00	123.00	274.00	32086.00	22022.00	2161.90	178.00	7194.00	45664.00
Rango	140976.00	121.00	274.00	31578.00	22016.00	2161.40	176.00	7133.00	45660.00
Datos	42562	42562	42562	42562	42562	42562	42562	42562	42562

	S	Sb	Sc	Th	V	Yb	Zn
Media	17179.61	1.58	11.20	8.02	106.06	1.36	120.37
Mediana	13959.00	1.02	8.50	7.80	93.00	0.90	56.00
Moda	50.00	1.00	4.30	3.80	68.00	0.60	44.00
Desv Stand	14187.26	11.19	8.57	3.84	66.77	1.80	523.92
Mínimo	43.00	0.65	0.60	2.30	4.00	0.20	15.00
Máximo	205136.00	619.90	40.80	76.70	348.00	28.00	27830.00
Rango	205093.00	619.25	40.20	74.40	344.00	27.80	27815.00
Datos	42562	42562	42562	42562	42562	42562	42562

3. División de la muestra para conjunto de entrenamiento y de prueba

Los 42562 datos correspondientes al conjunto de datos originales son divididos aleatoriamente en dos subconjuntos (Fig. 4.4): un subconjunto de 25477 muestras para entrenamiento, que será utilizado para construir los modelos geoestadísticos y la clasificación; y, un subconjunto de 17085 muestras para prueba, en este subconjunto se pronosticará los tipos de roca, alteraciones y zonas minerales a partir de la información de entrenamiento.

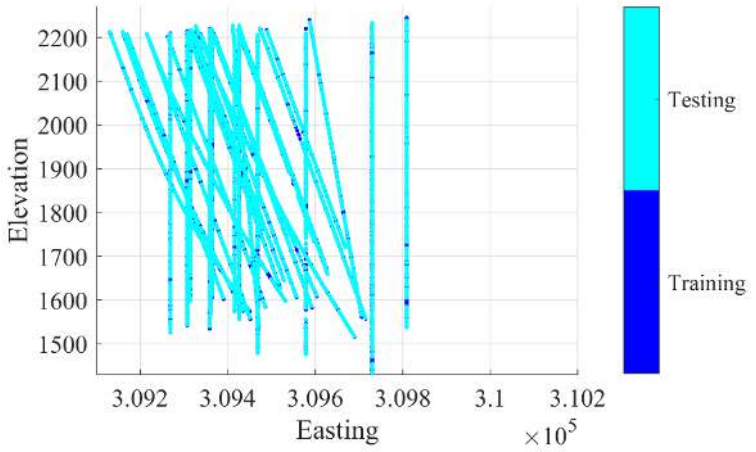

Figura 4.4. Representación espacial de la división aleatoria para las muestras de las perforaciones en un subconjunto de entrenamiento (25477 datos) y un subconjunto de prueba (17085 datos)

4. Análisis variográfico

Los datos de concentraciones geoquímicas son transformados a datos gaussianos con la finalidad de condicionar las simulaciones gaussianas que serán construidas. Los mapas variográficos (Fig. 4.5) no indican una marcada anisotropía en el plano horizontal, por lo que se opta por calcular variogramas experimentales (25 directos y 300 cruzados) en el plano horizontal y la dirección vertical, con los parámetros de la Tabla 4.4.

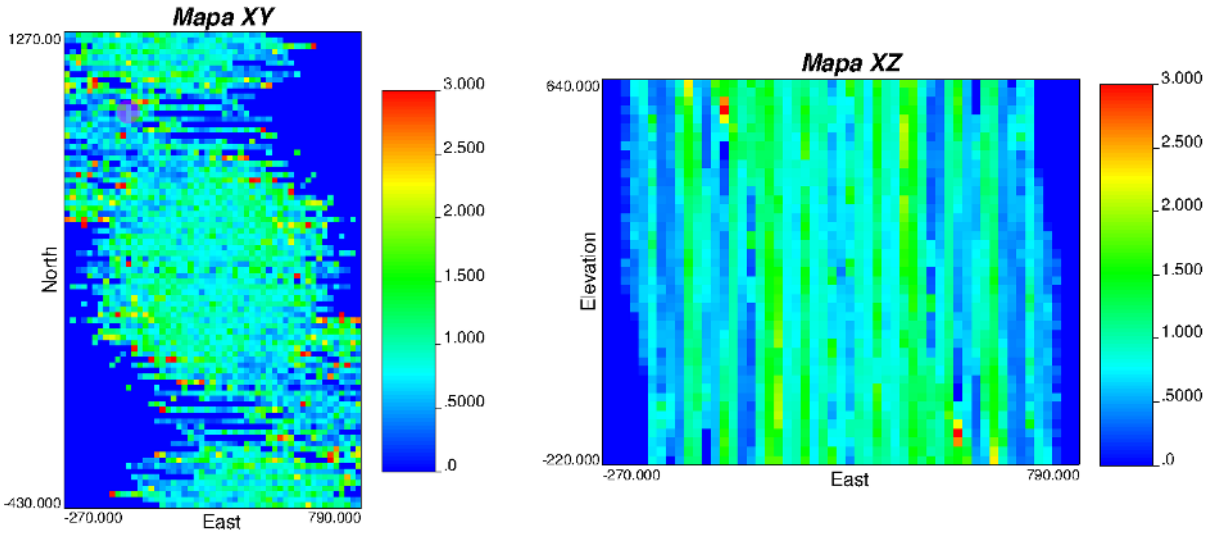

Figura 4.5. Ejemplo de mapa variográfico para los datos transformados de plata, en el cual no se evidencia una anisotropía en las direcciones horizontales

Tabla 4.4. Parámetros para los variogramas experimentales en las direcciones horizontal y vertical

Dirección	Horizontal	Vertical
Azimut (°)	0.0	0.0
Tolerancia Azimut (°)	90.0	90.0
Dip (°)	0.0	90.0
Tolerancia dip (°)	20.0	20.0
Paso (m)	20.0	18.0
N° pasos	35	22
Tolerancia en paso (m)	10.0	9.0

Los variogramas experimentales son ajustados a un modelo teórico con cuatro estructuras anidadas (un efecto pepita y tres estructuras esféricas de alcances de correlación entre 50 m a 1000 m), con ayuda de un algoritmo de mínimos cuadrados para el ajuste de las mesetas (Tabla 4.5, Fig. 4.6). A modo de ilustración, los variogramas directos y cruzados representados en la Fig. 4.6 corresponden a elementos químicos que poseen una correlación lineal positiva y son denominados elementos raros característicos en los pórfidos de cobre.

Tabla 4.5. Estructuras anidadas utilizadas para ajustar los variogramas directos y cruzados

Estructura anidada	Tipo de estructura	Alcance horizontal (m)	Alcance vertical (m)
1	Efecto pepita	0	0
2	Esférica	50	70
3	Esférica	120	400
4	Esférica	250	1000

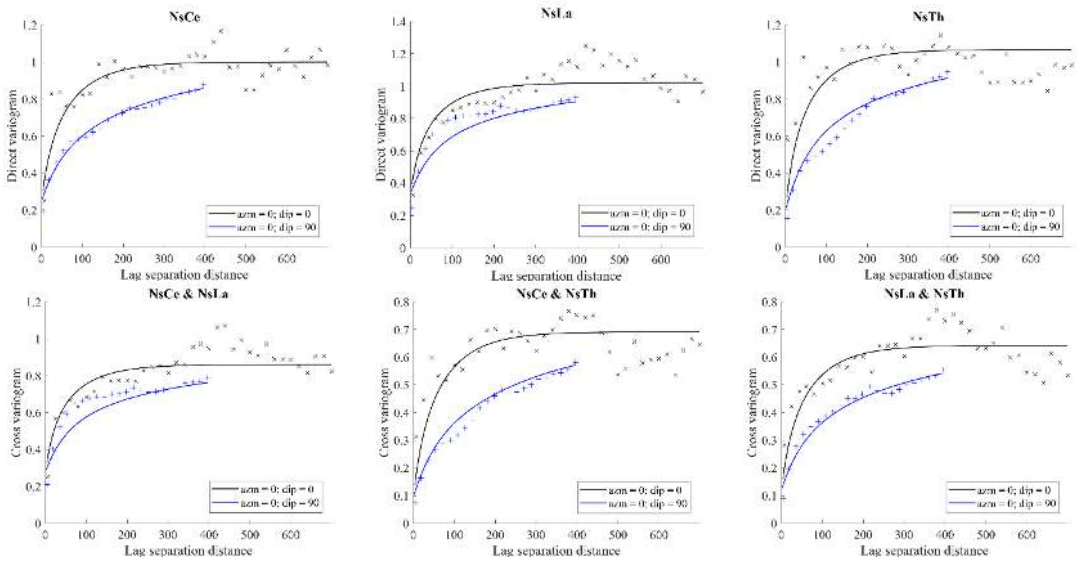

Figura 4.6. Variogramas experimentales directos y cruzados (cruces) y variogramas teóricos (líneas solidas), para las variables La, Ce, Th

5. Simulación condicional

Las funciones aleatorias gaussianas asociadas a las concentraciones geoquímicas son simuladas en los sitios de prueba, filtrando el efecto pepita que está asociado con ruido o errores de medición. La simulación no condicional (*SNC*) filtrada (*fil*) en un sitio x, ya sea de entrenamiento o de prueba, se escribe como:

$$\boldsymbol{Y}_{fil}^{SNC}(x) = \sum_{s=2}^{4} \mathbf{A}_s \boldsymbol{W}_s^{SNC}(x) \tag{4.1}$$

donde $\boldsymbol{W}_s$ representan los factores del análisis factorial geoestadístico asociados a las estructuras anidadas esféricas (con índices s = 2, 3 y 4), ver las ecuaciones 2.24 y 2.26 en el capítulo 2. La simulación de estos factores se realiza mediante bandas rotantes (Emery, 2008).

Asimismo, se simulan las mismas funciones aleatorias gaussianas en los sitios de entrenamiento (x_1, …, x_n), pero esta vez con el efecto pepita:

$$\boldsymbol{Y}^{SNC}(x_\alpha) = \sum_{s=1}^{4} \mathbf{A}_s \boldsymbol{W}_s^{SNC}(x_\alpha) \tag{4.2}$$

Lo anterior permite calcular los "residuos" de la simulación (diferencia entre valores simulados y valores de los datos transformados) en estos sitios de entrenamiento, los cuales son usados para el condicionamiento de las simulaciones. La simulación condicional (*SC*) es obtenida por la adición de la simulación no condicional y del cokriging de los residuos:

$$\boldsymbol{Y}_{fil}^{SC}(x) = \boldsymbol{Y}_{fil}^{SNC}(x) + \sum_{\alpha=1}^{n} \boldsymbol{\Lambda}_\alpha^T(x)[\boldsymbol{Y}(x_\alpha) - \boldsymbol{Y}^{SNC}(x_\alpha)] \tag{4.3}$$

con

$$\begin{pmatrix} \boldsymbol{C}(x_1 - x_1) & \cdots & \boldsymbol{C}(x_1 - x_n) \\ \vdots & \ddots & \vdots \\ \boldsymbol{C}(x_n - x_1) & \cdots & \boldsymbol{C}(x_n - x_n) \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda}_1(x) \\ \vdots \\ \boldsymbol{\Lambda}_n(x) \end{pmatrix} = \begin{pmatrix} \boldsymbol{C}_{fil}(x_1 - x) \\ \vdots \\ \boldsymbol{C}_{fil}(x_n - x) \end{pmatrix} \tag{4.4}$$

y

$$\boldsymbol{C}_{fil}(h) = \sum_{s=2}^{4} \mathbf{B}_s \, \rho_s(h) \tag{4.5}$$

En la ecuación 4.5, ρ_s (con s = 2, 3 y 4) denotan las estructuras esféricas del modelo lineal de corregionalización y $\mathbf{B}_s$ son las respectivas matrices de corregionalización. La diferencia con la metodología tradicional de condicionamiento (ecuación 2.27 en el capítulo 2) radica en el hecho que el sistema de cokriging (4.4) no considera el efecto pepita en el miembro derecho, pero sí lo considera en el miembro izquierdo; este último punto se debe a que los residuos calculados en los sitios de entrenamiento no están filtrados del efecto pepita.

Se construye un total de 300 realizaciones de las funciones aleatorias gaussianas asociadas a las concentraciones geoquímicas, tanto en los sitios de entrenamiento como en los sitios de prueba. Las simulaciones en los primeros sitios permitirán entrenar los algoritmos de clasificación, los cuales serán posteriormente aplicados a los sitios de prueba. Para evaluar el impacto del ruido en

los resultados de la clasificación, se construyen otras 300 realizaciones sin filtrar el efecto pepita, es decir, sin eliminar la componente pepítica en el sistema de cokriging (4.4).

6. Clasificación utilizando arboles de decisión

Siguiendo la metodología señalada en la sección 1.4 del capítulo 3, se utiliza un modelo de Árbol de decisión (Mitchell, 1997) con el método de división CHAID (Chi-square automatic interaction detectors) (Kass, 1980; Ibarguren et al., 2016), para obtener las predicciones de las tres variables categóricas de interés para el modelo geológico: tipo de roca, tipo de alteración y zona mineral. En cada caso, el árbol de decisión es entrenado con los valores simulados en el subconjunto de entrenamiento, luego aplicado a los valores simulados en el subconjunto de prueba, desembocando en 300 realizaciones de cada variable categórica obtenidas con filtrado del efecto pepita y otras 300 obtenidas sin filtrado.

La Tabla 4.6 indica las tasas de acierto de la clasificación (con dos estrategias, una basada en el promedio de las 300 realizaciones, otra basada en la clasificación más probable, ver sección 1.4 del capítulo 3), tanto en los sitios de entrenamiento, donde se han medido las concentraciones geoquímicas, como en los sitios de prueba, donde no se han medido.

Se desprende una mejora sistemática en la clasificación cuando se realiza el filtrado de ruido, tanto para la predicción del tipo de roca (hasta 9 puntos porcentuales), del tipo de alteración (6 puntos porcentuales) y zona mineral pasando (1 a 2 puntos porcentuales, con un acierto superior al 95%). La mejora se evidencia no solamente en los sitios de entrenamiento, en los cuales existen mediciones de las concentraciones geoquímicas, sino que también en los sitios de prueba, es decir, la metodología propuesta es capaz de mejorar la clasificación en situaciones de extrapolación. De las dos estrategias de clasificación, aquella basada en la clasificación más probable siempre logra resultados iguales o mejores que aquella basada en el promedio de las realizaciones, sobre todo en el tipo de roca (4 puntos porcentuales de diferencia entre ambas estrategias).

Tabla 4.6. Precisión en la predicción del tipo de roca, alteración y zona mineral utilizando la clasificación por arboles de decisión. Los valores corresponden a los subconjuntos de entrenamiento (25477 muestras) y de prueba (17085 muestras), filtrando o no el efecto pepita en la simulación de las funciones aleatorias asociadas a las concentraciones geoquímicas

Variable categórica	Estrategia de clasificación	Subconjunto de entrenamiento		Subconjunto de prueba	
		Sin ruido	Con ruido	Sin ruido	Con ruido
Tipo de roca	**Más probable**	85.93%	76.98%	74.76%	68.48%
	Promedio de realizaciones	81.72%	76.98%	70.66%	62.45%
Tipo de alteración	**Más probable**	88.52%	82.29%	81.69%	79.11%
	Promedio de realizaciones	88.50%	82.29%	80.32%	77.06%
Zona mineral	**Más probable**	98.64%	96.08%	96.42%	94.02%
	Promedio de realizaciones	97.46%	96.08%	95:29%	94.44%

6.1. Predicción del tipo de roca

6.1.1. Clasificación en el subconjunto de entrenamiento

Por una parte, se aplica el clasificador CHAID al promedio de las 300 realizaciones obtenidas en el subconjunto de entrenamiento, entregando una precisión del 81.72%. Por otra parte, utilizando el CHAID a los valores simulados de cada simulación por separado, se determina el tipo de roca más probable como aquel que más se repite en las 300 clasificaciones, logrando una precisión de 85.93% (Tabla 4.7; Fig. 4.7).

Tabla 4.7. Números de datos clasificados correctamente (línea diagonal) y datos clasificados incorrectamente (fuera de la diagonal) para cada roca observada (fila) y cada roca predicha (columna), para el subconjunto de datos de entrenamiento. Los números corresponden a los tipos de roca más probable sobre 300 realizaciones construidas filtrando ruido (efecto pepita)

Roca observada	Roca pronosticada (clasificación más probable)											Total	Precisión %
	Ans	Ban	Brc	Dac	Dio	Dio_D	Gab	Qdi_D	Qmd	Ton	Tuf		
Andesita (Ans)	**6114**	105	0	0	0	0	38	0	778	0	0	7035	86.90%
Basalto (Ban)	669	**3247**	0	0	0	0	28	0	513	0	0	4457	72.85%
Brecha (Brc)	0	0	**2**	0	0	0	0	0	1	0	0	3	66.66%
Dacita (Dac)	0	2	0	**110**	0	0	0	0	11	0	0	123	89.43%
Diorita (Dio)	45	0	0	0	**209**	0	0	0	82	0	0	336	62.20%
Dique dioritico (Dio_D)	44	3	0	0	0	**175**	4	0	383	0	0	609	28.73%
Gabro (Gab)	116	107	0	0	0	0	**755**	0	229	0	0	1207	62.55%
Diorita de cuarzo (Qdi_D)	11	2	0	0	0	0	0	**99**	12	0	0	124	79.84%
Cuarzo monzodiorítico (Qmd)	321	49	0	0	0	0	21	0	**11182**	0	0	11573	96.62%
Tonalita (Ton)	0	0	0	0	0	0	0	0	7	**0**	0	7	0.00%
Toba (Tuf)	2	1	0	0	0	0	0	0	0	0	**0**	3	0.00%
Total	**7322**	**3516**	**2**	**110**	**209**	**175**	**846**	**99**	**13198**	**0**	**0**	**25477**	**85.93%**
											Precisión de la Clasificación		**85.93%**

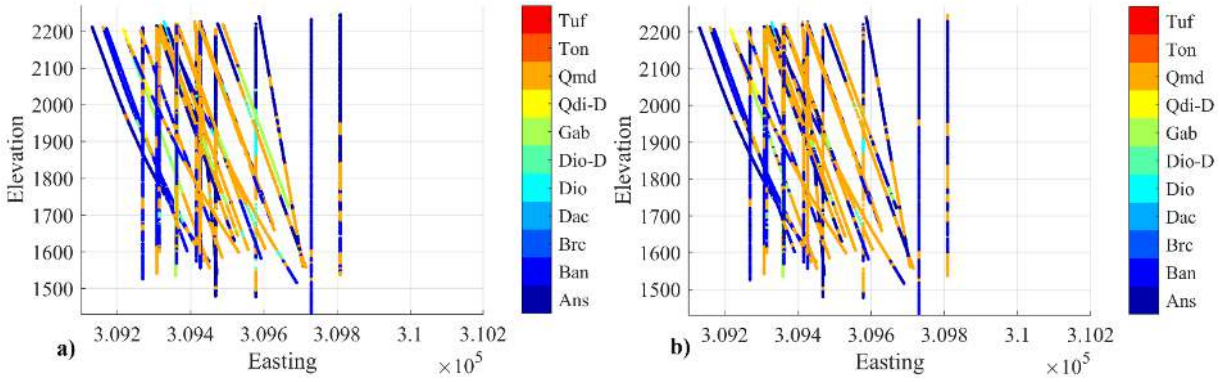

Figura 4.7. a) disposición espacial de los 11 tipos de rocas en las muestras de entrenamiento; **b)** disposición de los tipos de rocas luego de realizar la simulación con filtrado de ruido y la clasificación con CHAID.

6.1.2. Clasificación en el subconjunto de prueba

Debido a una mayor incertidumbre en los sitios donde están ubicados las muestras del subconjunto de prueba, en los cuales los valores de concentraciones geoquímicas no son conocidos al momento de clasificar, las puntuaciones de precisión bajan. El clasificador entrega una precisión de 68.48% sin realizar el filtrado de ruido; mientras que al filtrar el ruido la precisión sube a un 74.76% (Tabla 4.8). La roca que más mejora en la clasificación es el Cuarzo monzodiorítico Qmd, que es la roca predominante (Fig. 4.8).

Tabla 4.8. Números de datos clasificados correctamente (línea diagonal) y datos clasificados incorrectamente (fuera de la diagonal) para cada roca observada (fila) y cada roca predicha (columna), para el subconjunto de prueba. Los números corresponden a los tipos de roca más probable sobre 300 realizaciones construidas filtrando ruido.

Roca observada	Roca pronosticada (clasificación más probable)											Total	Precisión %
	Ans	Ban	Brc	Dac	Dio	Dio_D	Gab	Qdi_D	Qmd	Ton	Tuf		
Andesita (Ans)	**3780**	111	0	0	0	0	33	0	804	0	0	4728	79.94%
Basalto (Ban)	1007	**1416**	0	0	0	0	11	0	528	0	0	2962	47.80%
Brecha (Brc)	0	0	**0**	0	0	0	0	0	0	0	0	0	0.00%
Dacita (Dac)	10	1	0	**19**	0	0	0	0	50	0	0	80	23.75%
Diorita (Dio)	75	0	0	0	**38**	0	0	0	101	0	0	214	17.76%
Dique dioritico (Dio_D)	79	5	0	0	0	**28**	4	0	302	0	0	418	6.69%
Gabro (Gab)	164	143	0	0	0	0	**302**	0	215	0	0	824	36.65%
Diorita de cuarzo (Qdi_D)	9	1	0	0	0	0	0	**3**	75	0	0	88	0.02%
Cuarzo monzodiorítico (Qmd)	511	58	0	0	0	0	10	0	**7187**	0	0	7766	92.54%
Tonalita (Ton)	0	0	0	0	0	0	0	0	5	**0**	0	5	0.00%
Toba (Tuf)	0	0	0	0	0	0	0	0	0	0	**0**	0	0.00%
Total	**5635**	**1735**	**0**	**19**	**38**	**28**	**360**	**3**	**9267**	**0**	**0**	**17085**	**74.76%**
											Precisión de la clasificación		**74.76%**

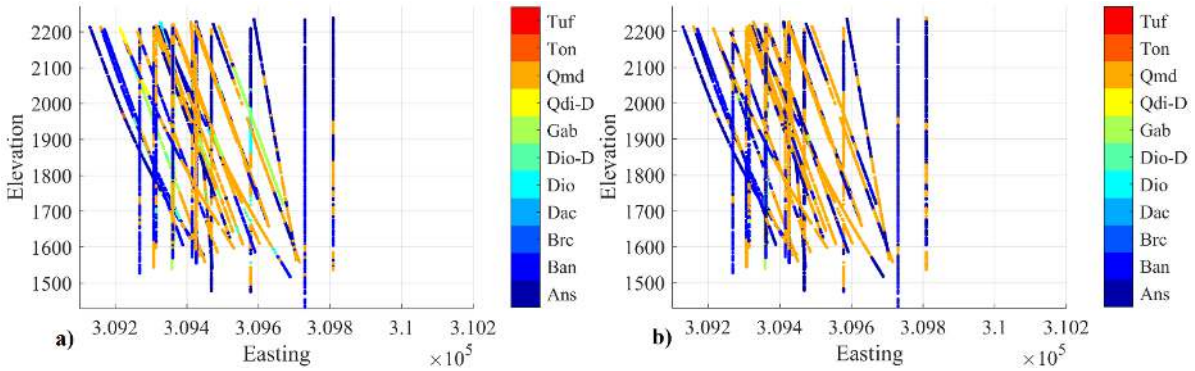

Figura 4.8. a) disposición espacial de los 11 tipos de rocas en las muestras de prueba; **b)** disposición de los tipos de rocas luego de realizar la simulación con filtrado de ruido y la clasificación con CHAID

6.2. Predicción del tipo de alteración

6.2.1. Clasificación en el subconjunto de entrenamiento

El tipo de alteración más probable en las 300 clasificaciones obtenidas por CHAID entrega una precisión de 88.52% (Tabla 4.9, Fig. 4.9), seis puntos porcentuales más comparado con la clasificación sin filtrar el ruido que da una precisión de 82.29%. El tipo de alteración que más mejora su puntuación corresponde a la alteración potásica, que es la más abundante.

Tabla 4.9. Números de datos clasificados correctamente (línea diagonal) y datos clasificados incorrectamente (fuera de la diagonal) para cada alteración observada (fila) y cada alteración predicha (columna), para el subconjunto de entrenamiento. Los números corresponden a los tipos de alteraciones más probable sobre 300 realizaciones construidas filtrando ruido.

Alteración observada	Predicción de la alteración (clasificación más probable)										Total	Precisión (%)
	Arg	Cal	Car	Chl	Phy	Pot	Prp	Qse	Ser	Slc		
Argílica (Arg)	**47**	0	0	0	5	49	0	0	0	0	101	46.53%
Calcosilicatada (Cal)	0	**17**	0	0	0	2	0	0	0	0	19	89.47%
Carbonatada (Car)	0	0	**0**	0	0	2	0	0	0	0	2	0.00%
Clorítica (Chl)	0	0	0	**16**	9	17	0	0	0	0	42	38.09%
Fílica (Phy)	0	0	0	0	**5587**	1831	0	0	0	0	7418	75.32%
Potásica (Pot)	0	0	0	0	614	**16504**	0	0	0	0	17118	96.41%
Propilítica (Prop)	0	0	0	0	57	178	**245**	0	0	0	480	51.04%
Cuarzo sericita (Qse)	0	0	0	0	0	2	0	**0**	0	0	2	0.00%
Sericita (Ser)	0	0	0	0	14	81	0	0	**131**	0	226	57.96%
Silicatada (Slc)	0	0	0	0	0	0	0	0	0	**4**	4	100.00%
Total	**47**	**17**	**0**	**16**	**6286**	**18666**	**245**	**0**	**131**	**4**	**25412**	**88.52%**
										Precisión de la clasificación		**88.52%**

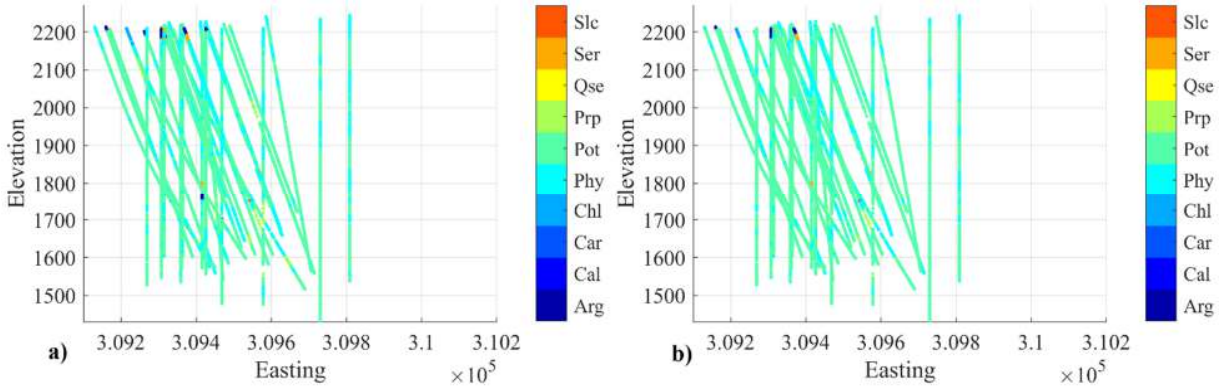

Figura 4.9. a) disposición espacial de los 10 tipos de alteraciones en las muestras de entrenamiento; **b)** disposición luego de realizar la simulación con filtrado de ruido y la clasificación con CHAID.

6.2.2. Clasificación en el subconjunto de prueba

Al extrapolar la clasificación al subconjunto de entrenamiento, la precisión disminuye a 79.11% (en base a simulaciones con ruido) y 81.69% (simulaciones con filtrado de ruido) (Tabla 4.10, Fig. 4.10). La alteración potásica es la que más mejora su puntuación.

Tabla 4.10. Números de datos clasificados correctamente (línea diagonal) y datos clasificados incorrectamente (fuera de la diagonal) para cada alteración observada (fila) y cada alteración predicha (columna), para el subconjunto de prueba. Los números corresponden a los tipos de alteraciones más probable sobre 300 realizaciones construidas filtrando ruido.

Alteración observada	Predicción de la alteración (clasificación más probable)										Total	Precisión (%)
	Arg	Cal	Car	Chl	Phy	Pot	Prp	Qse	Ser	Slc		
Argílica (Arg)	**0**	0	0	0	9	63	0	0	0	0	72	0.00%
Calcosilicatada (Cal)	0	**0**	0	0	0	13	0	0	0	0	13	0.00%
Carbonatada (Car)	0	0	**0**	0	0	2	0	0	0	0	2	0.00%
Clorítica (Chl)	0	0	0	**0**	20	14	0	0	0	0	34	0.00%
Fílica (Phy)	0	0	0	0	**2991**	2016	0	0	0	0	5007	59.74%
Potásica (Pot)	0	0	0	0	539	**10880**	0	0	0	0	11419	95.28%
Propilítica (Prop)	0	0	0	0	60	237	**36**	0	0	0	333	10.81%
Cuarzo sericita (Qse)	0	0	0	0	0	0	0	**0**	0	0	0	0.00%
Sericita (Ser)	0	0	0	0	29	114	0	0	**11**	0	154	7.14%
Silicatada (Slc)	0	0	0	0	3	0	0	0	0	**0**	3	0.00%
Total	**0**	**0**	**0**	**0**	**3651**	**13339**	**36**	**0**	**11**	**0**	**17037**	**81.69%**
							Precisión de la clasificación					**81.69%**

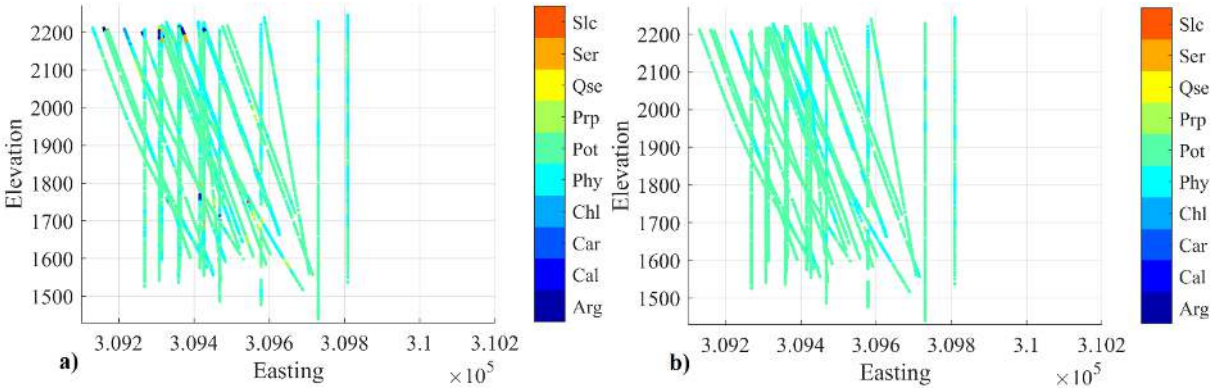

Figura 4.20. a) disposición espacial de los 10 tipos de alteraciones en las muestras de prueba; **b)**disposición luego de realizar la simulación con filtrado de ruido y la clasificación con CHAID.

6.3. Predicción de la zona de mineral

6.3.1. Clasificación en el subconjunto de entrenamiento

El método CHAID aplicado a las simulaciones con filtrado de ruido entrega una precisión de 98.64% (Tabla 4.11; Fig. 4.11) para la zona mineral más probable, comparado con el 96.08% de

precisión cuando el clasificador es aplicado a las simulaciones sin filtrar el ruido. La zona mineral hipógena es la que más mejora en su puntuación.

Tabla 4.11. Números de datos clasificados correctamente (línea diagonal) y datos clasificados incorrectamente (fuera de la diagonal) para cada zona mineral observada (fila) y cada zona mineral predicha (columna), para el subconjunto de entrenamiento. Los números corresponden a los tipos de alteraciones más probable sobre 300 realizaciones construidas filtrando ruido.

Zona mineral observada	Predicción de zona mineral (clasificación más probable)								Total	Precisión %
	Hyp	Lea	Lea_Oxi	Lea_Sup	Oxi	Oxi_Sup	Sup	Sup_Hyp		
Hipógeno (hyp)	**23558**	2	1	0	0	0	3	0	23564	99.97%
Lixiviado (Lea)	30	**357**	0	0	0	0	5	0	392	91.07%
Mixto lixiviado/óxido (Lea_Oxi)	20	0	**96**	0	0	0	10	0	126	76.19%
Mixto lixiviado/supérgeno (Lea_Sup)	0	0	0	**68**	0	0	6	0	74	91.89%
Óxido (Ox)	20	0	0	0	**83**	0	2	0	105	79.05%
Mixto óxido/supérgeno (Oxi_Su)	9	0	0	0	0	**109**	3	0	121	90.08%
Supérgeno (Su)	206	4	0	0	0	0	**632**	0	842	75.06%
Mixto supérgeno/hipógeno (Sup_Hy)	2	0	0	0	0	0	21	**41**	64	64.06%
Total	**23845**	**363**	**97**	**68**	**83**	**109**	**682**	**41**	**25288**	**98.64%**
								Precisión de la clasificación		**98.64%**

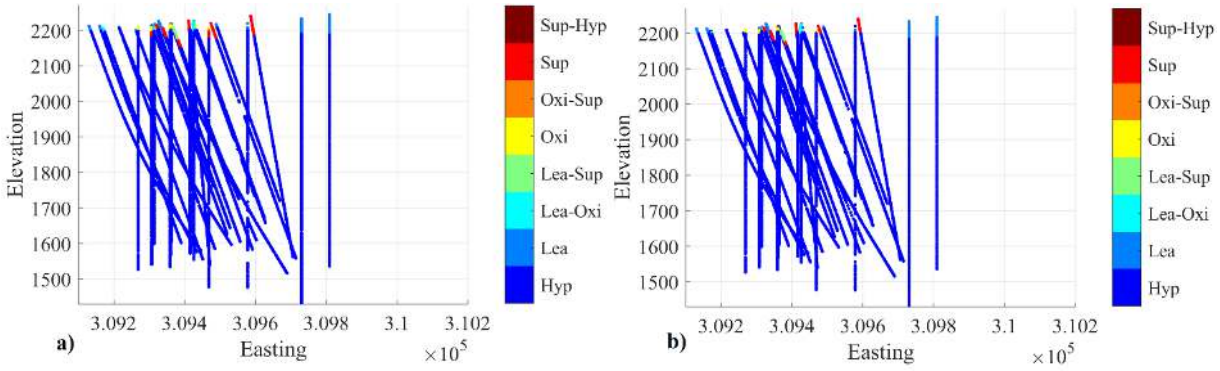

Figura 4.31. a) disposición espacial de las 8 zonas minerales en las muestras de entrenamiento; **b)** disposición luego de realizar la simulación con filtrado de ruido y la clasificación con CHAID.

6.3.2. Clasificación en el subconjunto de prueba

En este caso, el clasificador sin realizar el filtrado de ruido entrega una precisión de 94.02%; mientras que, al realizar el filtrado, la precisión sube un punto a 96.42% (Tabla 4.12; Fig. 4.12). Nuevamente, la zona mineral más mejora en la clasificación corresponde a hipógeno. En cambio, las zonas mixto lixiviado/supérgeno (Lea-Sup) y mixto supérgeno/hipógeno (Sup_Hyp) pierden todas sus muestras, pasando a hacer parte de las zonas hipógenas, lixiviadas y supérgenas y de las zonas hipógenas y supérgenas, respectivamente.

Tabla 4.12. Números de datos clasificados correctamente (línea diagonal) y datos clasificados incorrectamente (fuera de la diagonal) para cada zona mineral observada (fila) y cada zona mineral predicha (columna), para el subconjunto de prueba. Los números corresponden a los tipos de alteraciones más probable sobre 300 realizaciones construidas filtrando ruido.

Zona mineral observada	Predicción zona mineral (clasificación más probable)								Total	Precisión %
	Hyp	Lea	Lea_Oxi	Lea_Sup	Oxi	Oxi_Sup	Sup	Sup_Hyp		
Hipógeno (hyp)	**15705**	0	0	0	0	0	11	0	15716	99.93%
Lixiviado (Lea)	96	**152**	0	0	0	0	11	0	259	58.69%
Mixto lixiviado/óxido (Lea_Oxi)	46	0	**11**	0	0	0	25	0	82	13.41%
Mixto lixiviado/supérgeno (Lea_Sup)	4	12	0	**0**	0	0	33	0	49	0.00%
Óxido (Ox)	57	0	0	0	**7**	0	5	0	69	10.14%
Mixto óxido/supérgeno (Oxi_Su)	30	4	0	0	0	**39**	7	0	80	48.75%
Supérgeno (Su)	218	0	0	0	0	0	**347**	0	565	61.41%
Mixto supérgeno/hipógeno (Sup_Hy)	26	0	0	0	0	0	18	**0**	44	0.00%
Total	**16182**	**168**	**11**	**0**	**7**	**39**	**457**	**0**	**16864**	**96.42%**
								Precisión de la clasificación		**96.42%**

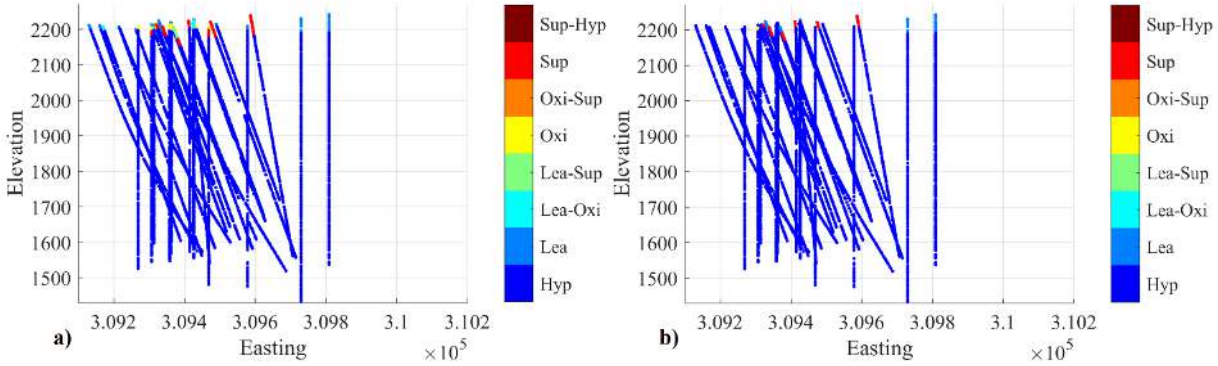

Figura 4.4. a) disposición espacial de las 8 zonas minerales en las muestras de prueba; **b)** disposición luego de realizar la simulación con filtrado de ruido y la clasificación con CHAID.

7. Discusión

7.1. Relaciones entre geoquímica, tipo de roca, alteración y zona mineral

Los depósitos de pórfido son clasificados típicamente en base a sus dotaciones de metales económicos (Kesler, 1973), así como en base a la composición de rocas magmáticas asociadas con mineralizaciones principales de cobre, molibdeno y renio (Sillitoe, 2010). Los pórfidos suelen estar asociados con el emplazamiento de complejos intrusivos intermedios a félsicos (Seedorff et al., 2005) como rocas de tipo dioritas (Dio), granodioritas a stocks de cuarzo-monzonita (Qmd), las cuales están asociadas a una alteración propilítica (Prp) (alteración de calcita, silicatos máficos) e incluso potásica (Pot); estos tipos de rocas y alteraciones hidrotermales tienen afinidad, en el presente caso de estudio, con concentraciones geoquímicas de Mg, Ca, Fe, Al, que son parte de minerales de la cloritización de la biotita, de silicatos máficos y de la epidota que remplaza a

plagioclasas; estas alteraciones suelen desarrollarse en zonas de mineral supergena (Sup) e hipogena (Hyp). En la superficie las rocas cuarzo-monzodioritas (Qmd) son muy alteradas, dando una alteración de tipo argilización (Arg) relacionada con zonas supergenas e hipogenas y con concentraciones geoquímicas como Ca, Cr, Th, Mg, Sc, Ce, Al, S, Ni y Pb, entre las principales. Además, la alteración silícica (Slc) puede ocurrir en la superficie en las rocas antes indicadas.

En las rocas máficas como andesita (Ans) y basalto (Ban), el conjunto de alteración potásica (Pot) es dominado por biotita rica en Mg, característico de valores altos de Al y magnetita, con poco cuarzo, feldespato potásico, anhidrita y sulfuros de Cu-Fe (Meyer and Hemley, 1967; Titley, 1982). En este tipo de rocas es común el emplazado de los stocks dominados por potásica (Pot) y propilítica (Prp) y seguido de alteraciones fílica (Phy), sílica (Chl) y argílica (Arg), que pueden desarrollarse en zonas minerales de tipo lixiviado (Lea) e hipógeno (Hyp).

Las vetas como diques de cuarzo dioritico (Qdi_D) y diques dioríticos (Dio_D) comúnmente definen un patrón radial y/o concéntrico alrededor de intrusiones centrales (e.g., Cannell et al., 2005) como la diorita (Dio). En consecuencia, ocurre la intrusión de múltiples lotes de magma andesítico y/o más máfico (Glazner et al., 2004); además los depósitos de pórfido son generalmente relacionados con magmas calco alcalinos intermedios a félsicos (Richards, 2009).

Las rocas que conforman un pórfido pueden enriquecerse en una variedad de metales traza (Thornton, 1995; Rose et al., 1979); sin embargo, la facilidad con la que estos elementos se liberan al medio ambiente desde las rocas durante la meteorización (denominada geodisponibilidad; Smith and Huyck, 1999) varía sustancialmente entre los tipos de rocas. Algunos de estas rocas, como las rocas intrusivas y/o volcánicas máficas a intermedias (Gab, Ban, Ans, Dio, Ton, Dac), contienen abundantes minerales de sulfuros como constituyentes primarios. Estos tipos de rocas a menudo tienen concentraciones elevadas de Fe, Mg, constituyentes de los minerales ferromagnesianos, y de metales traza (Goldschmidt, 1937; Guilbert and Park, 1986; Le Maitre, 1989; Thornton, 1995), como Yb, S, Co, Cr, Ni, Pb, Th, Cu, Ag, Sb, Zn, Se, Cd, Mn, As, y P, que se encuentran al menos en parte dentro de los sulfuros. En muchos casos, estas rocas fueron mineralizadas por los mismos procesos que forman los depósitos minerales.

Todas estas relaciones entre geoquímica, tipos de rocas, alteraciones y minerales explican las altas puntuaciones de clasificación obtenidas a partir de la información geoquímica (Tablas 4.7 a 4.12), así como el hecho de que la mayoría de los datos mal clasificados quedan asignados a otra categoría de características similares.

7.2. Fortalezas y debilidades de la metodología propuesta

Las principales enseñanzas que nos dejan los resultados obtenidos, tanto para tipo de roca, tipo de alteración y zona mineral, son las siguientes:

1) Existe una fuerte dependencia entre estas propiedades geológicas y las concentraciones geoquímicas, lo que explica las altas tasas de acierto logradas por la clasificación (entre 68% y 99%, según el caso).
2) La clasificación, tanto en las muestras de entrenamiento como en las muestras de prueba, mejora consistentemente al usar simulaciones filtradas del efecto pepita, en comparación con las realizaciones obtenidas de forma tradicional. Los variogramas modelados (Fig. 4.6) sugieren que el efecto pepita representa entre un 20% y un 30% de la variabilidad espacial

y corresponde mayormente a una variabilidad de pequeña escala y errores de medición en los ensayos geoquímicos que no se relacionan con los procesos geológicos. El filtrado del efecto pepita se basa en el análisis de corregionalización o análisis factorial geoestadístico (Goovaerts, 1992; Wackernagel, 2003), que permite separar la información geoquímica en componentes espaciales asociadas a diferentes escalas de variación; a pesar de que dicho análisis ha sido mayormente utilizado en el marco de problemas de predicción espacial (cokriging), es posible extenderlo a los problemas de simulación condicional. La mejora observada en la clasificación sugiere que las propiedades geológicas tales como el tipo de roca, el tipo de alteración y la zona mineral, poseen dependencias con las componentes espaciales de mayor escala, pero son estadísticamente y espacialmente independientes de la componente de pequeña escala, la cual corresponde a una fuente de variabilidad introducida (errores de medición) más que a una fuente de variabilidad natural. Por lo tanto, la asociación entre la geología y la geoquímica se manifiesta solamente a través de las primeras componentes.

3) La clasificación también mejora al considerar la clase más probable (aquella que más se repite en las 300 realizaciones construidas), en comparación con la clasificación obtenida sobre el promedio de las realizaciones. Esto sugiere un fuerte potencial de las técnicas de simulación condicional a los problemas de clasificación regionalizada que, hoy en día, se basan mayormente en predicciones obtenidas por kriging (similares al promedio de las realizaciones) que suavizan la variabilidad espacial, al momento de clasificar sitios para los cuales no existen mediciones.
4) Finalmente, se observa que las categorías más escasas tienden a desaparecer al momento de clasificar, en beneficio de las categorías más abundantes. Esto es una manifestación del dilema entre precisión y exactitud: para lograr una clasificación con la mejor tasa de acierto (mejor precisión), se tiende a privilegiar las categorías más abundantes o más probables, lo que provoco un sesgo (inexactitud) hacia estas categorías. Diversos autores (por ejemplo, Soares, 1992), han propuesto algoritmos que corrigen este sesgo, a costa de una pérdida de precisión.

Capitulo 5. Estudio de caso para datos de geoquímica de superficie

Este capítulo presenta una aplicación de la segunda propuesta metodológica descrita en el capítulo 3 para muestras superficiales de geoquímica y litología. Los contenidos de este capítulo han sido sometidos a publicación en la revista *Journal of Geochemical Exploration*.

> Guartán, J.A., Emery, X., 2020. *Predictive lithological mapping based on geostatistical joint modeling of lithology and geochemical element concentrations*. Journal of Geochemical Exploration, submitted (25/09/2020).

Predictive lithological mapping based on geostatistical joint modeling of lithology and geochemical element concentrations

Abstract

The spatial analysis and interpretation of lithological and geochemical sampling information are central in mineral prospecting and initial geological-mining exploration to delineate exploration targets and locate economic mineralization. This work compares two geostatistical approaches for the spatial prediction of lithological classes through a case study in mineral prospection, considering lithological and geochemical information at a set of surface samples. Both approaches calculate the probabilities of occurrence of the lithological classes at unsampled locations and select the most probable class as the predicted lithology. A split-sample technique is used to assess their performance, with the predictions being made at a testing data subset on the basis of the information of a training subset. The first approach relies on a cokriging of the lithological class indicators and yields an accuracy score (percentage of matches between true and predicted lithological classes) of 90.5%, while the second approach, consisting of a plurigaussian modeling of the classes, increases this score to 92.6%. Unlike the former approach, it also provides consistent outcomes of both the lithological classes and the geochemical covariates, which is valuable for mineral prospectivity mapping.

Key words: indicator cokriging; plurigaussian simulation; multivariate geostatistics; spatial prediction.

1. Introduction

The information of the surface and subsurface samples in mineral prospecting and exploration allows the management and planning of geological analyses to locate anomalies or continue studying and exploring an orebody. Both categorical (such as lithological classes) and quantitative (such as major and trace element abundances) properties observed or measured at these samples constitute search vectors to orebodies (Govett, 1983; Cameron et al., 2004; Urqueta et al., 2009; Cohen et al., 2010) and can be characterized and interpreted by means of multivariate statistics, data mining, GIS, spatial regression models and/or geostatistical methods (Sandjivy, 1984; Wackernagel and Sanguinetti, 1993; Jiménez-Espinosa et al., 1993; Jiménez-Espinosa and Chica-Olmo, 1999; Reis et al., 2003; Carranza, 2008; Grunsky, 2010; Afzal et al., 2012, 2016; Wang et al., 2013; Castillo et al., 2015; Wilford et al., 2016; Zuo and Wang, 2016; Adeli and Emery, 2021; Guartán and Emery, 2021).

Concerning the latter (geostatistical) methods, the modeling of quantitative and categorical variables is often achieved through a hierarchical or nested approach (Chilès and Delfiner, 2012), where the layout of the categories is first delineated, based on geomodeling techniques (Royer et al., 2015) or stochastic models (Armstrong et al., 2011; Mariethoz and Caers, 2014), prior to the spatial prediction or simulation of the quantitative variables in each category using the data belonging to this category (Talebi et al., 2016; Paithankar and Chatterjee, 2018; Maleki and Emery, 2020). Such an approach implies a controlling effect of the categorical variable on the quantitative variables, which turn out to be discontinuous when crossing the boundary between two categories. Numerous alternatives have been proposed to reduce or to avoid discontinuities across boundaries, including: the introduction of cross-correlations between the quantitative variables across

categories (Larrondo et al., 2004; Vargas-Guzmán, 2008; Mery et al., 2017), the use of overlapping (Ortiz and Emery, 2006) or probabilistic (Emery and González, 2007; Talebi et al., 2019b) categories, the modeling of quantitative variables and the posterior definition of the categories via statistical classification (Olea, 1999; Grunsky et al., 2014; Adeli et al., 2018; Talebi et al., 2019a; Adeli and Emery, 2021; Guartán and Emery, 2021), the joint prediction of the quantitative variables and category indicators (Dowd, 1993; Kasmaee et al., 2019) or of service variables defined as the product of a quantitative variable with an indicator (Séguret, 2013; Kasmaee et al., 2019), or the joint simulation of the quantitative and indicator variables, e.g., by combining a multigaussian model for the former and a plurigaussian model for the latter (Dowd, 1994; Emery and Silva, 2009; Maleki and Emery, 2015). However, the model complexity often limits the use of the joint simulation approach to a few quantitative variables. Some of the aforementioned approaches do not provide an outcome of the categorical variables, as the focus is on the prediction or simulation of the quantitative variables, while the others provide a deterministic or a probabilistic representation of the categories.

Having a spatial representation of the categorical variable (lithological classes) is of great importance in mineral prospection and initial geological-mining exploration studies, in which the lithology can constitute a search vector that complements the geochemical information for identifying exploration targets and for planning drilling campaigns aimed at locating and defining economic mineralization. For instance, the presence of intermediate or felsic volcanic rocks that are altered by the presence of an intrusive igneous body provides relevant information for the exploration of porphyry deposits and hydrothermal veins (Sillitoe, 2003).

In this context, the objective of this work is to compare two geostatistical approaches for the spatial prediction of lithological classes, based on lithological and geochemical information available at surface samples. The outline is the following: Section 2 presents the data set under consideration, the methodology used to assess the performance and compare the predictive models, and details the two proposed approaches. The first one consists of an indicator coding and a cokriging of the lithological classes, while the second one relies on a truncated plurigaussian simulation of the lithological classes and its extension to include the geochemical covariates. This second approach constitutes the main novelty of the work, through the elaboration of a mixed model that accounts for both the quantitative and categorical variables and allows not only the prediction of the lithological classes, but also the calculation of the probabilities of occurrence of these classes and the prediction of geochemical concentrations at any unsampled location. The prediction results are presented in Section 3 and discussed in Section 4, and conclusions are drawn in Section 5.

2. Materials and methods

2.1. Study area

This work focuses on the modeling of a data set from a mining exploration area located in southern Ecuador. The area is covered by Tertiary volcanic rocks and is composed of an undifferentiated sequence of the Saraguro group of subaerial, calcoalkaline, intermediate to acidic volcanic rocks. Andesitic to dacitic rocks predominate, but rhyolitic rocks are common too. They are in a stratigraphic sequence with a rhyolitic-breccia mass flow, tuffaceous sandstones and white fine tuffs of the Plancharumi formation, and are overlain by crystal tuffs of rhyolitic composition. Polymictic clast supported hydrothermal breccias showing silicified clay fragments and silica-clay

alteration in both matrix and clasts are also present. Moreover, several intrusive bodies occur: a microdiorite intrusive cross-cut by pyrite and quartz-pyrite veinlets (stockwork), a diorite porphyry intrusive, with clay-silica alteration and disseminated pyrite, and rhyolite subvolcanic intrusive domes (Baldock, 1982).

The geochemical database contains 3998 samples, as the result of a surface sampling with a horizontal spacing that varies from 100m × 400m to 50m × 50m (Fig. 5.1a). Each sample has information on the concentrations of 36 geochemical elements, reported in percentage (%) for major elements and in part per million (ppm) for trace elements (Table 5.1) and the prevailing lithological class. The sum of the 36 concentrations varies from 0.697% to 25.98%, with an average of 5.117% over the 3998 samples (Fig. 5.1b). Since not all the geochemical elements have been measured (in particular, information on silicon and many trace elements is missing), this sum is not constant and far from 100%, so that no compositional treatment (e.g. log-ratio transformation) is required here prior to geostatistical modeling. As for the lithology, ten classes (Table 5.2) are considered based on an interpreted field map (Fig. 5.1a) prepared by exploration geologists. The lithology can be directly observed for most of the samples (approximately 80% of the total), for which the bedrock is exposed or the soil contains rock fragments allowing to identify the lithological class. For the remaining 20% of samples with an important surface soil or subsoil horizon preventing a direct observation of the bedrock, the lithology can be determined from the observation of neighboring outcrops.

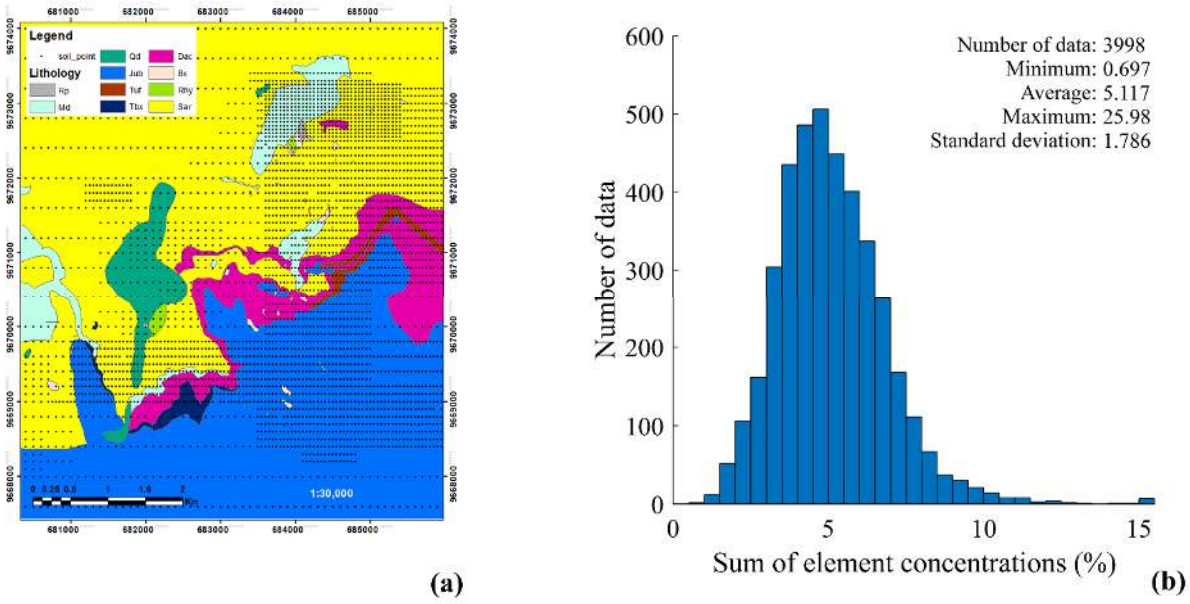

Figure 5.1. a) Interpreted geological map of the study area and location of surface samples. Sampling is generally on a grid of 100m × 400m to 50m × 50m (Source: Mining Company Cornerstone Ecuador S.A., 2012); b) histogram of the sum of 9 major and 27 trace element concentrations, expressed in percent.

Table 5.1. Quantitative variables with their coding and definition

Definition	Variable	Symbol	Variable	Symbol	Variable	Symbol
9 major elements (%)	Iron	Fe	Magnesium	Mg	Sodium	Na
	Calcium	Ca	Titanium	Ti	Potassium	K
	Phosphorus	P	Aluminum	Al	Sulfur	S
27 trace elements (ppm)	Molybdenum	Mo	Uranium	U	Barium	Ba
	Copper	Cu	Thorium	Th	Boron	B
	Lead	Pb	Strontium	Sr	Tungsten	W
	Zinc	Zn	Cadmium	Cd	Scandium	Sc
	Silver	Ag	Antimony	Sb	Thallium	Tl
	Nickel	Ni	Bismuth	Bi	Mercury	Hg
	Cobalt	Co	Vanadium	V	Selenium	Se
	Manganese	Mg	Lanthanum	La	Tellurium	Te
	Arsenic	As	Chromium	Cr	Gallium	Ga

Table 5.2. Rock types, codification and proportions in the study zone

Rock type	Lithology	Codification	Number of data	lithology proportions
Hydrothermal Breccia	Bx	1	24	0.60%
Tuffaceous Sandstone (Plancharumi formation)	Dac	2	390	9.75%
Tuff (Jubones Formation)	Jub	3	1274	31.86%
Microdiorite/Andesite Porphyry	Md	4	333	8.33%
Quartz diorite/Tonalite Porphyry	Qd	5	101	2.53%
Rhyolite	Rhy	6	11	0.27%
Rhyolite Porphyry	Rp	7	6	0.15%
Andesite/dacite/rhyolite (Saraguro group)	Sar	8	1774	44.37%
Rhyolite Breccia Tuffs (Plancharumi formation)	Tbx	9	36	0.92%
White fine Tuffs (Plancharumi formation)	Tuf	10	49	1.22%
		Total	**3998**	**100%**

A contact analysis based on the examination of lagged scatter plots (Maleki and Emery, 2020) reveals no change in the mean values and correlations of the geochemical concentrations across lithological class boundaries (Fig. 5.2a.), which supports the assumption of soft boundaries, i.e., with no clear-cut discontinuity of the geochemical concentrations when crossing the boundary between two lithological classes. On the other hand, the matrix of point biserial correlation coefficients between the class indicators and the geochemical concentrations indicates significant relationships between lithology and geochemistry (Fig. 5.2b). In particular, we highlight several correlations (uranium with Tuf, Dac, Md and Sar, vanadium with Md and Sar, cadmium with Sar, lead with Qd, Sar and Jub, and silver, molybdenum and thorium with Jub, Sar and Md) that will be useful in the following (Section 2.4) to elaborate a simplified lithology-geochemistry model.

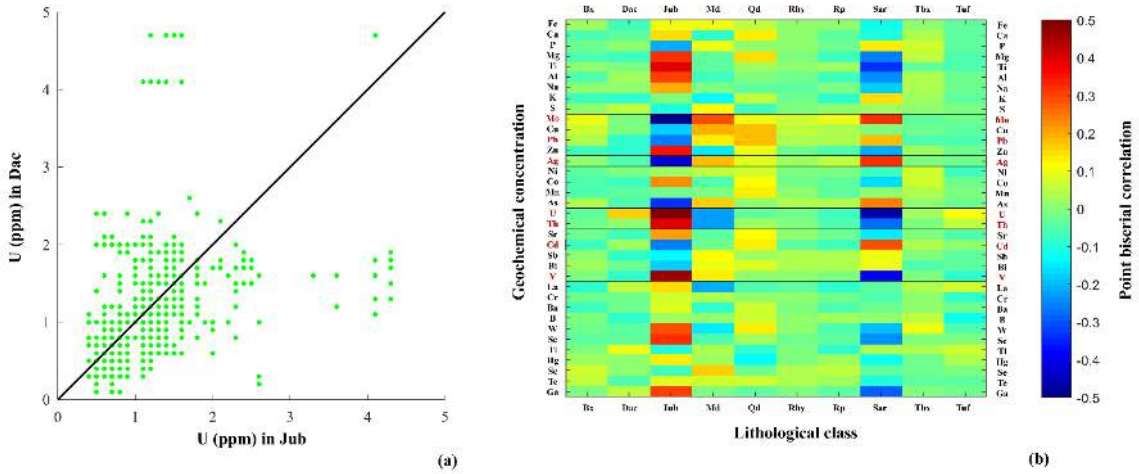

Figure 5.2. (a) An example of lagged scatter plot showing the uranium concentration measured in Dac (lithological class 2) as a function of the uranium concentration measured in Jub (lithological class 3), provided that both measurements are no more than 150 meters apart. The plot indicates a strong correlation across the boundary between the two lithological classes and no significant change in the mean value. (b) Color representation of the point biserial correlation matrix between class indicators and geochemical concentrations

2.2. Methodology

Our primary goal is to develop a geostatistical model to predict the lithological class everywhere in the area under study, knowing the lithological and geochemical information at a set of sampling locations. In order to assess the quality of the predictive models, the original sample (3998 data) is randomly split into two subsets: a training subset (3198 data) on which the model will be constructed, and a testing subset (800 data) on which predictions will be held and compared with reality (Fig. 5.3). The percentage of match between predicted and true lithological classes on the testing subset, referred to as the 'accuracy score hereinafter', will give an insight into the ability of the model to predict the lithology at any unsampled location. The methodology is summarized in the schematic diagram in Fig. 5.4 and detailed in the following subsections.

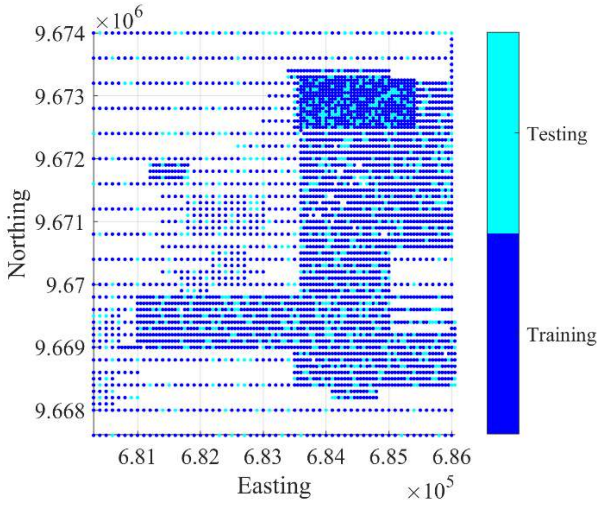

Figure 5.3. Random division of surface samples into a training subset (3198 samples) and a testing subset (800 samples)

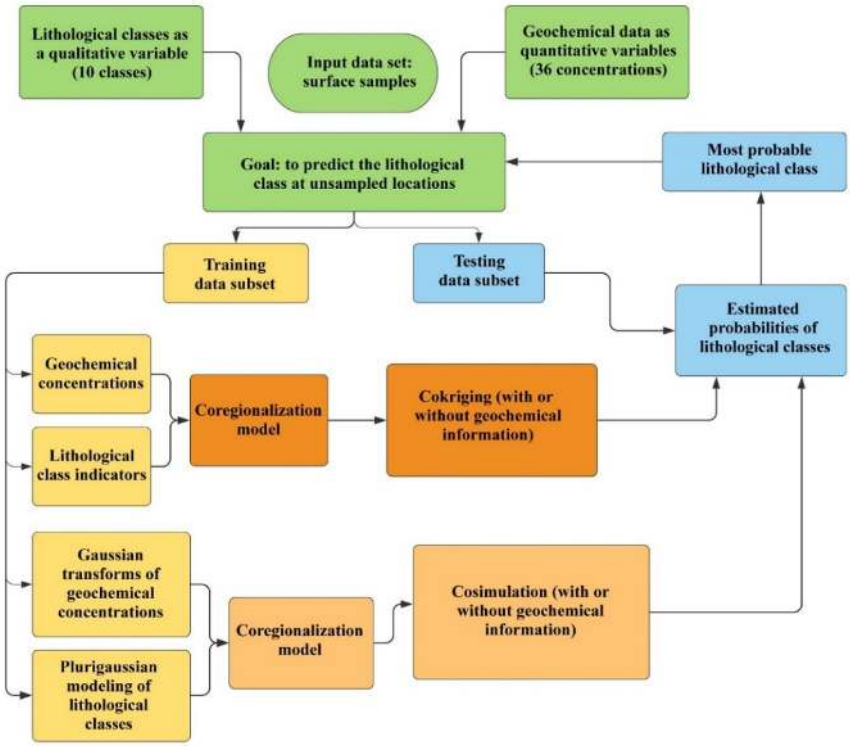

Figure 5.4. Schematic diagram of the methodological proposal

2.3. First approach: indicator cokriging

This approach relies on a direct coding of the lithological classes into indicator (binary) variables:

- $I_1(\mathbf{x})$ = 1 if location **x** belongs to Bx, 0 otherwise
- $I_2(\mathbf{x})$ = 1 if location **x** belongs to Dac, 0 otherwise
- $I_3(\mathbf{x})$ = 1 if location **x** belongs to Jub, 0 otherwise
- $I_4(\mathbf{x})$ = 1 if location **x** belongs to Md, 0 otherwise
- $I_5(\mathbf{x})$ = 1 if location **x** belongs to Qd, 0 otherwise
- $I_6(\mathbf{x})$ = 1 if location **x** belongs to Rhy, 0 otherwise
- $I_7(\mathbf{x})$ = 1 if location **x** belongs to Rp, 0 otherwise
- $I_8(\mathbf{x})$ = 1 if location **x** belongs to Sar, 0 otherwise
- $I_9(\mathbf{x})$ = 1 if location **x** belongs to Tbx, 0 otherwise
- $I_{10}(\mathbf{x})$ = 1 if location **x** belongs to Tuf, 0 otherwise.

The rationale is to predict the indicators at each location of the testing subset by cokriging based on the information of the training subset. The predicted indicators can be interpreted as estimates of the probabilities of occurrence of the lithological classes (Solow, 1986; Goovaerts, 1997); the predicted class can be taken as the one for which the predicted indicator is the greatest, which amounts to consider the most probable lithology (Goovaerts, 1997; Kasmaee et al., 2019; Talebi et al., 2019b). Note that other assignation schemes could be designed, for instance, to guarantee the

reproduction of the proportions of the lithological classes (Soares, 1992), at the cost of a lower precision, i.e., a greater mean squared error.

Because of the compositional constraint ($I_1 + \ldots + I_{10} = 1$), cokriging the ten indicators would lead to a singular system of equations and it is actually sufficient to predict nine out of the ten indicators, the remaining one being deduced as the complementary to one of the nine predictions (Goovaerts, 1997; Wackernagel, 2003). In the present case, cokriging is performed without considering indicator I_7, which corresponds to the scarcest class (Rp, present in only 0.15% of the data, as per Table 5.2).

The steps to follow for this approach are:

1. Calculate the experimental direct and cross-variograms of the indicator data and fit a linear model of coregionalization in order to get a valid set of theoretical direct and cross-variograms.
2. At each testing data location, predict the indicators by ordinary cokriging (i.e., cokriging with unknown mean values) (Goovaerts, 1997; Wackernagel, 2003) and select the lithological class for which the cokriged indicator is the greatest one.

As for Step 1, no anisotropy is identified in the experimental variograms, so that the fitting relies on omnidirectional direct and cross-variograms. A least-square algorithm (Goulard and Voltz, 1992) is used to fit the variogram models by nesting a nugget effect, two isotropic spherical and two isotropic exponential structures (Table 5.3; Fig. 5.5). Concerning Step 2, cokriging is performed using a moving neighborhood with a radius of 150 m in order to incorporate the training samples adjacent to the target testing sample.

Table 5.3. Nested structures used for the fitting of indicator direct and cross-variograms

Nested structure	Structure type	Range of correlation (m)
1	Nugget effect	0
2	Spherical	350
3	Spherical	500
4	Exponential	1500
5	Exponential	2500

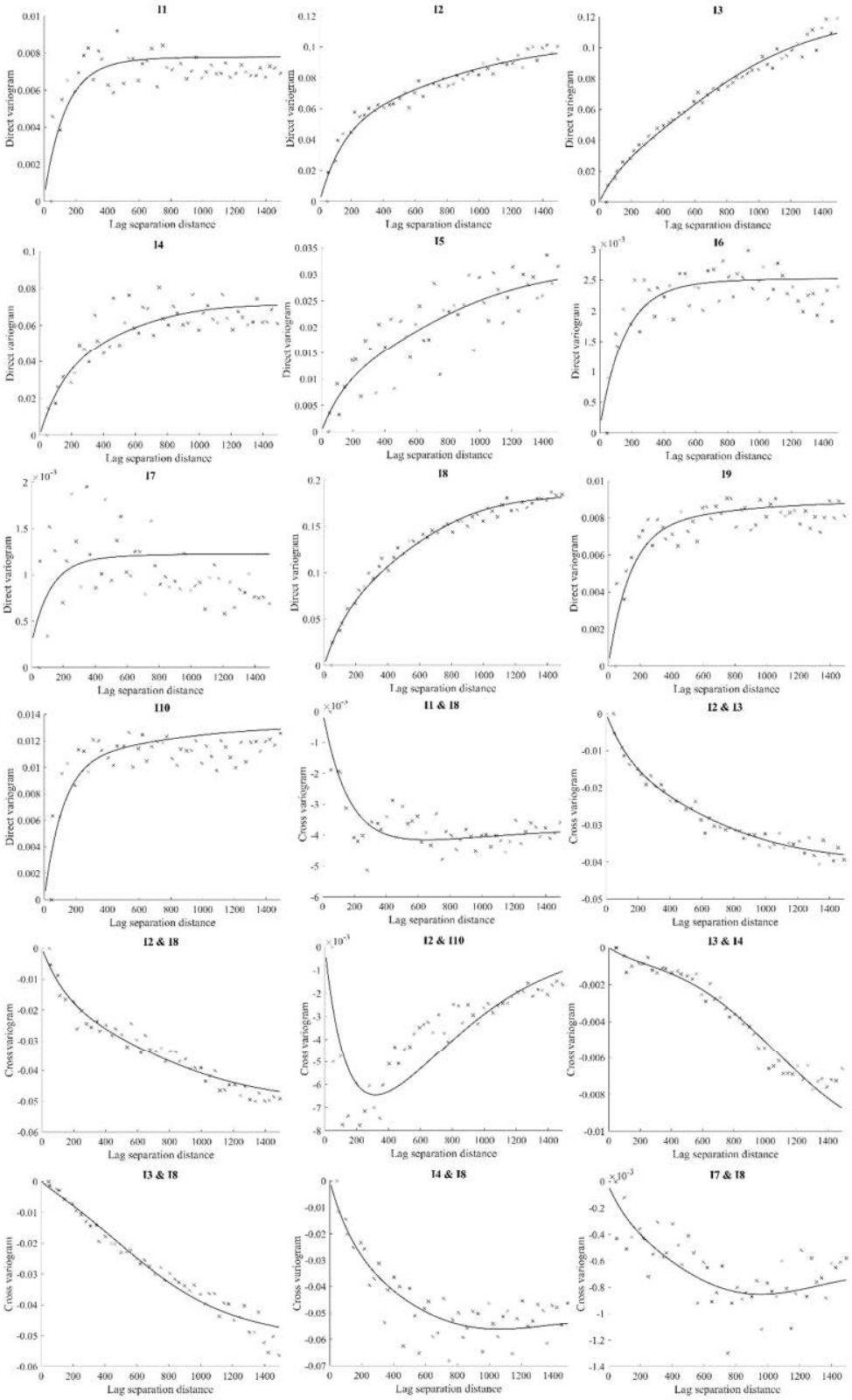

Figure 5.5. An example of omnidirectional experimental (crosses) and modeled (solid lines) direct and cross-variograms for lithology indicators I_1 to I_{10}. All the modeled variograms are isotropic

2.4. Second approach: mixed plurigaussian-multigaussian simulation

A second approach is proposed to determine the probabilities of occurrence of the lithological classes at the testing data locations, based on a modeling of the lithological and geochemical information with Gaussian random fields, as detailed in the next subsections.

2.4.1. Plurigaussian modeling of the lithological classes

2.4.1.1. Truncation rule

In this model (Armstrong et al., 2011), it is hypothesized that the lithology indicators are obtained by truncating a set of parent standard Gaussian random fields. The determination of the number of such Gaussian random fields and of the truncation rule is made in accordance with the interpreted contacts between lithological classes (Fig. 5.1a) and with their chronology (Madani and Emery, 2015): on the one hand, a younger class can cross-cut an older class (in space and in the flag that represents the truncation rule), while the reverse is not permitted; on the other hand, two classes that have a contact in space should also have a contact in the flag. As a result, a seven-dimensional flag is elaborated (Fig. 5.6), i.e., seven Gaussian random fields (denoted as $Y_1,\dots Y_7$) are used to codify the lithological classes, as expressed in Table 5.4.

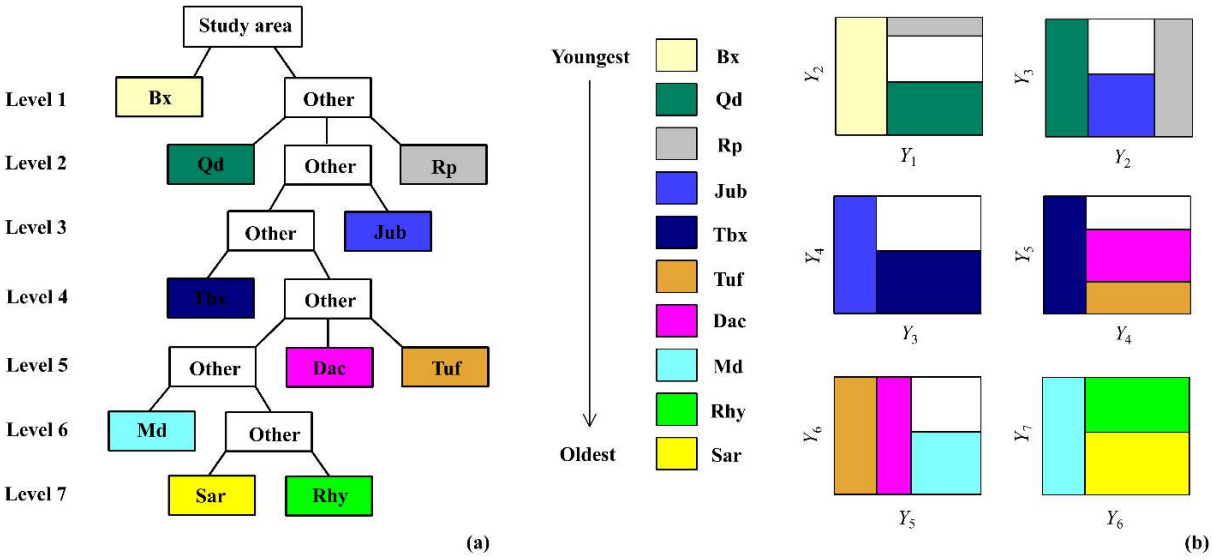

Figure 5.6. (a) Partition of the study area into ten lithological classes, based on the interpreted contacts (Fig. 1a) and chronological ordering. The first level splits the study area into Bx (youngest lithological class) and the remaining classes; among the latter, the second level distinguishes Qd, Rp and the rest, and so on until the seventh level that separates the oldest classes Rhy and Sar. This partition translates into a seven-dimensional truncation flag, each dimension being associated with a Gaussian random field Y_i (with $i = 1,\dots, 7$). Subfigure (b) displays two-dimensional sections of this multidimensional truncation flag showing the association between the lithological classes and the Gaussian random fields; horizontal and vertical lines symbolize the truncation thresholds acting on the latter random fields. Younger classes can cross-cut older ones, but the reverse is not permitted, in agreement with the lithological chronology. Also, the truncation rule prohibits contacts between Qd and Rp, as well as between Tuf and Md, Rhy or Sar, which agrees with the interpreted field map in Fig. 1a

Tabla 5.4. Plurigaussian truncation rule showing how the lithological classes are obtained from the truncation of the Gaussian random fields. The rule involves nine numerical thresholds ($y_1, y_2, y_2', y_3, y_4, y_5, y_5', y_6, y_7$)

Lithological class	Inequality constraints on the Gaussian random fields						
Bx	$Y_1<y_1$						
Qd	$Y_1\geq y_1$	$Y_2<y_2$					
Rp	$Y_1\geq y_1$	$Y_2\geq y_2'$					
Jub	$Y_1\geq y_1$	$y_2'>Y_2\geq y_2$	$Y_3<y_3$				
Tbx	$Y_1\geq y_1$	$y_2'>Y_2\geq y_2$	$Y_3\geq y_3$	$Y_4<y_4$			
Tuf	$Y_1\geq y_1$	$y_2'>Y_2\geq y_2$	$Y_3\geq y_3$	$Y_4\geq y_4$	$Y_5<y_5$		
Dac	$Y_1\geq y_1$	$y_2'>Y_2\geq y_2$	$Y_3\geq y_3$	$Y_4\geq y_4$	$y_5'>Y_5\geq y_5$		
Md	$Y_1\geq y_1$	$y_2'>Y_2\geq y_2$	$Y_3\geq y_3$	$Y_4\geq y_4$	$Y_5\geq y_5'$	$Y_6<y_6$	
Sar	$Y_1\geq y_1$	$y_2'>Y_2\geq y_2$	$Y_3\geq y_3$	$Y_4\geq y_4$	$Y_5\geq y_5'$	$Y_6\geq y_6$	$Y_7<y_7$
Rhy	$Y_1\geq y_1$	$y_2'>Y_2\geq y_2$	$Y_3\geq y_3$	$Y_4\geq y_4$	$Y_5\geq y_5'$	$Y_6\geq y_6$	$Y_7\geq y_7$

2.4.1.2.Truncation thresholds

The above truncation rule depends on nine numerical thresholds ($y_1, y_2, y_2', y_3, y_4, y_5, y_5', y_6, y_7$) that can be determined on the basis of the proportions of each lithological class. Under the assumption that the seven Gaussian random fields are statistically independent, the first two-dimensional section in Figure 6b gives:

- $\mathrm{P}\{\mathrm{Bx}\} = \mathrm{P}\{Y_1(\mathbf{x}) < y_1\} = G(y_1)$
- $\mathrm{P}\{\mathrm{Qd}\} = \mathrm{P}\{Y_1(\mathbf{x}) \geq y_1\} \times \mathrm{P}\{Y_2(\mathbf{x}) < y_2\} = [1 - G(y_1)]\,G(y_2)$
- $\mathrm{P}\{\mathrm{Rp}\} = \mathrm{P}\{Y_1(\mathbf{x}) \geq y_1\} \times \mathrm{P}\{Y_2(\mathbf{x}) \geq y_2'\} = [1 - G(y_1)][1 - G(y_2')]$

with P standing for the probability and G for the standard Gaussian cumulative distribution function. Similarly, for the remaining classes, one has:

- $\mathrm{P}\{\mathrm{Jub}\} = [1 - G(y_1)][G(y_2') - G(y_2)]G(y_3)$
- $\mathrm{P}\{\mathrm{Tbx}\} = [1 - G(y_1)][G(y_2') - G(y_2)][1 - G(y_3)]G(y_4)$
- $\mathrm{P}\{\mathrm{Tuf}\} = [1 - G(y_1)][G(y_2') - G(y_2)][1 - G(y_3)][1 - G(y_4)]G(y_5)$
- $\mathrm{P}\{\mathrm{Dac}\} = [1 - G(y_1)][G(y_2') - G(y_2)][1 - G(y_3)][1 - G(y_4)][G(y_5') - G(y_5)]$
- $\mathrm{P}\{\mathrm{Md}\} = [1 - G(y_1)][G(y_2') - G(y_2)][1 - G(y_3)][1 - G(y_4)][1 - G(y_5')]G(y_6)$
- $\mathrm{P}\{\mathrm{Sar}\} = [1 - G(y_1)][G(y_2') - G(y_2)][1 - G(y_3)][1 - G(y_4)][1 - G(y_5')][1 - G(y_6)]G(y_7)$
- $\mathrm{P}\{\mathrm{Rhy}\} = [1 - G(y_1)][G(y_2') - G(y_2)][1 - G(y_3)][1 - G(y_4)][1 - G(y_5')][1 - G(y_6)][1 - G(y_7)]$

The above equations provide a one-to-one relationship between the truncation thresholds and the class probabilities, allowing to determine the former knowing the latter. By identifying the probabilities with the proportions informed in Table 5.2, one finds:

$$\begin{cases} y_1 = -2.51562 & y_2 = -1.95508 & y_2' = -2.95312 \\ y_3 = -0.45020 & y_4 = -2.06250 & y_5 = -1.89062 \\ y_5' = -0.79492 & y_6 = -1.04102 & y_7 = 2.29688 \end{cases}$$

2.4.1.3.Coding lithological classes into indicators

As for the indicator cokriging approach, the lithological classes are coded into indicators associated with the underlying Gaussian random fields. The indicator can be unknown if the Gaussian random field is not involved in the definition of the class, otherwise it is either 0 or 1 (Table 5.5).

Tabla 5.5. Coding of lithological classes into indicators of Gaussian random fields

	$Y_1<y_1$	$Y_2<y_2$	$Y_2>y_2'$	$Y_3<y_3$	$Y_4<y_4$	$Y_5<y_5$	$Y_5<y_5'$	$Y_6<y_6$	$Y_7<y_7$
Bx	1								
Qd	0	1	0						
Rp	0	0	1						
Jub	0	0	0	1					
Tbx	0	0	0	0	1				
Tuf	0	0	0	0	0	1	1		
Dac	0	0	0	0	0	0	1		
Md	0	0	0	0	0	0	0	1	
Sar	0	0	0	0	0	0	0	0	1
Rhy	0	0	0	0	0	0	0	0	0

2.4.1.4.Covariance analysis

The experimental nonergodic covariances (Isaaks and Srivastava, 1988) of seven indicator variables ($Y_1<y_1$, $Y_2<y_2$, $Y_3<y_3$, $Y_4<y_4$, $Y_5<y_5'$, $Y_6<y_6$ and $Y_7<y_7$) are calculated on the training data, again in an omnidirectional manner insofar as no clear anisotropy can be identified. The formula relating the covariance of a Gaussian random field with that of its indicators (Matheron, 1989; Chilès and Delfiner, 2012) is then used to convert the experimental indicator covariances into experimental covariances of the Gaussian random fields, following the approach of Emery and Cornejo (2010), and the latter are fitted with combinations of isotropic spherical, exponential and cubic structures with ranges varying between 100 and 10,000 m, in such a way that the total sill (variance) is equal to 1 for each Gaussian random field (Fig. 5.7; Table 5.6). The cross-covariances are identically zero since the Gaussian random fields are assumed independent.

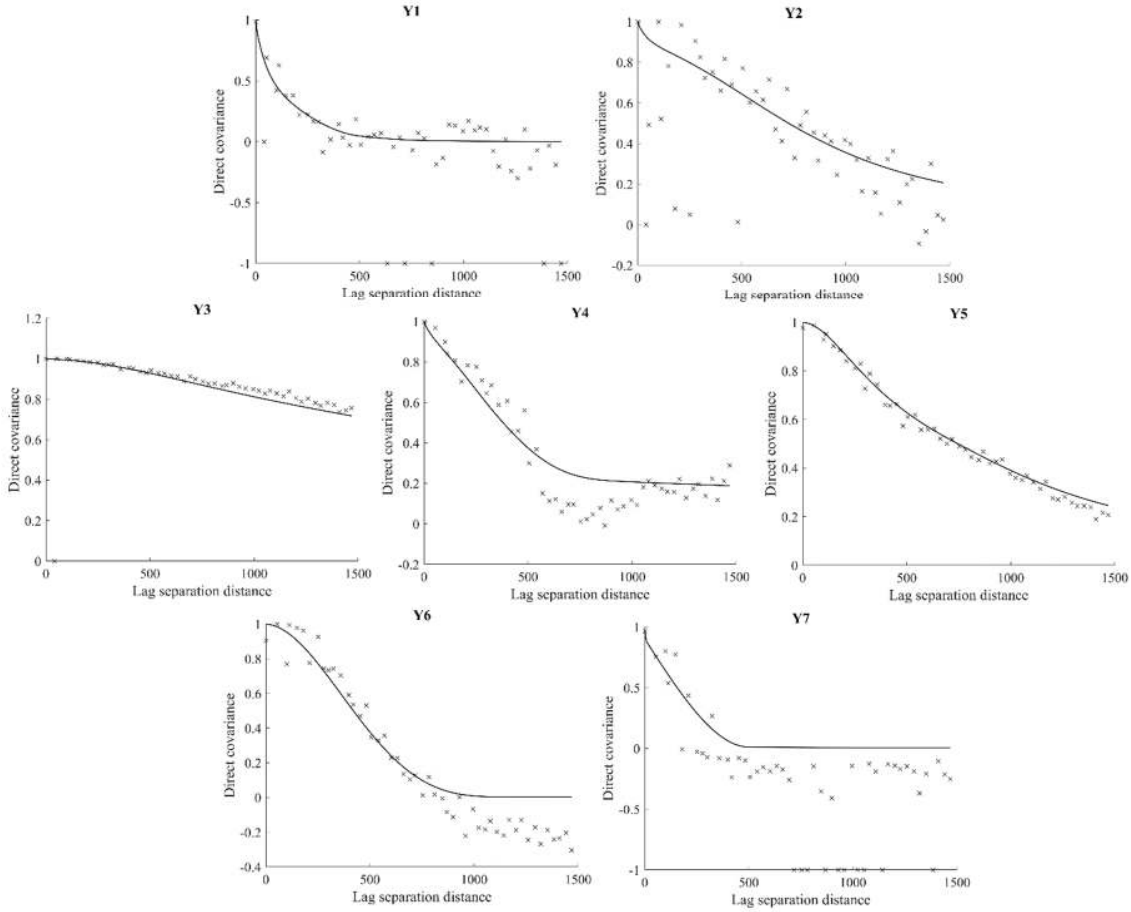

Figure 5.7. Omnidirectional experimental (crosses) and modeled (solid lines) direct covariances of the Gaussian random fields associated with the lithology indicators. All the modeled covariances are isotropic. The experimental covariances are derived from that of the lithological class indicators, based on the formula linking the covariance of a Gaussian random field with the covariance of the truncated random field

Table 5.6. Nested structures used for fitting the direct covariances of the Gaussian random fields

Nested structure	Structure type	Range of correlation (m)	Nested structure	Structure type	Range of correlation (m)
1	Nugget effect	0	12	Cubic	350
2	Spherical	100	13	Cubic	400
3	Spherical	350	14	Cubic	500
4	Spherical	500	15	Cubic	800
5	Spherical	800	16	Cubic	1,000
6	Spherical	1,000	17	Cubic	1,200
7	Spherical	1,600	18	Cubic	1,600
8	Spherical	10,000	19	Cubic	2,000
9	Exponential	150	20	Cubic	3,000
10	Exponential	450	21	Cubic	5,000
11	Exponential	1,500	22	Cubic	10,000

2.4.1.5.Simulation at testing data locations

At each testing data location, the Gaussian random fields Y_1, …, Y_7 are simulated conditionally to the adjacent training data (located at no more than 150 meters from the target). The simulation process is iterative, based on Gibbs sampling (Armstrong et al., 2011). For conciseness, it is explained for the first Gaussian random field Y_1, assuming that $\mathbf{x}_0$ is the target testing data location and $\mathbf{x}_1 \ldots \mathbf{x}_n$ are the selected adjacent training data locations:

(1) Set $Y_1^{(0)}(\mathbf{x}_\alpha) = y_1$ for $\alpha = 0, \ldots, n$.
(2) Calculate the $(n+1) \times (n+1)$ variance-covariance matrix $\mathbf{C}_1$ of $(Y_1(\mathbf{x}_0), \ldots, Y_1(\mathbf{x}_n))$
(3) Invert $\mathbf{C}_1$ to obtain the precision matrix $\mathbf{B}_1 = \mathbf{C}_1^{-1}$
(4) Iteration: for $k = 1, \ldots, K$
    (a) Select an index α between 0 and n.
    (b) Set $Y_1^{(k)}(\mathbf{x}_\beta) = Y_1^{(k-1)}(\mathbf{x}_\beta)$ for $\beta = 0, \ldots, \alpha - 1, \alpha+1, \ldots, n$.
    (c) Predict $Y_1(\mathbf{x}_\alpha)$ by simple kriging based on $\{Y_1(\mathbf{x}_\beta): \beta = 0, \ldots, \alpha - 1, \alpha+1, \ldots, n\}$. Denote by y_1^* and σ_1^* the kriging prediction and the standard deviation of the kriging error, respectively.
    (d) If $\alpha = 0$ (location $\mathbf{x}_\alpha$ is the target testing data), set $Y_1^{(k)}(\mathbf{x}_\alpha) = y_1^* + \sigma_1^* N_1$, where N_1 is a standard Gaussian random variable.
    (e) If $\alpha \geq 1$ (location $\mathbf{x}_\alpha$ is a training data):
        (i) If Bx prevails at $\mathbf{x}_\alpha$, set $Y_1^{(k)}(\mathbf{x}_\alpha) = y_1^* + \sigma_1^* N_1$, with N_1 a standard Gaussian random variable subject to $y_1^* + \sigma_1^* N_1 < y_1$
        (ii) Otherwise, set $Y_1^{(k)}(\mathbf{x}_\alpha) = y_1^* + \sigma_1^* N_1$, with N_1 a standard Gaussian random variable subject to $y_1^* + \sigma_1^* N_1 \geq y_1$.

At Step (4a), the locations are visited following random permutations and the number K of iterations is chosen such that each location is visited 100 times. The kriging weights and kriging variance needed at Step (4c) are directly obtained from the precision matrix $\mathbf{C}_1$ calculated at Step (3) (Dubrule 1983), which avoids solving many kriging systems of equations. The procedure is repeated in order to generate three hundred realizations of the Gaussian random fields Y_1, …, Y_7 at each testing location, which are converted into as many realizations of the lithological class via the truncation rule (Table 5.4).

2.4.1.6.Prediction of lithology at testing data locations

The 300 realizations are used to calculate the probabilities of occurrence of each lithological class at each testing data location. Then, following the same reasoning as in indicator cokriging, the most probable lithological class (i.e., the one that appears more frequently among the 300 realizations) is retained as the prediction. The number of realizations (300) has been chosen purposely large in order to make sure that the probabilities are accurately estimated and that the predicted class matches the truly most probable class at all, or practically all, the target locations.

2.4.2. Use of geochemical concentrations as covariates

2.4.2.1.Gaussian anamorphosis

To incorporate the 36 geochemical concentrations as covariates in the simulation process, it is necessary to transform these variables into standard Gaussian random fields. Such a transformation is held on each variable separately, based on the 3198 training data. An alternative to the separate transformation of each variable is a joint transformation, e.g., with the flow anamorphosis approach (Talebi et al., 2019b). This option has been discarded for two reasons. The first one is the 'curse of dimensionality': it is illusory to believe that a sample of 3198 data could give a precise representation of a 36-variate distribution; instead, all the data are likely to be far away from both the center and the tails and to be concentrated in the 'middle' of such a multivariate distribution (Giraud, 2014). The second reason is that a joint transformation makes the transformed Gaussian random fields be a mix of the 36 original geochemical concentrations: the associations between geochemistry and lithology become difficult to interpret and it is no longer possible to reduce the dimensionality of the problem (i.e., to jointly simulate the lithological classes and a reduced subset of geochemical concentrations, deemed the most interesting) without simulating the whole set of Gaussian random fields; as most of these random fields are spatially cross-correlated, this yields great complications in the covariance analysis stage, see the next subsection..

2.4.2.2.Covariance analysis

The high amount of Gaussian random fields (7 + 36 = 43) and the fact that their experimental covariances do not always share the same structural features, in particular, the behavior at the origin is often smooth for $Y_1 \dots Y_7$ (Fig. 5.7) but not for the 36 remaining fields, prevents the fitting of a full linear model of coregionalization involving the 43 direct covariances and 903 cross-covariances, even when resorting to automatic or semi-automatic fitting algorithms, with more reason because of the constraints of unit sill for the first seven direct covariances and zero sill for the associated cross-covariances.

The proposed solution is to select a subset of (Gaussian transforms of) geochemical concentrations, the ones exhibiting the most significant spatial cross-correlations with the Gaussian random fields $Y_1 \dots Y_7$ associated with the lithological classes. This way, the covariance modeling reduces to several small problems that can be solved by manual or semi-automatic fitting procedures. These are:

- one variable alone: Y_1
- three groups of two variables: $\{Y_5, Y_U\}$, $\{Y_6, Y_V\}$ and $\{Y_7, Y_{Cd}\}$
- one group of three variables: $\{Y_2, Y_4, Y_{Pb}\}$
- one group of four variables: $\{Y_3, Y_{Ag}, Y_{Mo}, Y_{Th}\}$.

The pertinence of such a simplification is corroborated by the fact that the selected geochemical covariates (U, V, Cd, Pb, Ag, Mo, Th) are the ones that exhibit the strongest correlations with the lithological classes Dac, Jub, Md, Qd, Sar and Tuf derived from the Gaussian random fields Y_2 to Y_7 (Fig. 5.2b). The absence of association between Y_1 with a geochemical concentration is not a relevant issue, insofar as Y_1 defines the Bx lithological class (Fig. 5.6b), which only accounts for 0.60% of the data (Table 5.2). The fitted covariance models (Fig. 5.8), based on the same basic nested structures as in Table 5.6, is such that the direct covariances of $Y_1 \dots Y_7$ are the same as that shown in Fig. 5.7. The cross-covariance between Y_2 and Y_4 is not identically zero, but it has a zero value at the origin, so that both random fields have independent values when they are taken at the same location; as a consequence, the determination of the truncation thresholds remains valid.

2.4.2.3.Prediction at testing data locations

The Gaussian random fields are simulated at each testing data location, following the same scheme as before (Section 2.4.1.5), except that the Gaussian random fields associated with the geochemical covariates are now considered. This implies larger covariance and precision matrices at Steps (2) and (3) and the use of simple cokriging at Step (4c) instead of simple kriging. Again, 300 realizations are generated and the most probable lithological class is used as the prediction at the target testing data location.

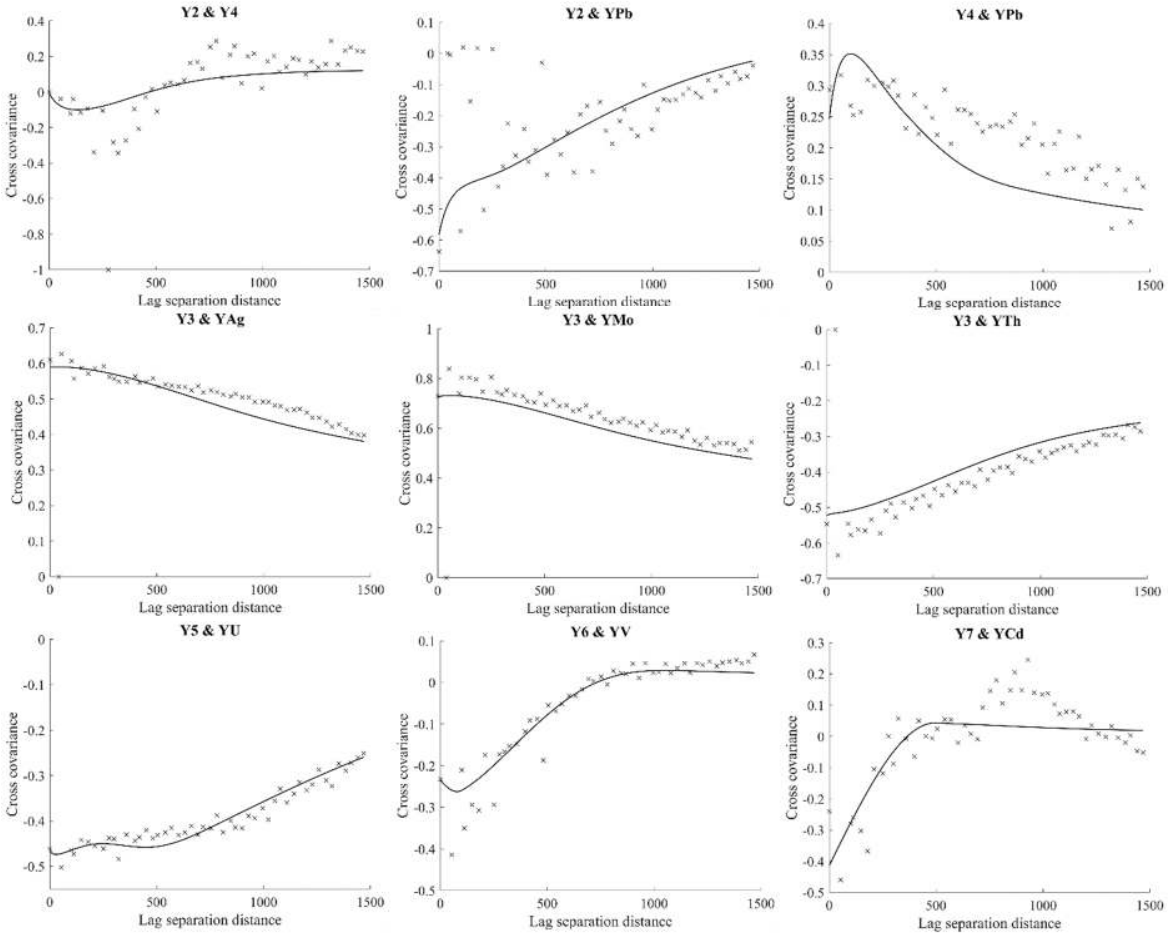

Figure 5.8. Omnidirectional experimental (crosses) and modeled (solid lines) direct and cross-covariances of the Gaussian random fields associated with the lithology indicators and Gaussian transforms of geochemical concentrations. All the modeled covariances are isotropic. For the sake of brevity, the direct and cross-covariances between Gaussian transforms of geochemical concentrations are not represented here

3. Results

With indicator cokriging, the predicted lithology matches the true lithology for 90.38% of the testing data (Fig. 5.9; Table 5.7). The results remain practically unchanged when including the geochemical concentrations as covariates in the cokriging system (here, with a linear model of coregionalization consisting of the direct and cross-variograms of the 9 lithology indicators and the 36 geochemical concentrations), the overall accuracy score being 90.50% in this case (Table 5.8).

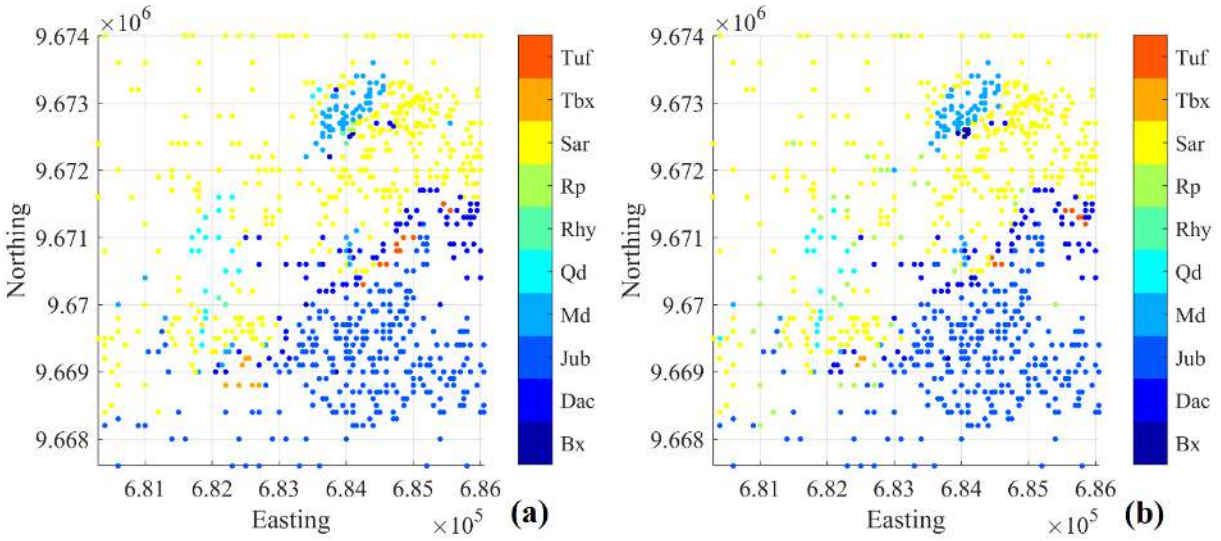

Figure 5.9. (a) Testing data locations with their respective lithology (10 classes); (b) predictive lithological map obtained with indicator cokriging

Table 5.7. Confusion matrix, applying indicator cokriging to determine the most probable lithological classes. The accuracy score is defined as the ratio between the number of correct predictions to the total number of predictions for a given class or overall

Observed lithology	Predicted lithology										Total	Accuracy (%)
	Bx	Dac	Jub	Md	Qd	Rhy	Rp	Sar	Tbx	Tuf		
Bx	**2**	0	2	0	0	0	0	1	0	0	5	40.00%
Dac	1	**61**	2	2	0	0	3	4	0	5	78	78.21%
Jub	1	0	**250**	0	0	0	2	0	0	0	253	98.81%
Md	0	2	0	**60**	0	0	2	3	0	0	67	89.55%
Qd	0	0	0	0	**15**	0	3	2	0	0	20	75.00%
Rhy	1	0	0	0	0	**1**	1	1	0	0	4	25.00%
Rp	2	0	0	1	0	0	**0**	0	0	0	3	0.00%
Sar	0	0	0	3	2	0	20	**328**	0	0	353	92.92%
Tbx	0	0	1	0	0	0	3	0	**3**	0	7	42.86%
Tuf	0	5	2	0	0	0	0	0	0	**3**	10	30.00%
Total	**7**	**68**	**257**	**66**	**17**	**1**	**34**	**339**	**3**	**8**	**800**	**90.38%**

Table 5.8. Confusion matrix, applying indicator cokriging with geochemical concentrations as covariates to determine the most probable lithological classes

Observed lithology	Predicted lithology										Total	Accuracy (%)
	Bx	Dac	Jub	Md	Qd	Rhy	Rp	Sar	Tbx	Tuf		
Bx	**2**	0	2	0	0	0	0	1	0	0	5	40.00%
Dac	1	**61**	2	2	0	0	3	4	0	5	78	78.21%
Jub	1	0	**251**	0	0	0	1	0	0	0	253	99.21%
Md	0	2	0	**60**	0	0	2	3	0	0	67	89.55%
Qd	0	0	0	0	**14**	0	4	2	0	0	20	70.00%
Rhy	1	0	0	0	0	**1**	1	1	0	0	4	25.00%
Rp	2	0	0	1	0	0	**0**	0	0	0	3	0.00%
Sar	0	1	0	3	2	0	18	**329**	0	0	353	93.20%
Tbx	0	0	1	0	0	0	3	0	**3**	0	7	42.86%
Tuf	0	5	2	0	0	0	0	0	0	**3**	10	30.00%
Total	**7**	**69**	**258**	**66**	**16**	**1**	**32**	**340**	**3**	**8**	***800***	**90.50%**

In contrast, plurigaussian simulation reaches an accuracy score of 92.38% (Fig. 5.10a; Table 5.9). When incorporating the geochemical information, this score increases to 92.63% (Fig. 5.10b; Table 5.10), with only a few changes in the predictive lithological map with respect to the map obtained without the geochemical covariates. These changes mainly concern the Qd, Rhy Sar and Dac lithological classes, the three latter having similar textural characteristics and being part of the undifferentiated Saraguro Group of felsic volcanic rocks.

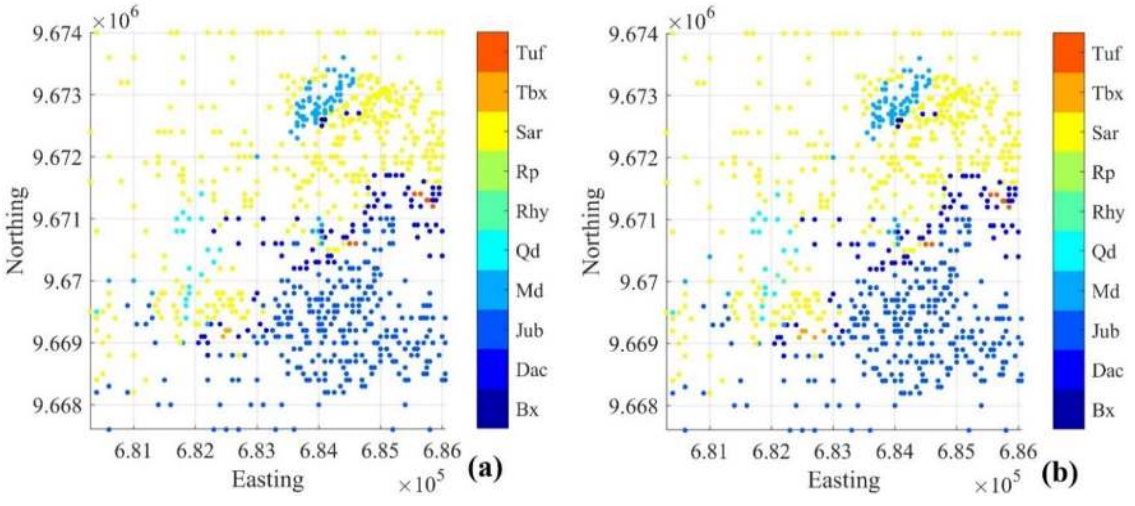

Figure 5.10. Predictive lithological map obtained with a set of 300 plurigaussian realizations, without (a) and with (b) the use of geochemical covariates

Table 5.9. Confusion matrix, applying plurigaussian modeling to determine the most probable lithological classes

Observed lithology	Predicted lithology (most likely classification)										Total	Accuracy (%)
	Bx	Dac	Jub	Md	Qd	Rhy	Rp	Sar	Tbx	Tuf		
Bx	**1**	0	2	0	0	0	1	1	0	0	5	20.00%
Dac	0	**66**	3	2	0	0	0	3	0	4	78	84.62%
Jub	0	0	**252**	0	0	0	0	1	0	0	253	99.60%
Md	0	2	0	**59**	0	0	0	6	0	0	67	88.06%
Qd	0	1	0	0	**14**	0	0	5	0	0	20	70.00%
Rhy	0	0	0	1	1	**0**	0	2	0	0	4	0.00%
Rp	2	0	0	1	0	0	**0**	0	0	0	3	0.00%
Sar	0	5	1	3	3	0	0	**341**	0	0	353	96.60%
Tbx	0	1	2	0	0	0	0	1	**3**	0	7	42.86%
Tuf	0	4	3	0	0	0	0	0	0	**3**	10	30.00%
Total	**3**	**79**	**263**	**66**	**18**	**0**	**1**	**360**	**3**	**7**	**800**	**92.38%**

Table 5.10. Confusion matrix, applying plurigaussian modeling with geochemical covariates to determine the most probable lithological classes

Observed lithology	Predicted lithology (most likely classification)										Total	Accuracy (%)
	Bx	Dac	Jub	Md	Qd	Rhy	Rp	Sar	Tbx	Tuf		
Bx	**1**	0	2	0	0	0	0	2	0	0	5	20.00%
Dac	0	**63**	4	2	0	0	0	3	2	4	78	80.77%
Jub	0	0	**251**	0	0	0	0	1	1	0	253	99.21%
Md	0	2	0	**59**	0	0	0	6	0	0	67	88.06%
Qd	0	0	0	0	**15**	0	0	5	0	0	20	75.00%
Rhy	1	0	0	0	0	**2**	0	1	0	0	4	50.00%
Rp	2	0	0	1	0	0	**0**	0	0	0	3	0.00%
Sar	0	3	1	3	2	0	0	**344**	0	0	353	97.45%
Tbx	0	1	2	0	0	0	0	1	**3**	0	7	42.86%
Tuf	0	4	3	0	0	0	0	0	0	**3**	10	30.00%
Total	**4**	**73**	**263**	**65**	**17**	**2**	**0**	**363**	**6**	**7**	**800**	**92.63%**

4. Discussion

4.1. Indicator cokriging vs plurigaussian predictions

Using indicator cokriging to predict lithology, there is little variation in the accuracy scores for the two methodologies carried out (with and without geochemical covariates), which are in the order

of 90%. The lithological information (primary variable) at the training data screens out the influence of the geochemical information (covariates), which can be explained because all the variables are equally sampled (isotopic sampling design) and ordinary cokriging tends to give a little contribution to the covariates owing to the unbiasedness restriction (Goovaerts, 1997).

Furthermore, the indicator approach suffers from consistency problems at two stages:

(1) In the variogram modeling stage, a linear model of coregionalization is used. Although this model yields positive semidefinite covariance matrices, there is no guarantee that it is valid to represent indicator (binary) variables, insofar as the class of valid indicator covariances is a subset of the class of positive semidefinite functions (Armstrong, 1992; Chilès and Delfiner, 2012). In particular, the spherical model is likely not to be valid for an indicator variable in the two-dimensional space (Emery, 2010).
(2) The indicator predictions obtained by cokriging are not genuine probabilities, as they may be negative or greater than 1. Out of the 8000 predictions held (800 locations times 10 indicators), 242 were found to be negative, with a minimum of -0.073, and 26 were found to be greater than 1, with a maximum of 1.074. These inconsistencies affect 110 of the testing data locations (13.75% of the 800 testing data). Even if this does not prevent selecting the lithological class with the greatest cokriged indicator as the final prediction, one is not aware of the probability of a mistaken prediction and cannot draw probability maps of the lithological classes.

Although the plurigaussian model is more difficult to implement than indicator cokriging (see Section 4.4 next), it improves the accuracy score by about 2%, which is not negligible. Furthermore, the model is internally consistent (there is no issue with the validity of the chosen covariance functions) and agrees with several features of the lithological classes, in particular, their contact relationships and their chronology, through the choice of the truncation rule.

4.2. Abundant vs scarce lithological classes

Overall, the most abundant lithological classes (Sar, Jub, Dac, Md and Qd, which represent 96.8% of the data) are well predicted with both indicator cokriging and plurigaussian simulation, with accuracy scores above 70% (Tables 5.7 to 5.10) and a good match between the predictive maps and the interpreted field map (Fig. 5.9 and 5.10).

In contrast, the accuracy scores for the scarce lithological classes (Rp, Rhy, Bx, Tbx and Tuf, which represent 3.2% of the data) are systematically less than 50%, and even 0% for Rp and, sometimes, for Rhy. This suggests that the scarce classes are, to a great extent, unpredictable. At each testing data location, their probabilities of occurrence are, most often, low: when a scarce class is assigned to this location, it means that none of the classes has a high probability and that the assignation is subject to a high uncertainty. Some modifications of the assignation strategy can be introduced to reproduce the global proportions of the scarce lithological classes (Soares, 1992), but there is no reason to believe that they will improve the accuracy scores of Tables 7 to 10. A more promising alternative is to map not only the most probable lithological class, but also the class probabilities, in order to make the practitioner aware of the possibility of finding each class at each unsampled location. This alternative is possible with the plurigaussian approach that provides genuine probabilities (Fig. 5.11).

4.3. Incorporation of geochemical covariates

The point biserial correlation matrix (Fig. 2b), the truncation rule (Fig. 6b) and the cross-covariances (Fig. 5.8) bring out several associations between lithological classes and geochemical concentrations: Qd and Tbx with Pb concentration, Jub with Ag, Mo and Th concentrations, Tuf and Dac with U concentration, Md with V concentration, and Sar with Cd concentration. Such associations can be explained by the geochemical and physical processes operating within the Earth crust and mantle. In particular, mobile (U) and immobile (Th) incompatible trace elements (Goldschmidt, 1937) are concentrated in primary magmas are later incorporated into mineral phases that make up the continental crust (Taylor and McLennan, 1985); U can migrate in aqueous fluids, and Th in silicate molten (Condie and Aster, 2010). These incompatible trace elements are therefore characteristic components of felsic volcanic rocks such as the Jub, Tuf and Dac lithological classes. Pb concentrations are associated with intermediate to acidic granitic rocks (Hart et al., 2004, Rottura et al., 1991) such as tonalite porphyry (Qd) and rhyolite breccia tuffs (Tbx). The dacite and andesite rocks of the Sar lithology are related to hydrothermal alterations and to highly volatile chalcophile elements such as Cd resulting from supergene alteration (Lodders, 2003).

Despite the previous associations, the incorporation of geochemical covariates does not significantly improve the plurigaussian predictions (Tables 5.9 and 5.10). This can be explained because the covariates are known at the same training locations as the lithological classes (isotopic sampling) and because the spatial auto-correlations of $Y_1 \ldots Y_7$ (Fig. 5.7) are much stronger than their cross-correlations with the geochemical covariates (Fig. 5.8): the geochemistry brings secondary information that, to a great extent, is screened out by the lithological information, as it also happened with indicator cokriging. Perhaps in cases when the vegetation prevents a direct observation of the lithology (inaccessible bedrock) or when the weathering and alteration of the bedrock make the identification of the lithological class ambiguous, which would lead to a heterotopic sampling design, the geochemical information (fully sampled data) could improve the indicator cokriging and plurigaussian predictions based only on lithological information (undersampled data).

The mixed plurigaussian-multigaussian approach is however useful to provide not only a prediction or a simulation of the lithological classes, but also a simulation of the geochemical concentrations that reproduces the spatial cross-dependence between geochemistry and lithology. In the present case based on the variable simplification (Section 2.4.2.2), one simulates the lithology jointly with the U, V, Cd, Pb, Ag, Mo and Th concentrations, not only the lithology alone as in the plurigaussian model of Section 2.4.1. As mentioned in the introductory section, both the lithology and the geochemistry are relevant for vectoring ore in mineral prospection and initial geological-mining exploration, so a consistent joint model unifying both types of information contributes to a better mineral prospectivity mapping.

4.4. Problem dimensionality

The high number of lithological classes (10) and geochemical element concentrations (36) complicate the implementation of the proposed approaches, in particular, in what refers to the fitting of the spatial correlation structure and to the spatial prediction.

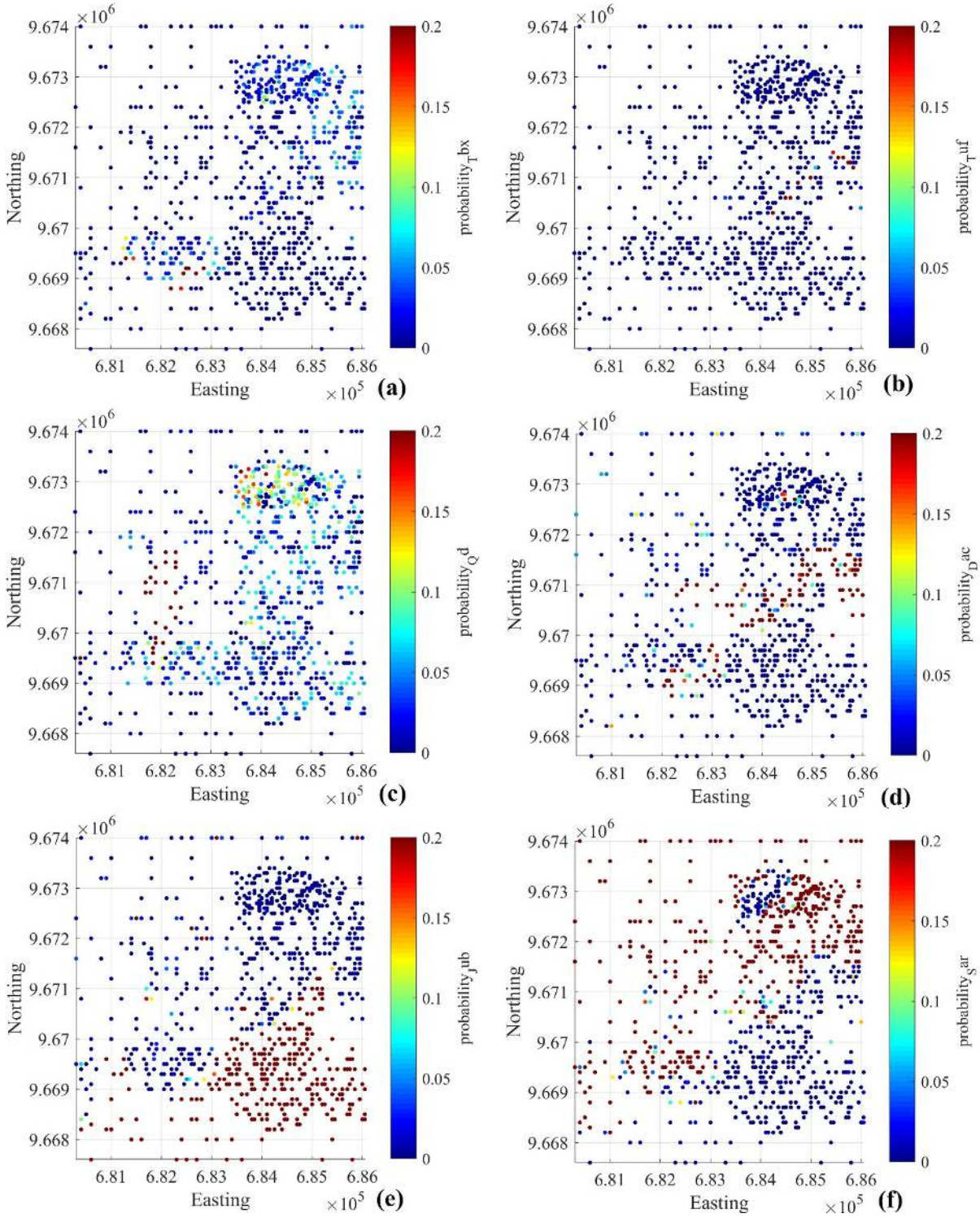

Figure 5.11. Probability maps estimated from a set of 300 plurigaussian realizations: (a) Rhyolite Breccia Tuffs (Tbx), (b) White fine Tuffs (Tuf), (c) Quartz diorite (Qd), (d) Tuffaceous sandstone (Dac), (e) Tuff with crystals (Jub), (f) Dacite/rhyolite (Sar)

Indicator cokriging works with 9 variables when considering only the lithological information, and 45 variables when adding the geochemical information. Apart from the use of a moving neighborhood during cokriging, this high number of variables implies a major effort in the fitting of a linear model of coregionalization. However, one can reach a satisfactory fitting by recourse to the semi-automated algorithm of Goulard and Voltz (1992), mainly because the behavior of all the variables (geochemical concentrations and lithological class indicators) is similar at short distances, exhibiting a nugget effect and/or a linear growth. The linear model of coregionalization is versatile enough to fit the correlation ranges and other structural features, such as the cross-dimple between the indicators I_2 and I_{10} (Fig. 5.5).

Things turn out to be more complicated for the mixed plurigaussian-multigaussian model (Section 2.4.2.2). A first issue is the lack of commercial software for plurigaussian simulation based on more than two underlying Gaussian random fields, reason for which proper programs, building on previous academic codes (Xu et al., 2006; Emery, 2007), had to be developed and adapted to account for quantitative covariates. Secondly, the calculated experimental direct and cross-covariances exhibit different short-scale behaviors, being quadratic for several underlying Gaussian random fields (in particular, Y_3, Y_5 and Y_6) (Fig. 5.7) and linear or linear with nugget for others. Despite many attempts, it has not been possible to satisfactorily fit the 43 direct covariances and the 903 cross-covariances with a linear model of coregionalization, reason for which the model had to be reduced by considering only a subset of the geochemical concentrations. This opens the need for more flexible covariance modeling strategies, for instance based on linear combinations of multivariate Matérn models, for which the behavior near the origin can vary from one variable to another (Gneiting et al., 2010); the validity conditions for this model are, however, still unknown to a great extent and the few sufficient conditions that have been established so far (Gneiting et al., 2010; Apanasovich et al., 2012) are restrictive. A third issue is the prohibitive computational burden to cosimulate the lithological classes and geochemical covariates at all the testing data locations, because the Gibbs sampler (Section 2.4.1.5) requires the use of a unique neighborhood to generate a random vector with the desired Gaussian distribution (Emery et al., 2014). For this reason, simulation has been performed at each testing data location separately, conditionally to a subset of neighboring training data. The model therefore provides the probabilities of the lithological classes at each target location, but not the joint probabilities over several locations. Talebi et al. (2017) propose a shortcut solution to perform joint simulation with the mixed plurigaussian-multigaussian model in the presence of many variables, by only considering the categorical data as conditioning information in the Gibbs sampler; this solution is approximate as it omits the influence of the quantitative covariates on the random fields associated with the categories.

5. Conclusions

Two geostatistical approaches are proposed to predict the lithological class at an unsampled location, based on lithological and geochemical information available at surface samples. The first approach relies on the cokriging of the lithological class indicators, with or without the geochemical covariates, and leads to an estimation of the probability of occurrence of each class, from which the most probable class can be selected as the final prediction. Although its accuracy score is high (around 90.5%), this approach suffers from consistency problems, with about 14% of the testing data having at least one negative estimated probability for some lithological class.

The second approach, consisting of a plurigaussian modeling of the lithological class, does not suffer from such problems and increases the accuracy score by about 2%. This approach also allows the incorporation of the geochemical covariates, so that it is able to provide joint simulations of both the lithology and geochemistry at any unsampled location, which is valuable for mineral prospectivity mapping. Because of the high dimensionality of the problem, some simplifications in the spatial correlation modeling (by the selection of a subset of 7 out of the 36 geochemical covariates) are necessary to obtain a satisfactory fit. This motivates the search for more general coregionalization model that are able to account for varied short-scale behaviors, correlation ranges and other structural features such as cross-dimples, and fitting strategies that can be used in highly multivariate contexts arising in geochemical exploration studies.

Acknowledgments

This research was funded by the National Agency for Research and Development of Chile, through grant ANID/CONICYT PIA AFB180004, by the Ministry of Higher Education, Science, Technology and Innovation of Ecuador (SENESCYT), through scholarship program "Open Call 2012 Second Phase" of the government of Ecuador, and by the Particular Technical University of Loja-Ecuador.

6. References

Adeli, A., Emery, X., 2021. Geostatistical simulation of rock physical and geochemical properties with spatial filtering and its application to predictive geological mapping. Journal of Geochemical Exploration, 220: 106661.

Adeli, A., Emery, X., Dowd, P., 2018. Geological modelling and validation of geological interpretations via simulation and classification of quantitative covariates. Minerals 8(1): 7.

Afzal, P., Fadakar Alghalandis, Y., Moarefvand, P., Rashidnejad Omran, N., Asadi Haroni, H., 2012. Application of power-spectrum–volume fractal method for detecting hypogene, supergene enrichment, leached and barren zones in Kahang Cu porphyry deposit, Central Iran. Journal of Geochemical Exploration 112: 131-138.

Afzal, P., Eskandarnejad Tehrani, M., Ghaderi, M., Hosseini, M.R., 2016. Delineation of supergene enrichment, hypogene and oxidation zones utilizing staged factor analysis and fractal modeling in Takht-e-Gonbad porphyry deposit, SE Iran. Journal of Geochemical Exploration 161: 119-127.

Apanasovich, T.V., Genton, M.G., Sun, Y., 2012. A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components. Journal of the American Statistical Association 107(497): 180–193.

Armstrong, M., 1992. Positive definiteness is not enough. Mathematical Geology 24(1): 135–143.

Armstrong, M., Galli, A., Beucher, H., Le Loc'h, G., Renard, D., Doligez, B., Eschard, R., Geffroy, F., 2011. Plurigaussian Simulations in Geosciences. Springer, Berlin.

Baldock, J.W., 1982. Geología del Ecuador. Boletín de Explicación del Mapa Geológico (1:1.000.000) de la República del Ecuador. Resource document, Ministerio de Recursos Naturales y Energéticos, Quito, 54 pp.

Cameron, E.M., Hamilton, S.M., Leybourne, M.I., Hall, G.E.M., McClenaghan, M.B., 2004. Finding deeply buried deposits using geochemistry. Geochemistry: Exploration, Environment, Analysis 4, 7–32.

Carranza, E.J.M., 2008. Geochemical Anomaly and Mineral Prospectivity Mapping in GIS. Elsevier, Amsterdam.

Castillo, P.I.C., Townley, B.K., Emery, X., Puig, A.F., Deckart, K., 2015. Soil gas geochemical exploration in covered terrains of northern Chile: data processing techniques and interpretation of contrast anomalies. Geochemistry: Exploration, Environment, Analysis 15(2–3): 222–233.

Chilès, J.P., Delfiner, P., 2012. Geostatistics: Modeling Spatial Uncertainty. Wiley, New York.

Cohen, D.R., Kelley, D.L., Anand, R., Coker, W.B., 2010. Major advances in exploration geochemistry, 1998-2007. Geochemistry: Exploration, Environment, Analysis 10(1): 3–16.

Condie, K.C., Aster, R.C., 2010. Episodic zircon age spectra of orogenic granitoids: The supercontinent connection and continental growth. Precambrian Research 180: 227–236.

Dowd, P.A., 1993. Geological and structural control in kriging. In: Soares, A. (ed.) Geostatistics Tróia'92. Kluwer Academic, Dordrecht, pp. 923–936.

Dowd, P.A., 1994. Geological controls in the geostatistical simulation of hydrocarbon reservoirs. Arabian Journal for Science and Engineering 19(2B): 237–247.

Dubrule, O., 1983. Cross-validation of kriging in a unique neighborhood. Journal of the International Association for Mathematical Geology 15(6): 687–699.

Emery, X., 2007. Simulation of geological domains using the plurigaussian model: New developments and computer programs. Computers & Geosciences 33(9): 1189–1201.

Emery, X., 2010. On the existence of mosaic and indicator random fields with spherical, circular, and triangular variograms. Mathematical Geosciences 42: 969–984.

Emery, X., Arroyo, D., Peláez, P., 2014. Simulating large Gaussian random vectors subject to inequality constraints by Gibbs sampling. Mathematical Geosciences 46(3): 265–283.

Emery, X., Cornejo, J., 2010. Truncated Gaussian simulation of discrete-valued, ordinal coregionalized variables. Computers & Geosciences 36(10): 1325–1338.

Emery, X., González, K.E., 2007. Probabilistic modelling of mineralogical domains and its application to resources evaluation. Journal of the South African Institute of Mining and Metallurgy 107(12): 803–809.

Emery, X., Silva, D.A., 2009. Conditional co-simulation of continuous and categorical variables for geostatistical applications. Computers & Geosciences 35(6): 1234–1246.

Giraud, G., 2014. Introduction to High-Dimensional Statistics. Chapman and Hall/CRC.

Gneiting, T., Kleiber, W., Schlather, M., 2010. Matérn cross-covariance functions for multivariate random fields. Journal of the American Statistical Association 105(491): 1167–1177.

Goldschmidt, V.M., 1937. The principles of distribution of chemical elements in mineral and rocks. Journal of the Chemical Society: 655–672.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, Oxford.

Goulard, M., Voltz, M., 1992. Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. Mathematical Geology 24(3): 269–286.

Govett, G.J.S., 1983. Handbook of Exploration Geochemistry vol. 3: Rock Geochemistry in Mineral Exploration. Elsevier, Amsterdam.

Grunsky, E.C., 2010. The interpretation of geochemical survey data. Geochemistry: Exploration, Environment, Analysis 10(1): 27–74.

Grunsky, E.C., Mueller, U.A., Corrigan, D., 2014. A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: applications for predictive geological mapping. Journal of Geochemical Exploration 141: 15–41.

Guartán, J.A., Emery, X., 2021. Regionalized classification of geochemical data with filtering of measurement noises for predictive lithological mapping. Natural Resources Research, in press.

Hart, T.R., Gibson, H.L., Lesher, C.M., 2004. Trace element geochemistry and petrogenesis of felsic volcanic rocks associated with volcanogenic massive Cu-Zn-Pb sulfide deposits. Economic Geology 99(5): 1003–1013.

Isaaks, E.H., Srivastava, R.M., 1988. Spatial continuity measures for probabilistic and deterministic geostatistics. Mathematical Geology 20: 313–341.

Jiménez-Espinosa, R., Chica-Olmo, M., 1999. Application of geostatistics to identify gold-rich areas in the Finisterre-Fervenza region, NW Spain. Applied Geochemistry 14: 133–145.

Jiménez-Espinosa, R., Sousa, A.J., Chica-Olmo, M., 1993. Identification of geochemical anomalies using principal component analysis and factorial kriging analysis. Journal of Geochemical Exploration 46(3): 245–256.

Kasmaee, S., Raspa, G., de Fouquet, C., Tinti, F., Bonduà, S., Bruno, R., 2019. Geostatistical estimation of multi-domain deposits with transitional boundaries: A sensitivity study for the Sechahun iron mine. Minerals 9(2): 115.

Larrondo, P., Leuangthong, O., Deutsch, C.V., 2004. Grade estimation in multiple rock types using a linear model of coregionalization for soft boundaries. In: Magri, E., Ortiz, J., Knights, P., Henríquez, F., Vera, M., Barahona, C. (eds.) Proceedings of the First International Conference on Mining Innovation. Gecamin Ltda, Santiago, Chile, pp. 187–196.

Lodders, K., 2003. Solar system abundances and condensation temperatures of the elements. The Astrophysical Journal 591: 1220–1247.

Madani, N., Emery, X., 2015. Simulation of geo-domains accounting for chronology and contact relationships: application to the Río Blanco copper deposit. Stochastic Environmental Research and Risk Assessment 29: 2173–2191.

Maleki, M., Emery, X., 2015. Joint simulation of grade and rock type in a stratabound copper deposit. Mathematical Geosciences 47: 471–495.

Maleki, M., Emery, X., 2020. Geostatistics in the presence of geological boundaries: exploratory tools for contact analysis. Ore Geology Reviews 120: 103397.

Mariethoz, G., Caers, J., 2014. Multiple-point Geostatistics: Stochastic Modeling with Training Images. Wiley, New York.

Matheron, G., 1989. The internal consistency of models in geostatistics. In: Armstrong, M. (ed.) Geostatistics. Springer, Dordrecht, pp. 21–38.

Mery, N., Emery, X., Cáceres, A., Ribeiro, D., Cunha, E., 2017. Geostatistical modeling of the geological uncertainty in an iron ore deposit. Ore Geology Reviews 88: 336–351.

Olea, R.A., 1999. Geostatistics for Engineers and Earth Scientists. Springer, New York.

Ortiz, J.M., Emery, X., 2006. Geostatistical estimation of mineral resources with soft geological boundaries: a comparative study. Journal of the South African Institute of Mining and Metallurgy 106(8): 577–584.

Paithankar, A., Chatterjee, S., 2018. Grade and tonnage uncertainty analysis of an African copper deposit using multiple-point geostatistics and sequential Gaussian simulation. Natural Resources Research 27: 419–436.

Reis, A.P., Sousa, A.J., Fonseca, E.C., 2003. Application of geostatistical methods in gold geochemical anomalies identification (Montemor-O-Novo, Portugal). Journal of Geochemical Exploration 77(1): 45–63.

Rottura, A., Del Moro, A., Pinarelli, L., Petrini, R., Peccerillo, A., Caggianelli, A., Piccarreta, G., 1991. Relationships between intermediate and acidic rocks in orogenic granitoid suites: petrological, geochemical and isotopic (Sr, Nd, Pb) data from Capo Vaticano (southern Calabria, Italy). Chemical Geology 92(1–3): 153–176.

Royer, J.J., Mejia, P., Caumon, G., Collon, P., 2015. 3D and 4D geomodelling applied to mineral resources exploration – an introduction. In: Weihed, P. (ed.) 3D, 4D and Predictive modelling of Major Mineral Belts in Europe. Springer, Cham, pp. 73–89.

Sandjivy, L., 1984. The factorial kriging analysis of regionalized data – Its application to geochemical prospecting. In: Verly, G., Journel, A.G., Maréchal, A. (eds.) Geostatistics for Natural Resources Characterization. Reidel, Dordrecht, pp. 559–571.

Séguret, S.A., 2013. Analysis and estimation of multi-unit deposits: application to a porphyry copper deposit. Mathematical Geosciences 45: 927–947.

Sillitoe, R.H., 2003. Iron oxide-copper-gold deposits: an Andean view. Mineralium Deposita 38(7): 787–812.

Soares, A., 1992. Geostatistical estimation of multi-phase structures. Mathematical Geology 24(2): 148–160.

Solow, A., 1986. Mapping by simple indicator kriging. Mathematical Geology 18(3): 335–352.

Talebi, H., Hosseinzadeh-Sabeti, E., Azadi, M., Emery, X., 2016. Risk quantification with combined use of lithological and grade simulations: Application to a porphyry copper deposit. Ore Geology Reviews 75: 42–51.

Talebi, H., Lo, J., Mueller, U., 2017. A hybrid model for joint simulation of high-dimensional continuous and categorical variables. In: Gómez-Hernández, J., Rodrigo-Ilarri, J., Rodrigo-Clavero, M., Cassiraga, E., Vargas-Guzmán, J. (eds.) Geostatistics Valencia 2016. Springer, Cham, pp. 415–430.

Talebi, H., Mueller, U., Tolosana-Delgado, R., Grunsky, E.C., McKinley, J.M., de Caritat, P., 2019a. Surficial and deep earth material prediction from geochemical compositions. Natural Resources Research 28: 869–891.

Talebi, H., Mueller, U., Tolosana-Delgado, R., van den Boogaart, K.G., 2019b. Geostatistical simulation of geochemical compositions in the presence of multiple geological units: application to mineral resource evaluation. Mathematical Geosciences 51(2): 129–153.

Taylor, S.R., McLennan, S.M., 1985. The Continental Crust: Its composition and evolution: an examination of the geochemical record preserved in sedimentary rocks. Blackwell Science, Oxford.

Urqueta, E., Kyser, T.K., Clark, A.H., Stanley, C.R., Oates, C.J., 2009. Lithogeochemistry of the Collahuasi porphyry Cu-Mo and epithermal Cu-Ag (-Au) cluster, northern Chile: Pearce element ratio vectors to ore. Geochemistry: Exploration, Environment, Analysis 9(1): 9–17.

Vargas-Guzmán, J.A., 2008. Transitive geostatistics for stepwise modeling across boundaries between rock regions. Mathematical Geosciences 40(8): 861–873.

Wackernagel, H., 2003. Multivariate Geostatistics: An Introduction with Applications. Springer, Berlin.

Wackernagel, H., Sanguinetti, H., 1993. Gold prospecting with factorial cokriging in the Limousin, France. In: Davis, J.C., Herzfeld, U.C. (eds.) Computers in Geology: 25 Years of Progress. Oxford University Press, Oxford, pp. 33–43.

Wang, C., Carranza, E.J.M., Zhang, S., Zhang, J., Liu, X., Zhang, D., Duan, C., 2013. Characterization of primary geochemical haloes for gold exploration at the Huanxiangwa gold deposit, China. Journal of Geochemical Exploration 124: 40–58.

Wilford, J., Caritat, P. de, Bui, E., 2016. Predictive geochemical mapping using environmental correlation. Applied Geochemistry 66: 275–288.

Xu, C.S., Dowd, P.A., Mardia, K.V., Fowell, R.J., 2006. A flexible true plurigaussian code for spatial facies simulations. Computers & Geosciences 32(10): 1629–1645.

Zuo, R., Wang, J., 2016. Fractal/multifractal modeling of geochemical data: A review. Journal of Geochemical Exploration 164: 33–41.

Capitulo 6. Conclusiones

Los problemas de clasificación regionalizada se presentan en etapas de prospección, exploración temprana y exploración avanzada para identificar los diversos tipos de variables involucradas en los procesos geológicos que permitan mejorar los modelos interpretativos de un depósito, con la finalidad de mejorar la identificación y evaluación de recursos y la planificación geológica-minera. La incorporación de técnicas de estadística, aprendizaje automático y geoestadística permite interpretar los procesos geológicos y geoquímicos a través de una clasificación regionalizada de datos de muestreo, cuyos resultados ayudan a vectorizar la prospección mineral en las etapas iniciales de la exploración del recurso mineral.

En general, los enfoques basados en estadística multivariable, aprendizaje automático y/o minería de datos no aprovechan la correlación espacial de los datos de muestreo, realizando la clasificación a partir de la relación colocalizada de dependencia entre variables cuantitativas (geoquímica) y categóricas (geología). Este trabajo de tesis demuestra, a través de dos casos de estudio, que esta clasificación puede mejorar significativamente al utilizar técnicas geoestadísticas.

La primera propuesta consiste realizar la clasificación a partir de la relación de dependencia entre la variable categórica (tipo de roca o dominio geológico) y un conjunto de covariables cuantitativas (concentraciones geoquímicas). La mejora respecto a los enfoques tradicionales se logra a través del modelamiento de estas últimas en el espacio y se basa en los siguientes aspectos: primero, la clasificación se realiza en numerosos escenarios simulados de las concentraciones geoquímicas, que emulan la variabilidad espacial, quedando como predicción final en cada sitio del espacio la clase que más se repite a través del conjunto de escenarios. Segundo, se puede filtrar la variabilidad de pequeña escala (efecto pepita), correspondiente a ruido y errores de medición, en los escenarios simulados de las concentraciones geoquímicas, permitiendo realzar la dependencia con la variable categórica y mejorar las puntuaciones de clasificación. Esta mejora sugiere que la variabilidad de pequeña escala es una fuente de variabilidad introducida estadística y espacialmente independiente de las propiedades geológicas tales como el tipo de roca, el tipo de alteración o la zona mineral, es decir, que la asociación entre la geología y la geoquímica se manifiesta solamente a través de las componentes de variabilidad de gran escala.

En la segunda propuesta, la clasificación se logra principalmente a través de la simulación directa de la variable categórica, es decir, la predicción depende en mayor parte de los valores categóricos observados en sitios adyacentes. Esta segunda propuesta resulta ser de más difícil implementación, dada la mayor complejidad del modelo de funciones aleatorias categóricas (plurigaussiano), en comparación con el modelo de funciones aleatorias cuantitativas (multigaussiano) requerido en la primera propuesta.

Si bien mejora poco las puntuaciones de la clasificación, la incorporación de las concentraciones geoquímicas como covariables entrega escenarios que permiten determinar asociaciones y explicar los procesos geoquímicos y físicos que ocurrieron dentro de la corteza terrestre, así como las relaciones entre las concentraciones geoquímicas con los procesos geológicos, siendo valioso para el mapeo de prospección mineral y para la toma de decisiones (identificación de blancos de prospección, diseño de campañas de muestreo para explorar y delinear recursos minerales, etc.). Ahora bien, este enfoque integral, en el cual se cosimulan las variables categóricas y cuantitativas con un modelo mixto plurigaussiano-multigaussiano, se enfrenta a dificultades prácticas, tanto en el modelamiento variográfico, debido a la gran cantidad de covarianzas o variogramas directos y

cruzados a ajustar y las características (formas y alcances) diferentes de dichos variogramas, como en la generación de simulaciones condicionales con métodos iterativos (muestreador de Gibbs, cuya implementación correcta requiere el uso de una vecindad única). En este contexto, algunas simplificaciones (selección de las covariables más relevantes y restricción de la simulación a los datos condicionantes más cercanos) fueron necesarias para poder llevar a cabo la propuesta integral de cosimulación. Lo anterior pone en evidencia la existencia de un *trade-off* entre, por una parte, el costo asociado a la mayor complejidad del modelo y la necesidad de simplicaciones en el ajuste variográfico y en la cosimulación, y, por otra parte, la riqueza de los entregables, consistentes en escenarios simulados de la litología y de la geoquímica, que reproducen la inter-dependencia entre ambas fuentes de información y permiten mejorar el análisis espacial geólogico-minero para la definición de áreas de interés prospectivo.

Entre las ventajas de las dos metodologías propuestas, destaca el hecho de que ambas entregan un conjunto de clasificaciones (tantas como realizaciones de las variables cuantitativas o categóricas hayan), permitiendo medir la incertidumbre en la clasificación, mediante mapas de probabilidad, en sitios donde no se conoce los procesos geológicos ni las concentraciones geoquímicas. Asimismo, aunque esta situación no se dio en los casos de estudio, ambas propuestas son aplicables a bases de datos heterotópicas, cuando no se conocen todas las variables en todos los sitios con datos, mientras que las técnicas de análisis de datos convencionales suelen considerar solamente los datos en donde todas las variables han sido medidas y descartar los otros datos.

Las principales perspectivas que dejan este trabajo son:

1) El diseño de modelos variográficos multivariables que permiten el ajuste de covarianzas o variogramas directos y cruzados con formas y alcances diferentes, situación que se observó en el segundo caso de estudio al momento de correlacionar las variables cuantitativas con la variable categórica. El modelo lineal de corregionalización mostró severas limitaciones en este aspecto. En la actualidad, existen pocas alternativas, siendo una de ellas el modelo Matérn multivariable (Gneiting et al., 2010), pero aun se tienen condiciones restrictivas sobre los parámetros que entregan un modelo válido (Apanasovich et al., 2012).
2) El diseño de algoritmos para condicionar las simulaciones geoestadísticas a grandes bases de datos, ya sea por el gran número de muestras o por el gran número de variables. De particular interés es el muestreador de Gibbs que, por su naturaleza iterativa, requiere una vecindad única para converger a la distribución condicional deseada (Emery et al., 2014), siendo prohibitivo en tiempos de cálculo y requerimientos de memoria en casos altamente multivariables.
3) El diseño de modelos multi-categóricos, que permitan simular conjuntamente los tipos de roca, alteración y mineral (variables categóricas inter-dependientes), incorporando también las concentraciones geoquímicas como covariables. Siguiendo el trabajo de Madani and Emery (2017), se podría pensar incluso en modelos no estacionarios de modo de reproducir zonaciones geológicas.

Bibliografía

Adeli, A., Emery, X., 2020. Geostatistical simulation of rock physical and geochemical properties with spatial filtering and its application to predictive geological mapping. Journal of Geochemical Exploration, in press.

Adeli, A., Emery, X., Dowd, P., 2018. Geological modelling and validation of geological interpretations via simulation and classification of quantitative covariates. Minerals 8(1), 7.

Afzal, P., Fadakar Alghalandis, Y., Moarefvand, P., Rashidnejad Omran, N., Asadi Haroni, H., 2012. Application of power-spectrum–volume fractal method for detecting hypogene, supergene enrichment, leached and barren zones in Kahang Cu porphyry deposit, Central Iran. Journal of Geochemical Exploration 112, 131–138.

Afzal, P., Eskandarnejad Tehrani, M., Ghaderi, M., Hosseini, M.R., 2016. Delineation of supergene enrichment, hypogene and oxidation zones utilizing staged factor analysis and fractal modeling in Takht-e-Gonbad porphyry deposit, SE Iran. Journal of Geochemical Exploration 161, 119–127.

Alabert, F., Massonnat, G.J., 1990. Heterogeneity in a complex turbiditic reservoir: Stochastic modelling of facies and petrophysical variability. In: Proceedings of the Sixty Fifth Annual Technical Conference and Exhibition, SPE Paper no. 20604. New Orleans, pp. 775–790.

Armstrong, M., 1992. Positive definiteness is not enough. Mathematical Geology 24(1), 135–143.

Armstrong, M., Galli, A., Beucher, H., Le Loc'h, G., Renard, D., Doligez, B., Eschard, R., Geffroy, F., 2011. Plurigaussian Simulations in Geosciences. 2nd revised edition. Springer, Berlin, 176 pp.

Apanasovich, T.V., Genton, M.G., Sun, Y., 2012. A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components. Journal of the American Statistical Association 107(497), 180–193.

Bahar, A., Kelkar, M., 2000. Journey from well logs/cores to integrated geological and petrophysical properties simulation: a methodology and application. SPE Reservoir Evaluation and Engineering 3(5), 444–456.

Baldock, J.W., 1982. Geología del Ecuador. Boletín de Explicación del Mapa Geológico (1:1.000.000) de la República del Ecuador. Resource document, Ministerio de Recursos Naturales y Energéticos, Quito, 54 pp.

Barbosa, P., Oliveira, T., Silva, J., 2010. Regionalized classification of multivariate geochemical data from Jacupiranga Alkaline Complex (Ribeira de Iguape Valley/Sao Paulo, Brazil). Revista Brasileira de Geociencias 40(2), 212–219.

Barnett, R.M., Manchuk, J.G., & Deutsch, C.V., 2013. Projection pursuit multivariate transform. Mathematical Geosciences, 46(3), 337–359.

Bohling, G.C., 1997. GSLIB-style programs for discriminant analysis and regionalized classification. Computers & Geosciences 23(7), 739–761.

Breiman, L., 2001. Random forests. Machine Learning 45(1), 5032.

Cáceres, A., Emery, X., 2010. Conditional co-simulation of copper grades and lithofacies in the Río Blanco-Los Bronces copper deposit. In: Castro, R., Emery, X., Kuyvenhoven, R. (eds.), Proceedings of the IV International Conference on Mining Innovation MININ 2010. Gecamin Ltda, Santiago, Chile, pp. 311–320.

Camacho, J., Ferrer, A., 2014. Cross-validation in PCA models with the element-wise ekf algorithm: practical aspects, Chemom. Intell. Lab. Syst. 131, 37–50.

Cameron, E.M., Hamilton, S.M., Leybourne, M.I., Hall, G.E.M., McClenaghan, B., 2004. Finding deeply buried deposits using geochemistry: Geochemistry: Exploration, Environment, Analysis 4, 7–32.

Cannell, J., Cooke, D.R., Walshe, J.L., Stein, H., 2005. Geology, mineralization, alteration, and structural evolution of the El Teniente porphyry Cu–Mo deposit. Economic Geology 100, 979–1003.

Caritat, P. de, Grunsky, E.C., 2013. Defining element associations and inferring geological processes from total element concentrations in Australian catchment outlet sediments: multivariate analysis of continental scale geochemical data. Applied Geochemistry 33, 104–126.

Carranza, E.J.M., 2009. Geochemical Anomaly and Mineral Prospectivity Mapping in GIS. Amsterdam: Elsevier.

Carrasco, P., 2010. Nugget effect, artificial or natural? Journal of the Southern African Institute of Mining and Metallurgy 110(6), 299–305.

Casella, G., George, E., 1992. Explaining the Gibbs Sampler. The American Statistician 46(3), 167–174.

Castillo, P.I.C., Townley, B.K., Emery, X., Puig, A.F., Deckart, K., 2015. Soil gas geochemical exploration in covered terrains of northern Chile: data processing techniques and interpretation of contrast anomalies. Geochemistry: Exploration, Environment, Analysis 15(2–3): 222–233.

Chanderman, L., Dohm, C.E., Minnitt, R.C.A., 2017. 3D geological modelling and resource estimation for a gold deposit in Mali. Journal of the Southern African Institute of Mining and Metallurgy 117(2), 189–197.

Chentsov, N.N., 1957. Lévy Brownian motion for several parameters and generalized white noise. Theory of Probability and Its Applications 2(2), 265–266.

Chilès, J.P., Delfiner, P., 2012. Geostatistics: Modeling Spatial Uncertainty. Wiley, New York.

Cohen, D.R., Kelley, D., Anand, R., Coker, W.B., 2010. Major advances in exploration geochemistry, 1998-2007. Geochemistry: Exploration, Environment, Analysis 10, 3–16.

Condie, K.C., Aster, R.C., 2010. Episodic zircon age spectra of orogenic granitoids: The supercontinent connection and continental growth. Precambrian Research 180, 227–236.

Darsow, A., Schafmeister, M.T., Hofmann, T., 2009. An ArcGIS approach to include tectonic structures in point data regionalisation. Ground Water 47(4), 591-597.

Deer, W.A., Howie, R.A., Zussman, J., 1996. An introduction to rock forming minerals, 2nd ed. Prentice Hall, New York. Vols. I–V.

Debon, F., Lemmet, M., 1999. Evolution of Mg/Fe ratios in the Late Variscan plutonic rocks from the external crystalline massifs of the Alps (France, Italy, Switzerland). Journal of Petrology 40, 1151–1185.

Desassis, N., Renard, D., 2013. Automatic variogram modeling by iterative least squares: univariate and multivariate cases. Mathematical Geosciences 45(4), 453–470.

Deutsch, C.V., Journel, A.G., 1998. GSLIB: Geostatistical Software Library and User's Guide. Oxford University Press, New York.

Dowd, P.A., 1993. Geological and structural control in kriging. In: Soares, A. (ed.) Geostatistics Tróia'92. Kluwer Academic, Dordrecht, pp. 923–936.

Dowd, P.A., 1994. Geological controls in the geostatistical simulation of hydrocarbon reservoirs. Arabian Journal for Science and Engineering 19 (2B), 237–247.

Dowd, P.A., 1997. Structural controls in the geostatistical simulation of mineral deposits. In: Baafi, E.Y., Schofield, N.A. (eds.), Geostatistics Wollongong'96. Kluwer Academic, Dordrecht, pp. 647–657.

Dowd, P.A., Pardo-Igúzquiza, E., Xu, C., 2003. Plurigau: a computer program for simulating spatial facies using the truncated plurigaussian method. Computers & Geosciences 29(2), 123–141.

Dubrule, O., 1983. Cross-validation of kriging in a unique neighborhood. Journal of the International Association for Mathematical Geology 15(6), 687–699.

Dubrule, O., 1993. Introducing more geology in stochastic reservoir modelling. In: Soares, A. (ed.), Geostatistics Tróia'92. Kluwer Academic, Dordrecht, pp. 351–369.

Duke, J.H., Hanna, P.J., 2001. Geological interpretation for resource modelling and estimation. In: Edwards, A.C. (ed.), Mineral Resource and Ore Reserve Estimation —The AusIMM Guide to Good Practice. Australasian Institute of Mining and Metallurgy: Melbourne, Australia, pp. 147–156.

Emery, X., 2007. Simulation of geological domains using the plurigaussian model: New developments and computer programs. Computers & Geosciences 33(9), 1189–1201.

Emery, X., 2008. A turning bands program for conditional co-simulation of cross-correlated Gaussian random fields. Computers & Geosciences 34(12), 1850–1862.

Emery X., 2010. Iterative algorithms for fitting a linear model of coregionalization. Computers & Geosciences 36(9), 1150–1160.

Emery, X., 2010. On the existence of mosaic and indicator random fields with spherical, circular, and triangular variograms. Mathematical Geosciences 42, 969–984.

Emery, X., Arroyo, D., Peláez, P., 2014. Simulating large Gaussian random vectors subject to inequality constraints by Gibbs sampling. Mathematical Geosciences 46(3), 265–283.

Emery, X., Arroyo, D., Porcu, E., 2016. An improved spectral turning-bands algorithm for simulating stationary vector Gaussian random fields. Stochastic Environmental Research and Risk Assessment 30(7), 1863–1873.

Emery, X., Cornejo, J., 2010. Truncated Gaussian simulation of discrete-valued, ordinal coregionalized variables. Computers & Geosciences 36(10), 1325–1338.

Emery, X., González, K.E., 2007. Probabilistic modelling of mineralogical domains and its application to resources evaluation. Journal of the South African Institute of Mining and Metallurgy 107(12), 803–809.

Emery, X., Lantuéjoul, C., 2006. TBSIM: A computer program for conditional simulation of three-dimensional Gaussian random fields via the turning bands method. Computers & Geosciences 32(10), 1615–1628.

Emery, X., Maleki, M., 2019. Geostatistics in the presence of geological boundaries: Application to mineral resources modeling. Ore Geology Reviews 114, article 103124.

Emery, X., Séguret, S.A., 2020. Geoestadística de Yacimientos de Cobre Chilenos – 35 Años de Investigación Aplicada. Caligrama, Sevilla.

Emery, X., Silva, D., 2009. Conditional co-simulation of continuous and categorical variables for geostatistical applications. Computers & Geosciences 35, 1234–1246.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179-188.

Freulon, X., de Fouquet, C., Rivoirard, J., 1990. Simulation of the geometry and grades of a uranium deposit using a geological variable. In: Proceedings of the XXII International Symposium on Applications of Computers and Operations Research in the Mineral Industry, Technische Universitat Berlin, Berlin, pp. 649–659.

Fouedjio, F., Hill, E.J., Laukamp, C., 2018. Geostatistical clustering as an aid for ore body domaining: case study at the Rocklea Dome channel iron ore deposit, Western Australia. Applied Earth Science (Trans. Inst. Min. Metall. B) 127(1), 15–29.

Galli, A., Beucher, H., Le Loc'h, G., Doligez, B., 1994. The pros and cons of the truncated Gaussian method. In: Armstrong, M., Dowd, P.A. (eds.) Geostatistical simulations. Kluwer, Dordrecht, pp. 217–233.

Galli, A., Gerdil-Neuillet, F., Dadou, C., 1984. Factorial kriging analysis: a substitute to spectral analysis of magnetic data. In: Verly, G., David, M., Journel, A.G., Maréchal, A. (eds.), Geostatistics for Natural Resources Characterization. Dordrecht: Reidel, pp. 543–557.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6, 721–741.

Glazner, A.F., Bartley, J.M., Coleman, D.S., Gray, W., Taylor, Z.T., 2004. Are plutons assembled over millions of years by amalgamation from small magma chambers? GSA Today 14, 4–11.

Gneiting, T., Kleiber, W., Schlather, M., 2010. Matérn cross-covariance functions for multivariate random fields. Journal of the American Statistical Association 105(491), 1167–1177.

Goldschmidt, V.M., 1937. The principles of distribution of chemical elements in mineral and rocks. Journal of the Chemical Society, 655–672.

Goovaerts, P., 1992. Factorial kriging analysis: a useful tool for exploring the structure of multivariate spatial soil information. Journal of Soil Science 43, 597–619.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York, 480 pp.

Goulard, M., Voltz, M., 1992. Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. Mathematical Geology 24(3), 269–286.

Govett, G.J.S., 1983. Handbook of Exploration Geochemistry vol. 3: Rock Geochemistry in Mineral Exploration. Elsevier, Amsterdam.

Gringarten, E., Deutsch, C.V., 2001. Teacher's aide: variogram interpretation and modelling. Mathematical Geology 33(4), 507–534.

Grunsky, E.C., Smee, B.W., 1999. The differentiation of soil types and mineralization from multi-element geochemistry using multivariate methods and digital topography. Journal of Geochemical Exploration 67, 287–299.

Grunsky, E.C., 2010. The interpretation of geochemical survey data. Geochemistry: Exploration, Environment and Analysis 10, 27–74.

Grunsky, E.C., Corrigan, D., Mueller, U.A., Bonham-Carter, G.F., 2012. Predictive geologic mapping using lake sediment geochemistry in the Melville Peninsula. Geological Survey of Canada, Open File 7171. http://dx.doi.org/10.4095/291901 (1 sheet).

Grunsky, E.C., Mueller, U.A., Corrigan, D., 2014. A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: applications for predictive geological mapping. Journal of Geochemical Exploration 141:15–41.

Guartán, J.A., Emery, X., 2020. Regionalized classification of geochemical data with filtering of measurement noises for predictive lithological mapping. Natural Resources Research, in press.

Guilbert, J., Park, C., 1986, The Geology of Ore Deposits: W.H. Freeman and Co., New York, 985 pp.

Haldorsen, H.H., Damsleth, E., 1990. Stochastic modeling. Journal of Petroleum Technology 42, 404–412.

Hamilton, W.B., 1998. Archean magmatism and deformation were not products of plate tectonics. Precambrian Res. 91, 143–179.

Hardle, W., Simar, L., 2007. Applied Multivariate Statistical Analysis. Berlin: Springer.

Harff, J., Davis, J.C., 1990. Regionalization in geology by multivariate classification. Mathematical Geology 22(5), 573–588.

Hart, T.R., Gibson, H.L., Lesher, C.M., 2004. Trace element geochemistry and petrogenesis of felsic volcanic rocks associated with volcanogenic massive Cu-Zn-Pb sulfide deposits. Economic Geology 99(5), 1003–1013.

Hassan, A.E., Bekhit, H.M., Chapman, J.B., 2009. Using Markov Chain Monte Carlo to quantify parameter uncertainty and its effect on predictions of a groundwater flow model. Environmental Modelling and Software 24(6), 749–763.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009, The elements of statistical learning: Data mining, inference and prediction: Springer Series in Statistics.

Hofmann, T., Darsow, A., Schafmeister, M.T., 2010. Importance of the nugget effect in variography on modeling zinc leaching from a contaminated site using simulated annealing. Journal of Hydrology 389(1-2), 78–89.

Ibarguren, I., Lasarguren, A., Pérez, J.M., Muguerza, J., Arbelaitz, O., Gurrutxaga, I., 2016. BFPART: Best-first PART. Information Sciences 367–368, 927–952.

Isaaks, E.H., Srivastava, R.M., 1988. Spatial continuity measures for probabilistic and deterministic geostatistics. Mathematical Geology 20, 313–341.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Computing Surveys, 31(3), 264-323.

Jaquet, O., 1989. Factorial kriging analysis applied to geological data from petroleum exploration. Mathematical Geology 21(7), 683–691.

Jenny, H., 1941. Factors of Soil Formation: a System of Quantitative Pedology. McGraw Hill, New York, 281 pp.

Jiménez-Espinosa, R., Chica-Olmo, M., 1999. Application of geostatistics to identify gold-rich areas in the Finisterre-Fervenza region, NW Spain. Applied Geochemistry 14, 133–145.

Jiménez-Espinosa, R., Sousa, A.J., Chica-Olmo, M., 1993. Identification of geochemical anomalies using principal component analysis and factorial kriging analysis. Journal of Geochemical Exploration 46(3), 245–256.

Journel, A.G., Huijbregts, C.J., 1978. Mining geostatistics. Academic Press, London.

Kasmaee, S., Raspa, G., de Fouquet, C., Tinti, F., Bonduà, S., Bruno, R., 2019. Geostatistical estimation of multi-domain deposits with transitional boundaries: A sensitivity study for the Sechahun iron mine. Minerals 9(2), 115.

Kass, G.V., 1980. An Exploratory Technique for Investigating Large Quantities of Categorical Data. Applied Statistics 29(2), 119.

Kesler, S.E., 1973. Copper, molybdenum and gold abundances in porphyry copper deposits. Economic Geology 68, 106–112.

Kuhn, S., Cracknell, M.J., Reading, A.M., 2018. Lithologic mapping using random forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia. Geophysics 83(4).

Kuhn, S., Cracknell, M.J., Reading, A.M., 2019. Lithological mapping in the Central African Copper Belt using random forests and clustering: strategies for optimised results. Ore Geology Reviews 112, 103015.

Lantuéjoul, C., 2002. Geostatistical Simulation, Models and Algorithms. Springer, Berlin

Larocque, G., Dutilleul, P., Pelletier, B., Fyles, J.W., 2006. Conditional Gaussian co-simulation of regionalized components of soil variation. Geoderma 134, 1–16.

Larrondo, P., Leuangthong, O., Deutsch, C.V., 2004. Grade estimation in multiple rock types using a linear model of coregionalization for soft boundaries. In: Magri, E., Ortiz, J., Knights, P., Henríquez, F., Vera, M., Barahona, C. (eds.) Proceedings of the First International Conference on Mining Innovation. Gecamin Ltda, Santiago, Chile, pp. 187–196.

Leal-Pacheco, F.A., Barbosa-Landim, P.M., 2005. Two-way regionalized classification of multivariate datasets and its application to the assessment of hydrodynamic dispersion. Mathematical Geology 37(4), 393–417.

Le Loc'h, G., Beucher, H., Galli, A., Doligez, B., 1994. Improvement in the truncated Gaussian method: combining several Gaussian functions. In: ECMOR IV, Fourth European Conference on the Mathematics of Oil Recovery. Røros, Norway, 13pp. (unpublished conference proceedings).

Le Loc'h, G., Galli, A., 1997. Truncated plurigaussian method: theoretical and practical points of view. In: Baafi, E.Y., Schofield, N.A. (eds.), Geostatistics Wollongong'96. Kluwer Academic, Dordrecht, pp. 211–222.

Le Maitre, R.W., 1976. The chemical variability of some common igneous rocks. J. Petrol. 17, 589–637.

Le Maitre, R.W. ed. 1989. A classification of igneous rocks and glossary of terms. Oxford: Blackwell Scientific.

Leuangthong, O., Deutsch, C.V., 2003. Stepwise conditional transformation for simulation of multiple variables. Mathematical Geology 35(2), 155–173.

Leuangthong, O., Khan, K.D., Deutsch, C.V., 2008. Solved problems in Geostatistics. John Wiley & Sons, Hoboken, NJ, 208 pp.

Liu, Y., Carranza, E.J.M., Zhou, K.F., Xia, Q.L., 2019. Compositional balance analysis: an elegant method of geochemical pattern recognition and anomaly mapping for mineral exploration. Natural Resources Research 28, 1269–1283.

Lodders, K., 2003. Solar system abundances and condensation temperatures of the elements. The Astrophysical Journal 591, 1220–1247.

Lozano-Santacruz, R., Verma, S.P., Girón, P., Velasco- Tapia, F., Morán-Zenteno, D., Viera, F., Chávez, G. 1995. Calibración preliminar de fluorescencia de rayos X para análisis cuantitativo de elementos mayores en rocas ígneas. Actas INAGEQ 1: 203–208.

Madani, N., Emery, X., 2015. Simulation of geo-domains accounting for chronology and contact relationships: Application to the Río Blanco copper deposit. Stochastic Environmental Research and Risk Assessment 29(8), 2173–2191.

Madani, N., Emery, X., 2017. Plurigaussian modeling of geological domains based on the truncation of non-stationary Gaussian random fields. Stochastic Environmental Research and Risk Assessment 31(4), 893–913.

Maleki, M., Emery, X., 2015. Joint simulation of grade and rock type in a stratabound copper deposit. Mathematical Geosciences 47, 471–495.

Maleki, M., Emery, X., 2017. Joint simulation of stationary grade and non-stationary rock type for quantifying geological uncertainy in a copper deposit. Computers and Geosciences 109, 258–267.

Maleki, M., Emery, X., Cáceres, A., Ribeiro, D., Cunha, E., 2016. Quantifying the uncertainty in the spatial layout of rock type domains in an iron ore deposit. Computational Geosciences 20(5), 1013–1028.

Mariethoz, G., Caers, J., 2014. Multiple-point Geostatistics: Stochastic Modeling with Training Images. Wiley, New York.

Mariethoz, G., Renard, P., Cornaton, F., Jaquet. O., 2009. Truncated plurigaussian simulations of aquifer heterogeneity. Ground Water 47(1), 13024.

Matheron, G., 1962. Traité de Géostatistique Appliquée. Paris: Technip.

Matheron, G., 1963. Principles of Geostatistics. Economic Geology 58(8), 1246–1266.

Matheron, G., 1965. Les Variables Régionalisées et leur Estimation. Une Application de la Théorie des Fonctions Aléatoires aux Sciences de la Nature. Masson, Paris, 306 p.

Matheron, G. 1973. The intrinsic randon functions and their applications. Advances in Applied Probability 5(3), 439–468.

Matheron, G., 1989. The internal consistency of models in geostatistics. In: Armstrong, M. (ed.) Geostatistics. Springer, Dordrecht, pp. 21–38.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

McKay, G., Harris, J.R., 2016. Comparison of the data-driven random forests model and a knowledge-driven method for mineral prospectivity mapping: a case study for gold deposits around the Huritz Group and Nueltin Suite, Nunavut, Canada. Natural Resources Research 25(2), 125–143.

Merian, E., Anke, M., Ihnat, M., Stoeppler, M., 2004. Elements and Their Compounds in the Environment - Occurrence, Analysis and Biological Relevance. New York: Wiley.

Mery, N., Emery, X., Cáceres, A., Ribeiro, D., Cunha, E., 2017. Geostatistical modeling of the geological uncertainty in an iron ore deposit. Ore Geology Reviews 88, 336–351.

Meyer, C., Hemley, J.J., 1967. Wall rock alteration. In: Barnes, H.L. (ed.) Geochemistry of Hydrothermal Ore Deposits, pp. 166–235. New York, NY: Holt, Rinehart, and Winston.

Mitchell, T.M., 1997. Decision Tree Learning. Singapore: WCB/McGraw-Hill Inc.

Mueller, U.A., Grunsky, E.C., 2016. Multivariate spatial analysis of lake sediment geochemical data; Melville Peninsula, Nunavut, Canada. Applied Geochemistry 75, 247–262.

Natrella, M., 2010. NIST/SEMATECH e-Handbook of Statistical Methods. NIST/SEMATECH.

Olea, R.A., 1999. Geostatistics for Engineers and Earth Scientists. Kluwer Academic Publishers, Norwell, Massachusetts, USA.

Ortiz, J.M., Emery, X., 2006. Geostatistical estimation of mineral resources with soft geological boundaries: a comparative study. Journal of the South African Institute of Mining and Metallurgy 106(8), 577–584.

Paithankar, A., Chatterjee, S., 2018. Grade and tonnage uncertainty analysis of an African copper deposit using multiple-point geostatistics and sequential Gaussian simulation. Natural Resources Research 27, 419–436.

Poldevaart, A., 1955. Chemistry of the Earth Crust. In: Poldevaart, A. (ed.). Crust of the Earth. A Symposium. Geol. Soc. Am. Spec. Paper 62: 119–144.

Quinlan, J.R., 1993. C4.5, Programs for Machine Learning. San Mateo: Morgan Kaufmann.

Reis, A.P., Sousa, A.J., Fonseca, E.C., 2003. Application of geostatistical methods in gold geochemical anomalies identification (Montemor-O-Novo, Portugal). Journal of Geochemical Exploration 77(1), 45–63.

Richards, J.P., 2009. Postsubduction porphyry Cu–Au and epithermal Au deposits: Products of remelting of subduction-modified lithosphere. Geology 37, 247–250.

Robb, L.J., 2011. Introduction to Ore-Forming Processes. Blackwell Publishing

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews 71, 804–818.

Ronov, A.B., Yaroshevsky, A.A., 1976. A new model for the chemical structure of the Earth´s Crust. Geochem. Int. 13, 89–121.

Rose, A.W., Hawkes, H.E., Webb, J.S., 1979. Geochemistry in Mineral Exploration, 2nd ed.: Academic Press, New York, 657 pp.

Rossi, M.E., Deutsch, C.V., 2014. Mineral Resource Estimation. Springer Heidelberg, 332 pp.

Rottura, A., Del Moro, A., Pinarelli, L., Petrini, R., Peccerillo, A., Caggianelli, A., Piccarreta, G., 1991. Relationships between intermediate and acidic rocks in orogenic granitoid suites: petrological, geochemical and isotopic (Sr, Nd, Pb) data from Capo Vaticano (southern Calabria, Italy). Chemical Geology 92(1–3), 153–176.

Royer, J.J., Mejia, P., Caumon, G., Collon, P., 2015. 3D and 4D geomodelling applied to mineral resources exploration – an introduction. In: Weihed, P. (ed.) 3D, 4D and Predictive modelling of Major Mineral Belts in Europe. Springer, Cham, pp. 73–89.

Salminen, R., Batista, M.J., Bidovec, M., Demetriades, A., De Vivo, B., De Vos, W., Duris, M., Gilucis, A., Gregorauskiene, V., Halamic, J., Heitzmann, P., Lima, A., Jordan, G., Klaver, G., Klein, P., Lis, J., Locutara, J., Marsina, K., Mazreku, A., O´ Connor, P.J., Olsson, S.À., Ottesen, R.-T., Plant, J.A., Reeder, S., Salpeeur, I., Sandström, H., Siewers, U., Steenfelt, A., Travainen, T., 2005. Geochemical Atlas of Europe. Espoo: Geological Survey of Finland.

Sandjivy, L., 1984. The factorial kriging analysis of regionalized data - its application to geochemical prospecting. In: Verly, G., Journel, A.G., Maréchal, A. (eds.) Geostatistics for Natural Resources Characterization. Dordrecht: Reidel, pp. 559–571.

Seedorff, E., Dilles, J.H., Proffett, J.M. Jr., Einaudi, M.T., Zurcher, L., Stavast, W.J.A., Johnson, D.A., Barton, M.D., 2005. Porphyry deposits – Characteristics and origin of hypogene features. Society of Economic Geologists, Economic Geology 100th Anniversary Volume, 1905–2005, 251–298.

Séguret, S.A., 2013. Analysis and estimation of multi-unit deposits: application to a porphyry copper deposit. Mathematical Geosciences 45, 927–947.

Sillitoe, R.H., 2003. Iron oxide-copper-gold deposits: an Andean view. Mineralium Deposita 38(7), 787–812.

Sillitoe, R.H., 2010. Porphyry-copper systems. Economic Geology 105, 3–41.

Simpson, E.H., 1949. Measurement of diversity. Nature 163(4148), 688.

Sinclair, A.J., Blackwell, G.H., 2002. Applied Mineral Inventory Estimation. Cambridge University Press: New York, NY, USA.

Smith, K.S., Huyck, H.L.O., 1999, An overview of the abundance, relative mobility, bioavailability, and human toxicity of metals; in Plumlee, G.S., Logsdon, M.J. (eds.), The Environmental Geochemistry of Mineral Deposits, Part A. Processes, Techniques, and Health Issues: Society of Economic Geologists, Reviews in Economic Geology, v. 6A, pp. 29–70.

Soares, A., 1992. Geostatistical estimation of multi-phase structures. Mathematical Geology 24(2), 148–160.

Solow, A., 1986. Mapping by simple indicator kriging. Mathematical Geology 18(3), 335–352.

Stanley, C.R., Sinclair, A.J., 1989. Comparison of probability plots and the gap statistic in the selection of thresholds for exploration geochemistry data. Journal of Geochemical Exploration 32(1-3), 355–357.

Sun, T., Chen, F., Zhong, L.X., Liu, W.M., Wang, Y., 2019. GIS-based mineral prospectivity mapping using machine learning methods: a case study from Tongling ore district, eastern China. Ore Geology Reviews 109, 26–49.

Talebi, H., Hosseinzadeh-Sabeti, E., Azadi, M., Emery, X., 2016. Risk quantification with combined use of lithological and grade simulations: Application to a porphyry copper deposit. Ore Geology Reviews 75, 42–51.

Talebi, H., Lo, J., Mueller, U., 2017. A hybrid model for joint simulation of high-dimensional continuous and categorical variables. In: Gómez-Hernández, J., Rodrigo-Ilarri, J., Rodrigo-Clavero, M., Cassiraga, E., Vargas-Guzmán, J. (eds.) Geostatistics Valencia 2016. Springer, Cham, pp. 415–430.

Talebi, H., Mueller, U., Tolosana-Delgado, R., Grunsky, E.C., McKinley, J.M., de Caritat, P., 2018. Surficial and deep earth material prediction from geochemical compositions – a spatial predictive model. Natural Resources Research 28, 869–891.

Talebi, H., Mueller, U., Tolosana-Delgado, R., van den Boogaart, K.G., 2019. Geostatistical simulation of geochemical compositions in the presence of multiple geological units: application to mineral resource evaluation. Mathematical Geosciences 51(2), 129–153.

Taylor, S.R., McLennan, S.M., 1985. The Continental Crust: Its composition and evolution: an examination of the geochemical record preserved in sedimentary rocks. Blackwell Science, Oxford.

Tercan A.E., Sohrabian B., 2013. Multivariate geostatistical simulation of coal quality data by independent components. International Journal of Coal Geology 112, 53–66.

Thornton, I., 1995, Metals in the global environment—facts and misconceptions: Internatl Council on Metals in the Environment, Ottawa, 103 pp.

Titley, S.R., 1982. The style and progress of mineralization and alteration in porphyry copper systems: American Southwest. In: Titley, S.R. (ed.) Advances in Geology of the Porphyry Copper Deposits. Southwestern North America, pp. 139–184. Tucson, AZ: University of Arizona Press.

Tolosana-Delgado, R., Mueller, U., van den Boogaart, K.G., 2019. Geostatistics for compositional data: an overview. Mathematical Geosciences 51(4), 485–526.

Tolosana-Delgado, R., van den Boogaart, K.G., Pawlowsky-Glahn, V., 2011. Geostatistics for compositions. In: Buccianti, A., Pawlowsky-Glahn, V. (eds.) Compositional Data Analysis. Theory and Applications. Wiley, Chichester, pp. 73–86.

Tukey, J., 1977. Exploratory data analysis. Pearson, London.

Urqueta, E., Kyser, T.K., Clark, A.H., Stanley, C.R., Oates, C.J., 2009. Lithogeochemistry of the Collahuasi porphyry Cu-Mo and epithermal Cu-Ag (-Au) cluster, northern Chile: Pearce element ratio vectors to ore. Geochemistry: Exploration, Environment, Analysis 9(1), 9–17.

U.S. Geological Survey (USGS). 2006. FGDC digital cartographic standard for geologic map symbolization (post script implementation), Survey Techniques and Methods Rep. 11-A2, U.S. Geol. Surv., Reston, Va.

van den Boogaart, K.G., Mueller, U., Tolosana-Delgado, R., 2017. An affine equivariant multivariate normal score transform for compositional data. Mathematical Geosciences 49(2), 231–251.

Vapnik, V.N., 1995. The nature of statistical learning theory: Springer Verlag New York, Inc.

Vapnik, V.N., 1998. Statistical learning theory: John Wiley & Sons, Inc.

Vargas-Guzmán, J.A., 2008. Transitive geostatistics for stepwise modeling across boundaries between rock regions. Mathematical Geosciences 40(8), 861–873.

Wackernagel, H., 1988. Geostatistical techniques for interpreting multivariate spatial information. In: Chung, C.F., Fabbri, A.G., Sinding-Larsen, R. (eds.) Quantitative Analysis of Mineral and Energy Resources. vol. C-223 of NATO ASI Series, Reidel, Dordrecht, pp. 393–409.

Wackernagel, H., 1998. Principal component analysis for autocorrelated data: a geostatistical perspective. Technical Report N022/98/G. Centre de Geostatistique, Ecole de Mines de Paris, Fontainebleau, France (41 pp.)

Wackernagel, H. 2003. Multivariate geostatistics: An introduction with applications. 3rd edition, Springer, New York, 388 p.

Wackernagel, H., Sanguinetti, H., 1993. Gold prospecting with factorial cokriging in the Limousin, France. In: Davis, J.C., Herzfeld, U.C. (eds.) Computers in Geology: 25 Years of Progress. Oxford University Press, Oxford, pp. 33–43.

Wang, C., Carranza, E.J.M., Zhang, S., Zhang, J., Liu, X., Zhang, D., Duan, C., 2013. Characterization of primary geochemical haloes for gold exploration at the Huanxiangwa gold deposit, China. Journal of Geochemical Exploration 124, 40–58.

Waske, B., Benediktsson, J.A., Árnason, K., Sveinsson, J.R., 2009, Mapping of hyperspectral AVIRIS data using machine learning algorithms: Canadian Journal of Remote Sensing 35, S106–S116.

Wilford, J., Caritat, P. de, Bui, E., 2016. Predictive geochemical mapping using environmental correlation. Applied Geochemistry 66, 275–288.

Xiang, J., Xiao, K.Y., Carranza, E.J.M., Chen, J.P., Li, S., 2020. 3D mineral prospectivity mapping with random forests: a case study of Tongling, Anhui, China. Natural Resources Research 29(1), 395-414.

Xu, C., Dowd, P.A., Mardia, K.V., Fowell, R.J., 2006. A flexible true plurigaussian code for spatial facies simulations. Computers & Geosciences 32(10), 1629–1645.

Zuo, R., Wang, J., 2016. Fractal/multifractal modeling of geochemical data: A review. Journal of Geochemical Exploration 164, 33–41.

Zuo, R.G., Xiong, Y.H., 2018. Big data analytics of identifying geochemical anomalies supported by machine learning methods. Natural Resources Research 27(1), 5–13.

Anexo A. Una aplicación de la clasificación mediante simulación y filtraje de ruido a datos de geoquímica de superficie

En este anexo se aplica la propuesta No. 1 del capítulo 3 a los datos de muestras superficiales del capítulo 5. Los contenidos siguientes han sido aceptados para publicación en la revista *Natural Resources Research*:

> Guartán, J.A., Emery, X., 2020. *Regionalized classification of geochemical data with filtering of measurement noises for predictive lithological mapping*. Natural Resources Research, in press. DOI 10.1007/s11053-020-09779-0.

Regionalized classification of geochemical data with filtering of measurement noises for predictive lithological mapping

Abstract

A method for lithological mapping is proposed, which combines the geostatistical simulation of geochemical concentrations with a coregionalization analysis and a decision-tree classification algorithm. The method consists of classifying each target point based on simulated values of the geochemical concentrations, filtered from the short-scale spatial components corresponding to noise and measurement errors. The procedure is repeated over many simulations to finally give as a result the most probable lithology at each target point. An application to a set of geochemical samples of soils and surface rocks is presented, in which the lithology type is recorded from an interpretive geological field map. It shows a significant classification improvement when pre-processing the sampling data through geostatistical simulation with filtering of the nugget effect, with a rate of correctly classified data increasing between 3.5 and 11 percentage points depending on whether on a training or a testing data subset is considered. The lithological prediction allows generating geological maps as complementary activities to the exploration of mineral resources to be able to forecast and/or to validate the geology mapped at each point of the explored areas.

Keywords: geochemistry; geostatistical simulation; coregionalization analysis; nugget effect; decision trees.

1. Introduction

A geological map is the result of the characterization of rocks. The different types of lithology can be studied based on their mineralogy or based on geochemical prospecting, which analyzes the composition of surface or soil sediments (Jenny, 1941). Surface chemistry research at a regional scale is often directed to determine the geochemical background or baseline of simple elements (Stanley and Sinclair, 1989). Surficial geochemical concentrations vary from one location in space to another with a more or less pronounced continuity and can be used to describe local natural conditions such as geology or anthropogenic activities. Multi-element geochemistry therefore provides useful information on the lithology as an essential parameter for the geology of prospective areas (Grunsky et al. 2012).

The application of machine learning provides a systematic framework through which geochemical/geological processes are identified, which would be, among others, the use of classification techniques for mineral prospectivity modeling and lithological mapping (Carranza, 2009; Grunsky, 2010; Grunsky et al., 2014; Rodriguez-Galiano et al., 2015; McKay and Harris, 2016; Zuo and Xiong, 2018; Kuhn et al., 2019; Liu et al., 2019; Sun et al., 2019; Xiang et al., 2020), e.g., linear discriminant analysis, neural networks, support vector machines, decision trees or random forests (Fisher, 1936; Quinlan, 1993; Breiman, 2001; Hastie et al., 2008). These methods work on continuous and categorical variables observed at the sampling locations, but generally do not explicitly account for the joint spatial correlation of the observations when classifying, therefore they are likely to lose information conveyed by the data.

Regionalized variables such as the geochemical concentrations often exhibit spatial dependencies that can be modeled by interpreting these variables as realizations of some hypothetical random

fields (Matheron, 1962; Chilès and Delfiner, 2012). In this setting, the spatial dependencies between regionalized data are characterized by covariance functions or by variograms that measure the average contrast existing between two data based on their separation in space and that can be inferred from the sampling information. A common modeling approach is to represent covariances or variograms as the sum of different components, called nested structures, acting at different spatial scales and associated with different ranges of correlation (Wackernagel, 1988, 2003). In particular, very short-range components (in particular, the well-known nugget effect) often represent noise or measurement errors, which can prejudice the classification of continuous variables, such as the geochemical concentrations, for predicting categorical variables, such as the lithology.

In the present work, the geochemical data used are the result of a point sampling where the analytical errors for heavy metals, in general, are up to 5-10% (Salminen et al., 2005) and may significantly contribute to the total uncertainty. Since modern techniques use very small sample quantities, poorly homogenized field samples can increase the variability of the data (Merian et al., 2004) and generate a nugget effect, that is, a partial lack of spatial correlation at small distances (Gringarten and Deutsch, 2001; Carrasco, 2010). Therefore, geostatistical tools and techniques that allow estimating the short-scale (nugget effect) and large-scale spatial dependencies of geochemical data, building predictive models, quantifying uncertainty, and making regionalized classifications on the basis of available data by accounting for their spatial structure are becoming more and more imperative (Olea, 1999; Gringarten and Deutsch, 2001; Darsow et al., 2009; Hassan et al., 2009; Adeli et al., 2018).

In the following, the classification problem is addressed, which consists of predicting a categorical variable (lithology) knowing continuous variables (geochemical concentrations). Many examples and applications of this problem have been documented (Olea, 1999; Barbosa et. al, 2010; Grunsky, 2010; Grunsky et al., 2012; Grunsky et al., 2014; Adeli et al., 2018; Talebi et. al., 2019). Some of these deal with regionalized data, the spatial correlation of which the researcher can take advantage in order to improve the classification. An important contribution of this work with respect to existing literature is the improvement of the classification in the presence of noise (nugget effect), combining multivariable simulation methodologies with spatial component filtering and machine learning techniques, in view to obtain a predictive lithological model based on geochemical data from surface samples.

2. Materials and methods

2.1. Study area and available data

The study area is located in the southern region of Ecuador. Its dimensions are approximately 5500 meters along the east-west direction, 6500 meters in the north-south direction, with an elevation that ranges from 2490 to 4022 above mean sea level. It is covered by Tertiary volcanic rocks (Baldock, 1982) and composed of an undifferentiated sequence of the Saraguro group of subaerial, calcoalkaline, intermediate to acidic volcanic rocks. Andesitic to dacitic (Sar) rocks predominate, but rhyolitic (Rhy) rocks are common too. They are in a stratigraphic sequence with a rhyolitic-breccia mass flow (Tbx), tuffaceous sandstones (Dac) and white fine tuffs (Tuf) of the Plancharumi formation, and are overlain by crystal tuffs of rhyolitic composition (Jub). Polymictic clast

supported hydrothermal breccias (Bx) showing silicified clay fragments and silica-clay alteration in both matrix and clasts are also present. Moreover, several intrusive bodies occur in the area: a microdiorite intrusive (Md) cut by pyrite and quartz-pyrite veinlets (stockwork), a diorite porphyry intrusive (Qd), with clay-silica alteration and disseminated pyrite, and rhyolite subvolcanic intrusive domes (Rp).

The geochemical database is the result of a surface sampling with horizontal spacing that varies from 100 m × 400 m to 50 m × 50 m (Fig. A.1), obtaining a total of 4060 samples. Exploratory and preparatory data analyses were performed to remove duplicate or erroneous data, so that the database on which the subsequent models will be built contains 3998 samples, with information on the prevailing lithology and the concentrations of 36 elements, reported in percentage (%) for major elements and in part per million (ppm) for trace elements (Table A.1). Ten lithologies (Table A.2) are considered based on an interpretive geological field map (Fig. A.1) prepared by exploration geologists.

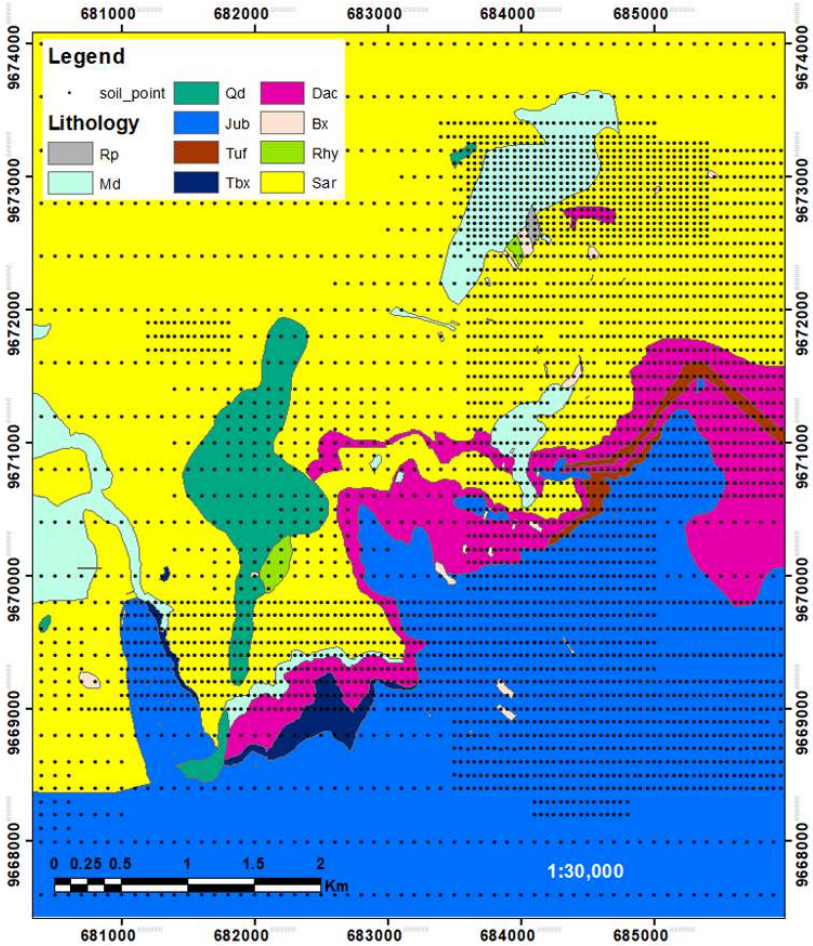

Figure A.1. Location of surface samples on an interpretive geological map of the study area. Sampling is generally on a grid of 100m × 100m, but in given sectors the sampling grid is denser (50m × 50m). The true lithology is observed only at the sample location and is interpreted (hence, subject to error) elsewhere. Source: Mining Company Cornerstone Ecuador S.A., 2012

Table A.8.1. Quantitative variables with their coding and definition

Definition	Variable	Symbol	Variable	Symbol	Variable	Symbol
9 major elements (%)	Iron	Fe	Magnesium	Mg	Sodium	Na
	Calcium	Ca	Titanium	Ti	Potassium	K
	Phosphorus	P	Aluminum	Al	Sulfur	S
27 trace elements (ppm)	Molybdenum	Mo	Uranium	U	Barium	Ba
	Copper	Cu	Thorium	Th	Boron	B
	Lead	Pb	Strontium	Sr	Tungsten	W
	Zinc	Zn	Cadmium	Cd	Scandium	Sc
	Silver	Ag	Antimony	Sb	Thallium	Tl
	Nickel	Ni	Bismuth	Bi	Mercury	Hg
	Cobalt	Co	Vanadium	V	Selenium	Se
	Manganese	Mg	Lanthanum	La	Tellurium	Te
	Arsenic	As	Chromium	Cr	Gallium	Ga

Table A.8.2. Lithological information and global statistics accounting for declustering weights obtained with the cell method (Goovaerts, 1997) using cells of size 400m × 400m

Rock type	Lithology	Codification	Number of Data	Declustered proportion
Hydrothermal Breccia	Bx	1	24	0.35%
Tuffaceous Sandstone (Plancharumi formation)	Dac	2	390	7.36%
Tuff (Jubones Formation)	Jub	3	1274	33.11%
Microdiorite/Andesite Porphyry	Md	4	333	5.22%
Quartz diorite/Tonalite Porphyry	Qd	5	101	3.06%
Rhyolite	Rhy	6	11	0.19%
Rhyolite Porphyry	Rp	7	6	0.06%
Andesite/dacite/rhyolite (Saraguro group)	Sar	8	1774	49.19%
Rhyolite Breccia Tuffs (Plancharumi formation)	Tbx	9	36	0.75%
White fine Tuffs (Plancharumi formation)	Tuf	10	49	0.71%
		Total	***3998***	***100%***

The distributions of concentrations for some geochemical elements (Co, Fe, Ti, Tl) are shown in Fig. A.2. The distributions of all concentrations are unimodal, positively skewed, with a few high extreme values for trace elements such as zinc, manganese, barium, molybdenum, arsenic, thorium, strontium, bismuth, antimony and tellurium. In the cases of sodium, boron and tungsten, some spikes (tied values, especially zeroes) are observed in their histograms, which may arise because of detection limits and/or precision problems in the measurement of the concentrations. The cumulated concentration of all 36 elements never exceeds 26% (the maximum arises for a sample with 24.7% Fe and the other eight major elements adding to 1.22%, while the minimum is 0.69% and the average is only 5.12%), so that the 36 geochemical concentrations under study can be considered as free of compositional or stoichiometric closure constraints and compositional data analysis techniques (e.g., Tolosana-Delgado et al., 2019) are not needed here. The behavior between these variables and the lithologies, as illustrated in the correlation circle of Fig. A.3a

obtained by Principal Component Analysis (PCA), shows the existence of two groups with highly positive correlations, the first one consisting of Zn, Mg, Ca, Co, Sc, Ba, Cr, Cd, W, Ni, Mn, K; the second group gathers As, Se, Mo, Ag, B, Hg, Sb, Fe, Te and Bi. These groups are related to the lithologies (Jub, Qd, Md, Tbx) and (Rp, Rhy, Bx), respectively (Table A.3), as indicated in the biplot of Fig. A.3b constructed with the lithologies as illustrative (supplementary) variables.

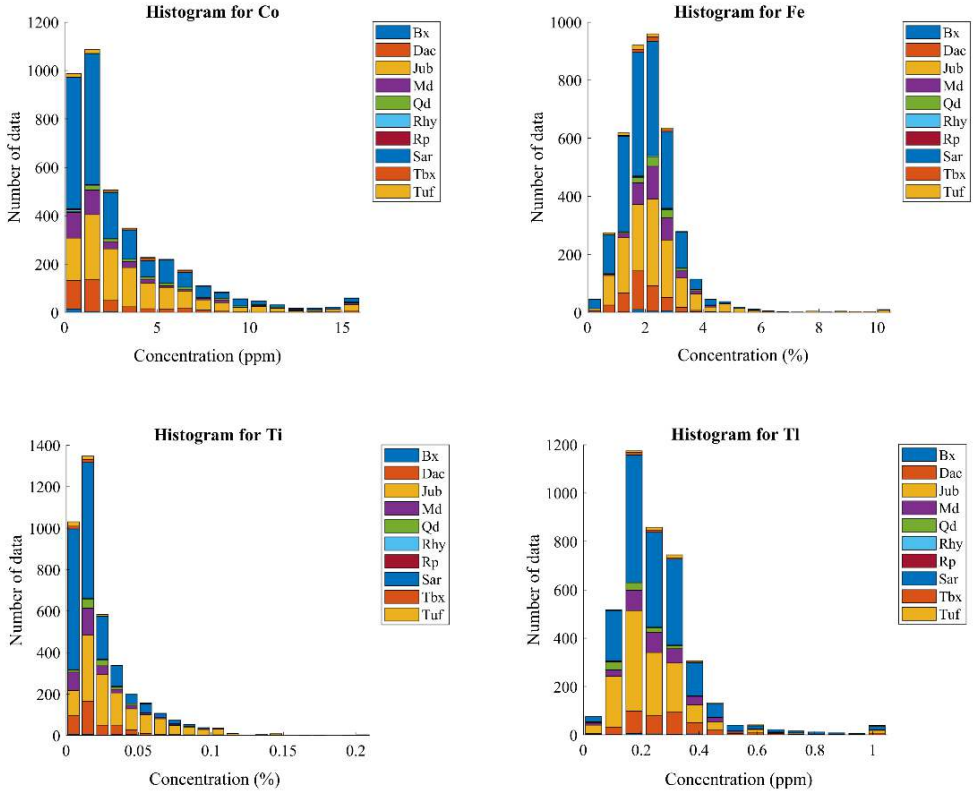

Figure A.2. Experimental distributions of four continuous variables (concentrations of Co, Fe, Ti, Tl)

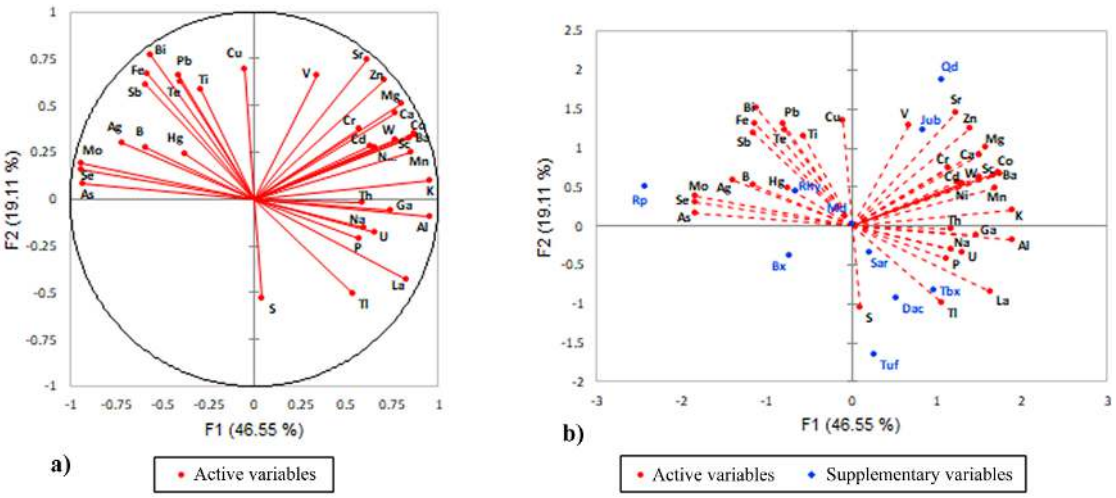

Figure A.3. a. Correlation circle between geochemical variables in the first factorial plane of principal component analysis that explains more than 65% of the total variability of the data; **b**. Correlation biplot between the quantitative variables and lithologies in the study area, where the lithologies were added as illustrative (supplementary) variables

Table A.8.3. Relations between geochemical concentrations and lithologies. The first value in each cell corresponds to the average of a geochemical concentration in a lithology, while the second value corresponds to its standard deviation

Concentrations	Lithology Bx	Lithology Dac	Lithology Jub	Lithology Md	Lithology Qd	Lithology Rhy	Lithology Rp	Lithology Sar	Lithology Tbx	Lithology Tuf
Iron (Fe)	2.347 / 0.63	1.946 / 0.79	2.388 / 1.52	2.458 / 1.05	2.337 / 0.6	2.323 / 0.54	2.681 / 1.10	2.034 / 0.99	2.164 / 0.44	1.903 / 0.97
Calcium (Ca)	0.04 / 0.04	0.097 / 0.15	0.119 / 0.15	0.093 / 0.18	0.232 / 0.27	0.095 / 0.08	0.023 / 0.02	0.12 / 0.22	0.164 / 0.20	0.057 / 0.07
Phosphorus (P)	0.046 / 0.03	0.055 / 0.03	0.045 / 0.03	0.065 / 0.04	0.053 / 0.03	0.051 / 0.01	0.037 / 0.02	0.058 / 0.03	0.074 / 0.03	0.044 / 0.03
Magnesium (Mg)	0.072 / 0.09	0.094 / 0.12	0.176 / 0.14	0.13 / 0.19	0.269 / 0.20	0.1 / 0.08	0.01 / 0.01	0.102 / 0.13	0.193 / 0.09	0.067 / 0.07
Titanium (Ti)	0.021 / 0.02	0.017 / 0.02	0.033 / 0.03	0.018 / 0.02	0.017 / 0.01	0.015 / 0.01	0.027 / 0.01	0.012 / 0.02	0.013 / 0.02	0.01 / 0.01
Aluminium (Al)	1.777 / 0.96	2.463 / 0.94	2.803 / 0.89	2.168 / 0.99	2.274 / 0.85	1.855 / 0.52	0.697 / 0.45	2.132 / 0.94	2.706 / 0.76	2.204 / 0.77
Sodium (Na)	0.006 / 0.01	0.006 / 0.01	0.008 / 0.01	0.006 / 0.01	0.004 / 0.01	0.001 / 0.00	0.003 / 0.00	0.005 / 0.01	0.009 / 0.00	0.004 / 0.00
Potassium (K)	0.046 / 0.04	0.068 / 0.04	0.082 / 0.06	0.062 / 0.04	0.094 / 0.06	0.051 / 0.03	0.011 / 0.01	0.087 / 0.05	0.085 / 0.03	0.064 / 0.04
Sulfur (S)	0.08 / 0.05	0.075 / 0.04	0.064 / 0.05	0.086 / 0.05	0.059 / 0.17	0.06 / 0.03	0.066 / 0.02	0.071 / 0.06	0.079 / 0.07	0.068 / 0.04
Molybdenum (Mo)	21.81 / 28.01	1.999 / 3.21	0.702 / 0.80	9.863 / 14.96	4.891 / 15.37	16.035 / 16.17	65.91 / 51.93	3.249 / 5.46	1.091 / 1.51	1.056 / 0.59
Copper (Cu)	46.348 / 39.84	20.794 / 11.42	19.73 / 14.30	36.904 / 41.60	90.833 / 197.84	36.507 / 11.96	41.693 / 8.20	23.649 / 24.62	16.267 / 4.95	17.623 / 9.74
Lead (Pb)	33.667 / 29.33	16.722 / 8.39	17.441 / 15.26	27.886 / 24.35	48.46 / 101.79	41.223 / 48.92	38.513 / 21.87	31.066 / 134.05	14.599 / 8.09	14.675 / 7.48
Zinc (Zn)	31.038 / 37.18	45.852 / 36.47	87.365 / 76.67	40.065 / 33.69	89.431 / 69.10	41.436 / 35.26	31.471 / 29.15	49.865 / 51.67	57.003 / 21.96	41.647 / 20.02
Silver (Ag)	0.292 / 0.61	0.133 / 0.10	0.109 / 0.15	0.309 / 0.43	0.265 / 0.25	0.688 / 1.68	0.553 / 0.35	0.345 / 1.51	0.179 / 0.25	0.115 / 0.05
Nickel (Ni)	2.346 / 1.51	3.102 / 2.35	3.01 / 1.92	3.796 / 3.52	3.789 / 2.64	2.482 / 0.76	2.214 / 1.63	3.018 / 2.27	3.917 / 1.08	2.078 / 0.97
Cobalt (Co)	1.45 / 1.61	2.578 / 3.17	4.79 / 17.20	2.963 / 4.00	5.769 / 5.32	1.709 / 1.94	0.743 / 0.30	2.727 / 3.84	4.706 / 2.19	3.012 / 4.27
Manganese (Mn)	83.75 / 85.50	265.151 / 603.62	297.401 / 667.93	322.16 / 595.78	570.089 / 437.67	204.545 / 179.67	28.571 / 18.34	335.065 / 507.88	364.583 / 240.56	308.755 / 524.16
Arsenic (As)	35.604 / 31.16	22.212 / 41.68	19.64 / 41.91	34.714 / 25.15	21.73 / 25.50	34.064 / 18.15	56.843 / 21.48	31.361 / 40.86	27.133 / 36.55	21.355 / 38.09
Uranium (U)	0.659 / 0.57	1.001 / 0.72	1.466 / 0.88	0.455 / 0.25	0.62 / 0.63	0.464 / 0.14	0.216 / 0.13	0.536 / 0.48	0.894 / 0.43	1.118 / 0.70

Concentrations	Lithology Bx	Lithology Dac	Lithology Jub	Lithology Md	Lithology Qd	Lithology Rhy	Lithology Rp	Lithology Sar	Lithology Tbx	Lithology Tuf
Thorium (Th)	1.383 / 2.07	1.804 / 1.77	2.987 / 2.55	0.649 / 0.45	1.611 / 1.34	1.018 / 0.57	0.371 / 0.24	1.075 / 0.90	1.139 / 0.72	2.465 / 1.81
Strontium (Sr)	11.971 / 8.75	16.488 / 21.14	25.621 / 29.18	15.429 / 20.61	36.002 / 32.93	18.482 / 12.67	8.2 / 4.32	18.089 / 26.97	14.428 / 11.28	10.837 / 8.93
Cadmium (Cd)	0.089 / 0.14	0.149 / 0.19	0.088 / 0.08	0.163 / 0.21	0.379 / 0.45	0.065 / 0.05	0.036 / 0.05	0.235 / 0.64	0.157 / 0.08	0.139 / 0.11
Antimony (Sb)	1.404 / 1.72	0.729 / 0.91	2.216 / 4.71	1.468 / 1.45	1.139 / 1.42	2.023 / 2.19	2.379 / 0.88	1.224 / 4.04	0.965 / 1.79	0.694 / 0.81
Bismuth (Bi)	0.497 / 0.25	0.455 / 0.40	0.807 / 1.60	0.634 / 0.61	0.735 / 0.96	0.792 / 0.33	1.049 / 0.37	0.579 / 2.90	0.259 / 0.15	0.293 / 0.18
Vanadium (V)	32.125 / 11.03	29.7 / 16.43	47.845 / 20.62	35.955 / 18.23	34.584 / 15.60	33.091 / 12.05	26.571 / 4.83	27.114 / 15.2	31.306 / 9.72	21.367 / 10.92
Lanthanum (La)	5.442 / 3.23	11.85 / 7.06	11.773 / 6.08	7.335 / 4.53	8.613 / 4.76	6.455 / 2.81	2.214 / 1.34	10.575 / 6.15	12.406 / 5.24	14.255 / 5.05
Chromium (Cr)	6.954 / 5.71	7.454 / 5.73	8.678 / 7.29	9.061 / 8.78	7.648 / 6.35	7.109 / 2.17	4.614 / 1.43	7.522 / 6.13	7.122 / 1.65	4.824 / 2.08
Barium (Ba)	72.471 / 54.69	125.52 / 138.47	147.752 / 133.28	101.597 / 104.14	153.045 / 124.50	116.118 / 89.76	42.214 / 26.17	122.727 / 101.78	115.997 / 46.46	96.818 / 58.14
Borom (B)	0.958 / 0.53	0.836 / 0.49	0.827 / 1.49	1.081 / 0.70	0.921 / 0.83	0.909 / 0.74	1.429 / 1.10	0.96 / 0.79	1.139 / 0.74	0.58 / 0.25
Tungsten (W)	0.044 / 0.06	0.046 / 0.12	0.058 / 0.04	0.037 / 0.05	0.102 / 0.27	0.048 / 0.03	0.01 / 0.00	0.048 / 0.17	0.096 / 0.03	0.033 / 0.09
Scandium (Sc)	1.421 / 0.58	1.905 / 1.18	2.492 / 1.36	1.858 / 1.16	2.169 / 1.77	1.582 / 0.70	1.086 / 0.30	1.704 / 1.02	1.714 / 0.69	1.829 / 0.98
Thallium (Tl)	0.254 / 0.16	0.306 / 0.21	0.26 / 0.46	0.27 / 0.14	0.188 / 0.09	0.35 / 0.50	0.077 / 0.06	0.262 / 0.18	0.632 / 1.09	0.352 / 0.17
Mercury (Hg)	0.119 / 0.08	0.089 / 0.05	0.237 / 0.45	0.093 / 0.05	0.059 / 0.07	0.175 / 0.36	0.196 / 0.25	0.091 / 0.13	0.131 / 0.40	0.115 / 0.09
Selenium (Se)	2.256 / 1.35	1.086 / 0.92	1.269 / 1.77	1.609 / 1.09	0.849 / 0.79	1.682 / 0.64	5.157 / 6.41	1.068 / 1.09	0.753 / 0.52	1.086 / 0.98
Tellurium (Te)	0.641 / 0.50	0.305 / 0.41	0.923 / 1.88	0.388 / 0.33	0.448 / 0.37	0.67 / 0.42	0.69 / 0.57	0.329 / 0.60	0.124 / 0.15	0.28 / 0.37
Gallium (Ga)	6.358 / 2.15	7.216 / 2.46	7.878 / 2.36	6.695 / 2.15	5.981 / 1.94	6.109 / 1.99	4.257 / 1.37	5.939 / 2.37	6.508 / 1.66	6.233 / 2.45

2.2. Methodology

Each surface sample provides information on its lithology (a category coded between 1 and 10), its spatial coordinates (easting and northing), and 36 quantitative variables (geochemical element concentrations). Our goal is to develop statistical and geostatistical models to predict the former (lithology) based on the information of the latter (coordinates and geochemical concentrations), trying to minimize the error between the predicted and observed lithologies. The methodology is summarized in the schematic diagram in Fig. A.4 and detailed in the next subsections.

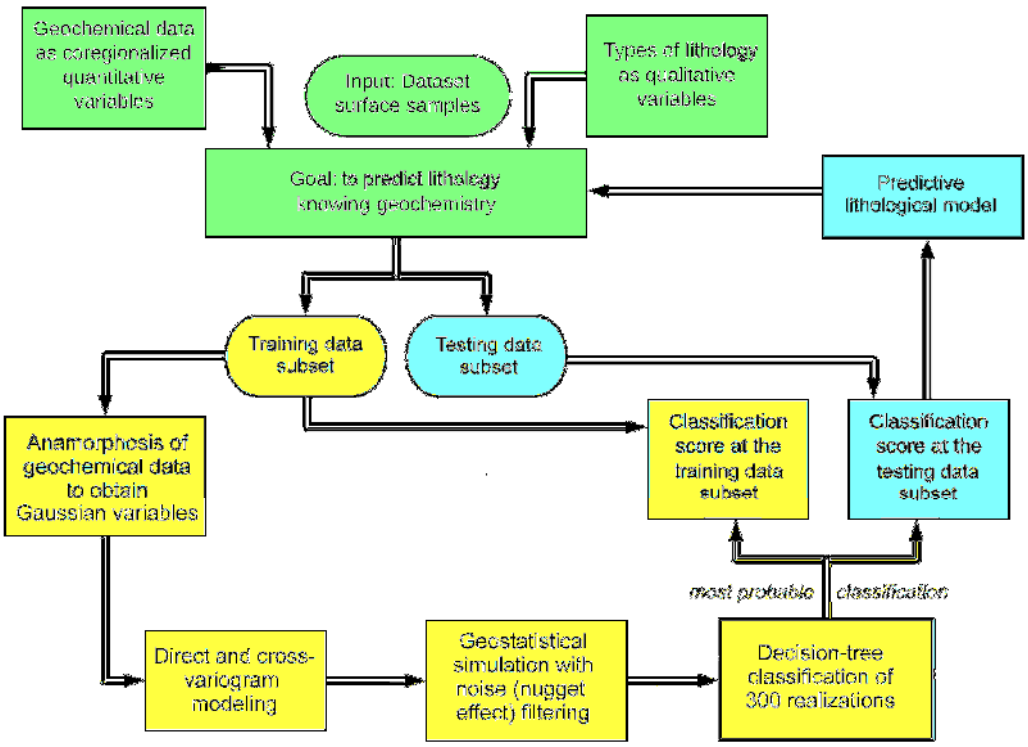

Figure A.4. Schematic diagram of the methodological proposal

2.2.1. Split sample testing

The original data set (3998 samples) is randomly split into two subsets (Fig. A.5):

- a training subset consisting of 3198 samples, which will be used to construct the geostatistical models and the classifiers presented next;
- a testing subset consisting of 800 samples, on which the lithology will be predicted by using only the information of the training subset and the classifiers fitted with this subset.

The accuracy of the proposed classifiers will be assessed on these two subsets. The assessment on the training subset will provide a measure of the ability of the classifier to determine a 'proxy' that predicts the lithology from the geochemical information. This measure is useful to discriminate between different possible classifiers and to choose the one that is most suitable to the data under study. However, since the proxy is established on the basis of all the training set information (both geochemistry and lithology), the accuracy score may not be representative of the ability to classify

a location with unknown lithology and, possibly, with unknown geochemistry too. The latter is the rationale for the use of a testing subset, which does not participate at all in the construction of the classifier: the accuracy score on this testing subset will reflect the performance of the classifier under real-world conditions, i.e., its ability to predict the lithology at unsampled locations.

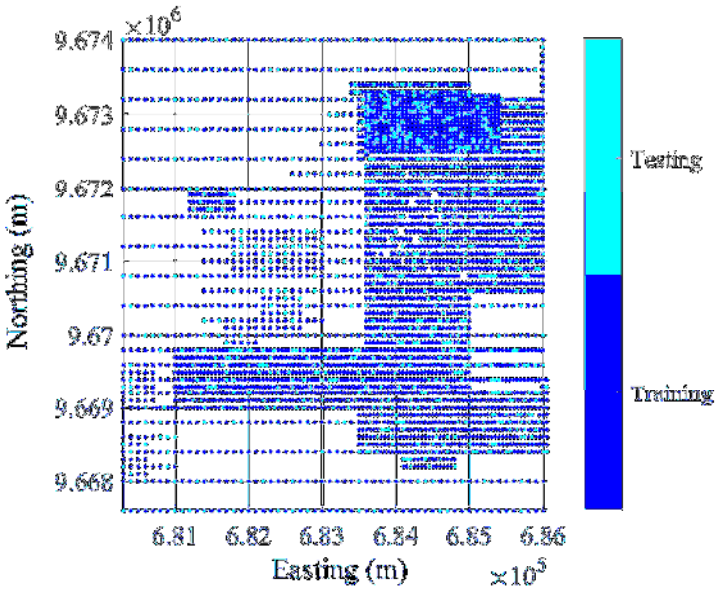

Figure A.5. Random division of surface samples into a training subset (3198 samples) and a testing subset (800 samples)

2.2.2. Gaussian anamorphosis

First of all, a multivariate geostatistical model is elaborated to allow the joint simulation of the continuous quantitative variables. In this context, the starting point is the transformation (anamorphosis) of the original variables (geochemical concentrations at the 3198 training data) into normally distributed variables with mean 0 and variance 1, since the simulation model considered is based on the assumption that the variables to be simulated are realizations of stationary Gaussian random fields (Chilès and Delfiner, 2012). The high number of variables (36) prevented us to use multivariate anamorphoses (Leuangthong and Deutsch, 2003; Barnett et al., 2013; van den Boogaart et al., 2017), so that the Gaussian transformation is performed by considering each variable separately.

2.2.3. Variogram analysis

The spatial correlation of the quantitative variables, now represented by Gaussian random fields, is modeled by nesting basic theoretical structures. Since it is not possible to identify preferential correlation directions, the fitting is made on the basis of the omnidirectional experimental direct (for each variable) and cross (between pairs of variables) variograms, calculated from the normal scores data at the training subset.

The high amount of variograms (36 direct and 630 cross-variograms) makes a manual fitting prohibitive due to the mathematical restrictions existing between all these variograms (Wackernagel, 2003). Accordingly, a semi-automatic algorithm is used to fit a linear model of

coregionalization, being only necessary to specify the number of basic nested structures, their types and ranges of correlation (Goulard and Voltz, 1992; Emery, 2010). In the present case, five spherical structures of ranges that vary from 100 m to 1450 m are considered, in addition to a nugget effect (Fig. A.6 and Table A.4).

Table A.8.4. Nested structures used in the fitting of direct and cross-variograms

Nested structure	Structure type	Range of correlation (m)
1	Nugget effect	0
2	Spherical	100
3	Spherical	200
4	Spherical	450
5	Spherical	1000
6	Spherical	1450

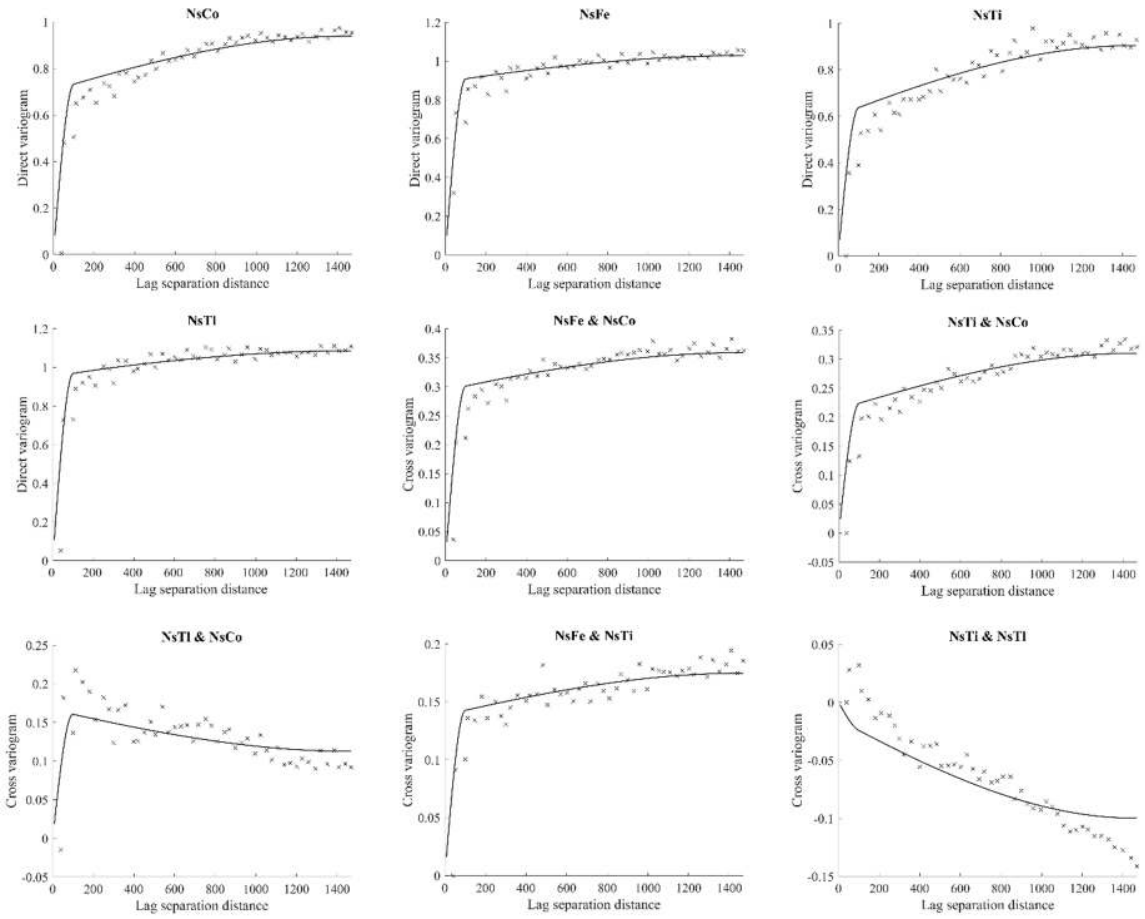

Figure A.6. An example of omnidirectional experimental direct and cross-variograms (crosses) and theoretical variogram models (solid lines) for quantitative variables Co, Fe, Ti, and Tl. All the modeled variograms are isotropic

Based on a visual assessment of Fig. A.6, the fit is considered satisfactory for all the Gaussian random fields: the fitted variogram models pass close to the calculated experimental points. The direct variograms tend to have a sill value close to one, which corroborates the validity of the stationarity assumption, at least local scale (quasi-stationarity) (Chilès and Delfiner, 2012). The cross-variograms are not identically zero and indicate the existence of spatial correlations between the quantitative variables, corroborating the dependencies observed in the correlation circle of Fig. A.3.

2.2.4. Noise filtering and conditional simulation

The nugget effect represents noise that is generally unknown to the researcher and is influenced by the sampling density and the sampling grid, as well as by complex geological environments, analytical errors, or an unknown short-scale site-specific variation (Hofmann et al., 2010; Carrasco, 2010). To minimize the impact of these types of noise on the regionalized classification, a filtering of the nugget effect is performed when simulating the Gaussian transforms of the geochemical concentrations.

Kriging and cokriging with filtering of spatial components are used for mapping regionalized properties in soil sciences, petroleum exploration, geophysical or geochemical prospecting (Galli et al., 1984; Sandjivy, 1984; Jaquet, 1989; Goovaerts, 1992; Wackernagel, 2003; Chilès and Delfiner, 2012; Castillo et al., 2015). However, very little is documented in the context of simulation techniques; we are only aware of the work of Larocque et al. (2006) who propose a sequential algorithm to simulate regionalized components and factors based on coregionalization analysis. Our procedure is slightly different (computationally less demanding and applicable to highly multivariate data sets) and consists of the following workflow:

1. Since the variogram model contains 6 nested structures, the information of the 36 Gaussian random fields can be decomposed into 216 (36 × 6) independent factors, using the so-called coregionalization analysis or geostatistical factor analysis (Goovaerts, 1992; Wackernagel, 2003). Such a factorization is one of the commonly used approaches to jointly simulate cross-correlated Gaussian random fields (Emery, 2008).
2. Each of the 216 factors is simulated at the training data locations, as well as at the testing data locations, using a continuous spectral turning-bands algorithm (Emery et al., 2016). This algorithm is preferred over other alternatives such as sequential or matrix decomposition methods, because of its applicability to highly multivariate simulation, low numerical complexity, minimal storage requirement and accuracy, being able to exactly reproduce the spatial correlation structure of each factor.
3. At each training data location, the 216 simulated factors are combined to reconstruct the 36 Gaussian random fields of interest, then the 36 ‘residuals’ (i.e., the differences between the simulated fields so reconstructed and the Gaussian values resulting from the Gaussian anamorphosis step) are calculated.
4. At both the training and testing data locations, the factors associated with the nested structures to be conserved are combined to reconstruct 36 variables filtered out from the nugget.
5. The filtered variables are then conditioned to the training data, by adding the simulated values obtained at step 4 and the simple cokriging (with zero mean) of the 36 residual variables obtained at step 3. In detail, cokriging takes as an input the unfiltered residuals calculated at the training locations only (step 3) and provides a prediction of the filtered

residuals at both the training and testing locations, which allow converting the non-conditional filtered simulation at these locations (step 4) into a conditional one. In comparison with the traditional conditioning procedure (e.g., Chilès and Delfiner, 2012), the difference lies in the right-hand side of the cokriging system of equations (corresponding to data-to-target direct and cross-covariance entries), where the nugget effect is removed and only the basic nested structures of interest are kept, while the left-hand side (corresponding to data-to-data direct and cross-covariance entries) contains all the basic nested structures, including the nugget effect, as in traditional cokriging without filtering. This modification of the right-hand side of the cokriging system is intended to predict the filtered residual variables at the training and testing data locations from the unfiltered residual values available at the training data locations.

The result consists of 300 realizations of the Gaussian random fields filtered from the nugget component and conditioned to the training data. These realizations are constructed at both the training and testing subsets and constitute as many geological scenarios in which geochemical measurements would have been taken without any measurement error or noise. The high number of scenarios (300) is intended to produce statistically robust comparisons of the classification results.

For comparative purposes, the simulation is also performed without filtering, that is, a traditional cosimulation of the 36 Gaussian random fields is performed. The procedure differs in the last two steps of the above workflow: at step 4, all the 216 factors are considered to reconstruct the 36 variables without filtering, while the nugget effect is not removed from the right-hand side of the cokriging system at step 5. The unfiltered residual data calculated at step 3 are used in the conditioning cokriging at step 5, regardless of whether the simulation is performed with or without filtering, the reason being that the conditioning normal scores data and the associated residuals available at the training locations always contain the nugget component (the exact unfiltered values at these locations are unknown); this explains why the nugget effect is never removed from the left-hand side of the cokriging system containing the data-to-data covariance entries with, in the present case, noisy residual data.

The realizations without filtering the nugget effect provide the same Gaussian values as the ones observed at the training data locations, whereas the realizations exhibit less short-scale variability and differ from the data observed at the training subset when filtering the nugget effect. At the testing subset, both the realizations with filtering and without filtering differ from the normal scores transforms of the measured geochemical data, insofar as these data have not been used in the simulation and in the conditioning process.

Note that no back-transformation of the simulated Gaussian values is performed. The reason is twofold: on the one hand, such a back-transformation is needless for the purpose of classification, as detailed next; on the other hand, it cannot be performed for the filtered simulation, insofar as the anamorphosis defined in Section 2.2.2 only applies to the unfiltered variables.

2.2.5. Classification using decision trees at the training data subset

Several classification algorithms (linear discriminant analysis, support vector machines and decision trees) are trained on the training data subset and their accuracy score are compared in order to select the most performant algorithm, which in the present case turned out to be the

Decision Tree model (Mitchell, 1997) with the CHAID (Chi-square automatic interaction detectors) division method (Kass, 1980; Ibarguren et al., 2016). In particular, CHAID yields an accuracy score about twenty percentage points higher than linear discriminant analysis (Table A.5). Having chosen the classifier, two strategies are considered for classification based on the simulated (Gaussian transforms of) geochemical variables. The first strategy applies CHAID using as a training set the average of the 300 simulated values at each training data location, which represents an average (expected) scenario almost identical to the simple cokriging prediction, with or without nugget effect filtering depending on the case under consideration (in the multivariate Gaussian framework, simple cokriging coincides with the average of infinitely many conditional realizations). As an alternative, the second strategy consists of repeating the classification 300 times, as many as there are scenarios of the geochemical variables, using in each repetition the simulated values of a single scenario as a training set; then, using the 300 classifications so obtained, the most probable lithology (i.e., the lithology that is most repeated among the 300 classifications) is assigned to each training data location.

Table A.8.5. Accuracy scores of several classifiers applied to the training data subset

Classification Algorithm	Accuracy score
Decision tree with CHAID division method	75.70%
Support vector machine	73.67%
Classification and regression tree (C&rt)	72.70%
Random forest	72.64%
Quick, Unbiased, Efficient Statistical Tree (QUEST)	68.20%
Linear discriminant analysis	55.80%

For each strategy, the accuracy of the classification is quantified by the percentage of correctly classified data (Tables A.6 and A.7). The entire procedure applies for the realizations with noise filtering, as well as for traditional realizations without filtering.

Table A.8.6. Filtering of noise (nugget effect, 1st nested structure) and its accuracy in predicting lithology after classifying using decision trees. Accuracy scores are calculated on the 3198 training data subset

Experiment	Filtered nested structures						Classification Accuracy			
	1st	2nd	3rd	4th	5th	6th	Most probable lithology	Classification on the average of the realizations	Classified data	Correctly classified data (most probable lithology)
1	No	No	No	No	No	No	75.70 %	75.70 %	3198	2421
	Yes	**No**	**No**	**No**	**No**	**No**	**86.80 %**	77.07 %	3198	2776
2	No	No	No	No	No	No	72.29%	72.29%	3198	2312
	Yes	**No**	**No**	**No**	**No**	**No**	**86.02%**	74.83%	3198	2751
3	No	No	No	No	No	No	76.33%	76.33%	3198	2441
	Yes	**No**	**No**	**No**	**No**	**No**	**84.33%**	73.64%	3198	2697
4	No	No	No	No	No	No	71.64%	71.64%	3198	2291
	Yes	**No**	**No**	**No**	**No**	**No**	**85.15%**	74.58%	3198	2723

Table A.8.7. Numbers of correctly classified data (diagonal line) and misclassified data (outside the diagonal) for each observed lithology (row) and each predicted lithology (column), for the training data subset. The numbers correspond to the most probable lithology over 300 realizations constructed by filtering noise (nugget effect)

Observed lithology	Predicted lithology (most probable lithology)										Total	Accuracy (%)
	Bx	Dac	Jub	Md	Qd	Rhy	Rp	Sar	Tbx	Tuf		
Bx	**9**	0	2	5	0	0	0	3	0	0	19	47.37%
Dac	0	**161**	43	1	0	0	0	114	0	0	319	50.47%
Jub	0	0	**1010**	0	0	0	0	8	2	0	1020	99.02%
Md	0	0	2	**150**	0	0	0	104	0	0	256	58.59%
Qd	0	0	1	1	**33**	0	0	44	2	0	81	40.74%
Rhy	0	0	0	2	0	**3**	0	6	0	0	11	27.27%
Rp	0	0	0	3	0	0	**3**	0	0	0	6	50.00%
Sar	0	2	16	8	4	0	0	**1386**	1	0	1417	97.81%
Tbx	0	0	14	0	0	0	0	6	**13**	0	33	39.39%
Tuf	1	11	8	0	0	0	0	8	0	**8**	36	22.22%
Total	**10**	**174**	**1096**	**170**	**37**	**3**	**3**	**1679**	**18**	**8**	**3198**	**86.80%**
										Classification Accuracy		**86.80%**

2.2.6. Classification using decision trees at the testing data subset

As previously discussed, the scores in Tables A.6 and A.7 may not be representative of the performance of the classifier at unsampled locations. For this reason, the classifiers fitted on the training data subset are applied to the simulated values (with or without noise) at the testing subset locations, which have been conditioned to the training data only. This exercise corresponds to a perfect extrapolation situation: predicting the lithology at locations for which the lithological and geochemical information is unknown. The results are indicated in Tables A.8 and A.9.

Table A.8.8. Filtering of noise (nugget effect, 1st nested structure) and its accuracy in predicting lithology after classifying using decision trees. Accuracy scores are calculated on the 800 testing data subset

Experiment	Filtered nested structures						Classification Accuracy			
	1st	2nd	3rd	4th	5th	6th	Most probable lithology	Classification on the average of the realizations	Classified data	Correctly classified data (most probable lithology)
1	No	No	No	No	No	No	69.25 %	53.63 %	800	554
	Yes	**No**	**No**	**No**	**No**	**No**	**73.13 %**	58.50 %	800	585
2	No	No	No	No	No	No	69.25%	57.38%	800	554
	Yes	**No**	**No**	**No**	**No**	**No**	**73.13%**	59.13%	800	585
3	No	No	No	No	No	No	68.63%	59.38%	800	549
	Yes	**No**	**No**	**No**	**No**	**No**	**70.38%**	56.75%	800	563
4	No	No	No	No	No	No	69.63%	51.00%	800	557
	Yes	**No**	**No**	**No**	**No**	**No**	**74.00%**	54.25%	800	592

Table A.8.9. Numbers of correctly classified data (diagonal line) and misclassified data (outside the diagonal) for each observed lithology (row) and each predicted lithology (column), for the testing data subset. The numbers correspond to the most probable lithology over 300 realizations constructed by filtering noise (nugget effect)

Observed lithology	Predicted lithology (most probable lithology)								Total	Accuracy (%)
	Bx	Dac	Jub	Md	Qd	Sar	Tbx	Tuf		
Bx	**0**	0	1	0	0	4	0	0	5	0.00%
Dac	0	**0**	21	0	0	50	0	0	71	0.00%
Jub	0	0	**229**	0	0	25	0	0	254	90.16%
Md	0	0	0	**0**	0	77	0	0	77	0.00%
Qd	0	0	0	0	**0**	20	0	0	20	0.00%
Sar	0	0	1	0	0	**356**	0	0	357	99.72%
Tbx	0	0	3	0	0	0	**0**	0	3	0.00%
Tuf	0	8	2	0	0	11	0	**0**	13	0.00%
Total	**0**	**0**	**257**	**0**	**0**	**543**	**0**	**0**	**800**	**73.13%**
								Classification Accuracy		**73.13%**

3. Results and analyses

3.1. Classification scores at the training data subset

The CHAID classification based on the normal scores data at the training subset reaches an accuracy score of 75.70%, with 2421 correctly classified out of 3198 data (Table A.6, first experiment). The same percentage is obtained with the simulation without nugget effect filtering, either when considering the average of the realizations or when considering the most probable lithology: by construction, all the realizations exactly reproduce the training data, so that all the 300 classifications are identical to the CHAID classification applied to the normal scores data. The same does not happen with the classifications made on the realizations that consider a filtering of the nugget effect. Classification in a single realization provides an accuracy score of between 69.23% and 79.77% (Fig. A.7), so it is generally not better than traditional classification (accuracy score of 75.70%). This can be explained because a single realization is not designed to accurately predict the actual values, but to provide a possible scenario of how these values could fluctuate in space. In spite of this, the classification based on the set of 300 realizations with filtering yields the following results (Table A.6, first experiment):

- the first strategy, consisting of classifying on the average of 300 realizations, has a mixed performance, with a slight improvement of the accuracy score (77.08%);
- the second strategy (Fig. A.8b) performs much better and achieves an accuracy score of 86.80% (that is, 11 percentage points higher than the classification based on the noisy normal scores data), thus proving that it takes better advantage of the spatial information. Unlike the first strategy, the most probable lithology accounts for the uncertainty in the real values of the geochemical concentrations, reflected in the variability existent between one scenario and another, whereas the average of the realizations smoothes out the spatial variability and loses valuable information.

To assess the robustness of the presented results, the entire experiment (anamorphosis, variogram analysis, simulation and classification) has been repeated with three other training data subsets obtained by randomly selecting 3198 out of the 3998 available data (Table A.6, experiments 2, 3 and 4). In all cases, the accuracy scores are comparable to the ones presented above (experiment 1), with an improvement of 8 to 14 percentage points for the most probable lithology in comparison with the classification on the noisy data, while the classification on the average of the realizations only shows a slight improvement (experiments 2 and 4) or deterioration (experiment 3). In the following, to avoid an excess of tables and figures, we focus on the results of experiment 1.

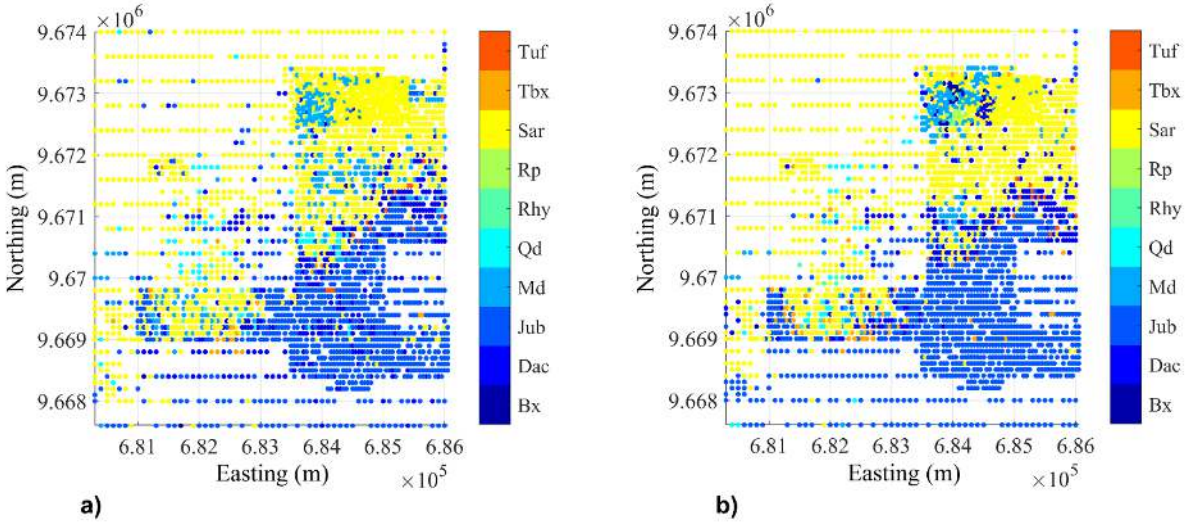

Figure A.7. Classifications at the training data subset for two realizations filtering the nugget effect component: a) realization 154 with an accuracy score of 69.23%, b) realization 298 with an accuracy score of 79.77%

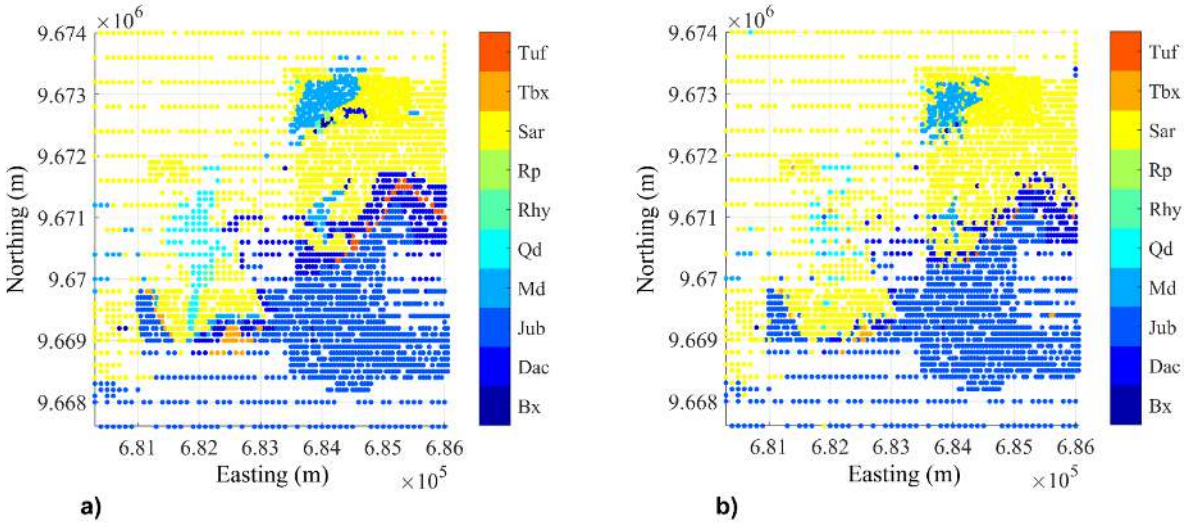

Figure A.8. a) Initial map of the training data locations with their respective lithology; b) predictive lithological map obtained with a set of 300 realizations filtering the nugget effect component

3.2. Classification scores at the testing data subset

The performance of the classifier is poorer when they are applied to the testing subset: the classification based on the unfiltered (noisy) simulation has an accuracy score of 69.25% when considering the most probable lithology, while the classification based on the filtered simulation has an accuracy score of 73.13% (Table A.8, experiment 1). The lower accuracy scores with respect to the previous subsection can be explained because the classification is performed at locations where not only the lithology is unknown, but also the concentrations of the geochemical elements (the classifiers take as input the simulated concentrations conditioned only on the information of the training data subset), therefore a deterioration of the accuracy scores was expectable. Said in other words, the simulated Gaussian values (with or without filtering) at a given testing data location have an increased variability across the 300 realizations due to the higher uncertainty at this location in comparison with a training data location, which translates into an increased variability in the 300 lithologies obtained with CHAID and a lower accuracy in the prediction of the true lithology. This increased variability can be illustrated by calculating, at each data location, the Simpson index (Simpson, 1949) giving the probability of finding the same lithology in two realizations taken at random from the 300 realizations: when using the second strategy (most probable lithology), the average index is 0.706 for the training data, while it is only 0.378 for the testing data (Fig. A.9), showing (on average) much more diversity of the 300 classifications at a testing data location than at a training data location.

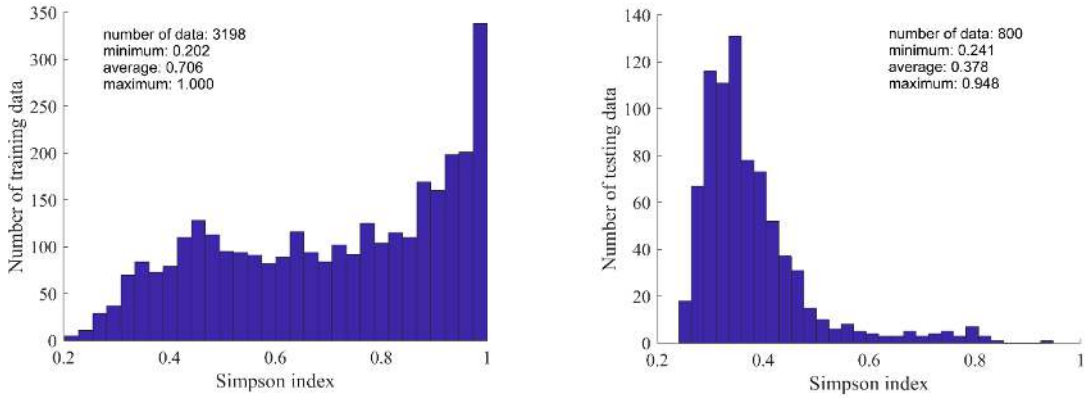

Figure A.9. Histogram of Simpson's diversity index calculated over the 300 simulated lithologies for a) the 3198 training data and b) the 800 testing data. The lower the index, the more diversity in the lithologies obtained over the 300 realizations at a given location, reflecting more uncertainty in the unknown lithology

In addition to a lower accuracy of the classifiers, one also observes a reduced difference in the scores obtained with and without noise filtering. However, the difference is still significant, at least for the second strategy (3.5 percentage points on average in experiments 1 to 4), and corroborates the improvement brought by filtering the undesired noise in the geochemical measurements. As it already happened for the training data subset, the first classification strategy based on the average of 300 realizations yields consistently worse accuracy scores (almost 15 percentage points lower than the second strategy based on the most probable lithology) (Table A.8) and, therefore, lacks interest in practice.

Remarkably, classifying the testing data based on their measured geochemical concentrations with the classifier fitted on the training data reaches an accuracy score of 58.75%, in-between the scores obtained with the two proposed strategies: knowing the geochemical concentrations improves the classification based on the average simulated concentration, but it is not better than the most probable lithology, even when the simulated concentrations are conditioned to the training data

only (thus, ignoring the testing data concentrations). In other words, the measurement noise (nugget effect) spoils the classification and is largely compensated by the filtering combined with the use of the most probable lithology as the prediction.

3.3. Classification of scarce lithologies

As indicated in Table A.7, the Rhyolitic-Rp porphyry lithology is scarce in the training data subset (Fig. A.8a) and in the classification half becomes part of the Microdioritic-Md porphyry (Fig. A.8b), more abundant and similar in relation to the texture of the rock. In the same way, an important part of the Rhyolites-Rhy and hydrothermal breccias-Tbx data (Fig. A.8a) are classified as belonging to the Saraguro-Sar and Jubones-Jub lithologies (Fig. A.8b), possibly due to the fact that both Rhy and Tbx are spatially scattered, each spot having a small extension, so they become part of neighboring formations with greater extensions, similar geochemical composition and petrographic characteristics after classification.

The above statements become more obvious when the classifier is applied to the testing subset, where the classification with the highest accuracy score (73.13%, most probable lithology calculated from simulation with noise filtering) only contains the two predominant lithology (Jub and Sar) (Table A.9 and Fig. A.10). The price to pay for minimizing the classification errors and getting an accurate prediction is a shift towards the most abundant lithologies to the detriment of scarce lithologies that tend to disappear in the classification results.

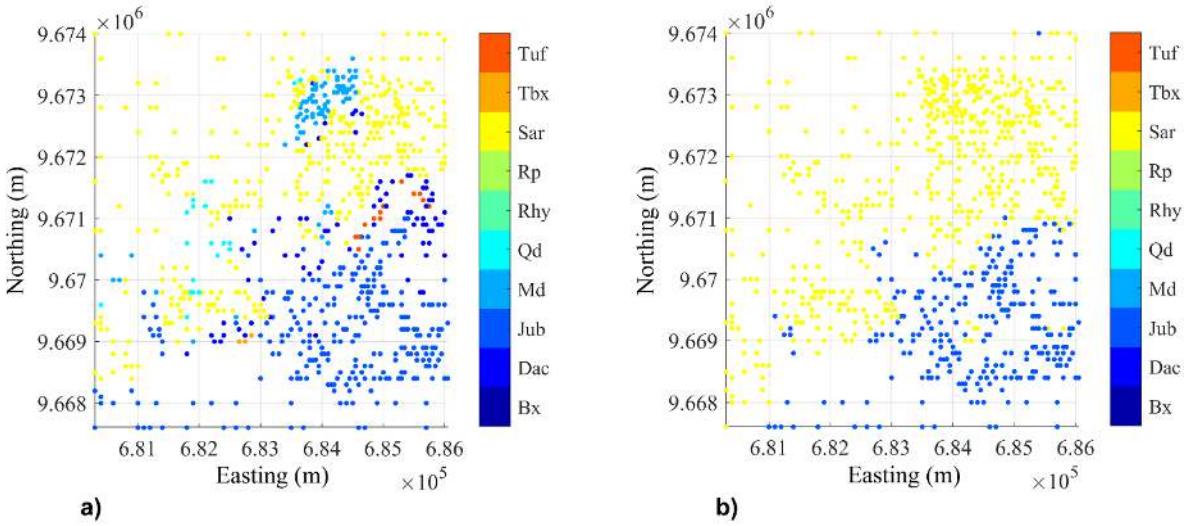

Figure A.10. a) Initial map of the testing data locations with their respective lithology (10 classes); b) predictive lithological map with 2 classes, obtained with a set of 300 realizations filtering the nugget effect component

3.4. Predictive lithological mapping

The Gaussian transforms of the geochemical concentrations can be simulated and classified into lithologies at any unsampled location, not only at the training or testing data locations, which yields a predictive lithological map of the entire domain under study. An example is provided in Fig. A.11a, corresponding to the most probable lithology obtained after noise filtering. Compared with the interpretive lithological map (Fig. A.1) and with the global proportions calculated with the

training and testing data (Table A.2), the shift towards the most abundant lithologies remains patent, and the scarcest lithologies are almost – although not completely – absent (Table A.10). This shift does not arise so clearly on the classifications obtained on the individual realizations, which exhibit proportions comparable to that of the data (Tables A.2 and A.10) (for each lithology, the data proportion always lies in-between the minimum and maximum proportions over the 300 classifications) and do not suffer from an under-representation of the less abundant lithologies.

Not all the information on the less abundant lithologies is lost however, insofar as the 300 classifications calculated on as many simulated scenarios allow to calculate the probability of occurrence of each lithology at each target location, instead of (in addition to) predicting a single lithology per target location (Fig. A.11c-l). The probability maps so obtained agree with the interpretive lithological map in Fig. A.1, as the highest probabilities of Bx, Dac, Jub, Md, Qd, Rhy, Rp and Sar mostly occur where these lithologies have been interpreted by exploration geologists. This agreement also holds, to a smaller extent, for the two remaining lithologies (Tbx and Tuf), for which the highest probabilities coincide with the interpretive map, although some intermediate probabilities of Tbx and Tuf arise in sectors interpreted as Sar and Dac, respectively. This can be explained because of the geochemical and lithological resemblances between, on the one hand, Tbx and Sar (the former may contain rhyolitic breccias and be similar to the latter that contains rhyolites) and, on the other hand, Tuf and Dac, both of which are tuffs that are part of the Plancharumi formation.

As a final note, the probability maps (Fig. A.11c-l) can be used to derive a predictive map that reproduces the global proportion of each lithology (as estimated in Table A.2) by using Soares' algorithm (Soares, 1992). Fig. A.11b shows the result of such a classification, which has a better global accuracy as it avoids the under-representation of scarce lithologies, but not necessarily a better local accuracy: taking the interpretive map (Fig. A.1) as the reference, one observes local improvements in Fig. A.11b in comparison with Fig. A.11a for the Md lithology or for Dac in the sector located in-between the two predominant lithologies Jub and Sar, but worse predictions for Qd or for Dac in the western and northeastern sectors of the domain under study.

Table A.8.10. Global proportions over the domain under study (14,964 grid nodes) for the 300 realizations with noise filtering and for the most probable lithology

Lithology	Individual realizations			Most probable lithology
	Minimum	Average	Maximum	
Bx	0.19%	1.40%	6.51%	0.06%
Dac	3.76%	8.62%	14.78%	1.13%
Jub	24.35%	32.14%	39.16%	33.79%
Md	2.71%	5.74%	12.02%	1.06%
Qd	1.70%	4.84%	10.76%	0.28%
Rhy	0.00%	0.89%	5.37%	0.02%
Rp	0.00%	0.14%	2.21%	0.04%
Sar	35.79%	43.41%	51.92%	63.45%
Tbx	0.31%	1.84%	4.14%	0.11%
Tuf	0.02%	0.97%	4.12%	0.06%

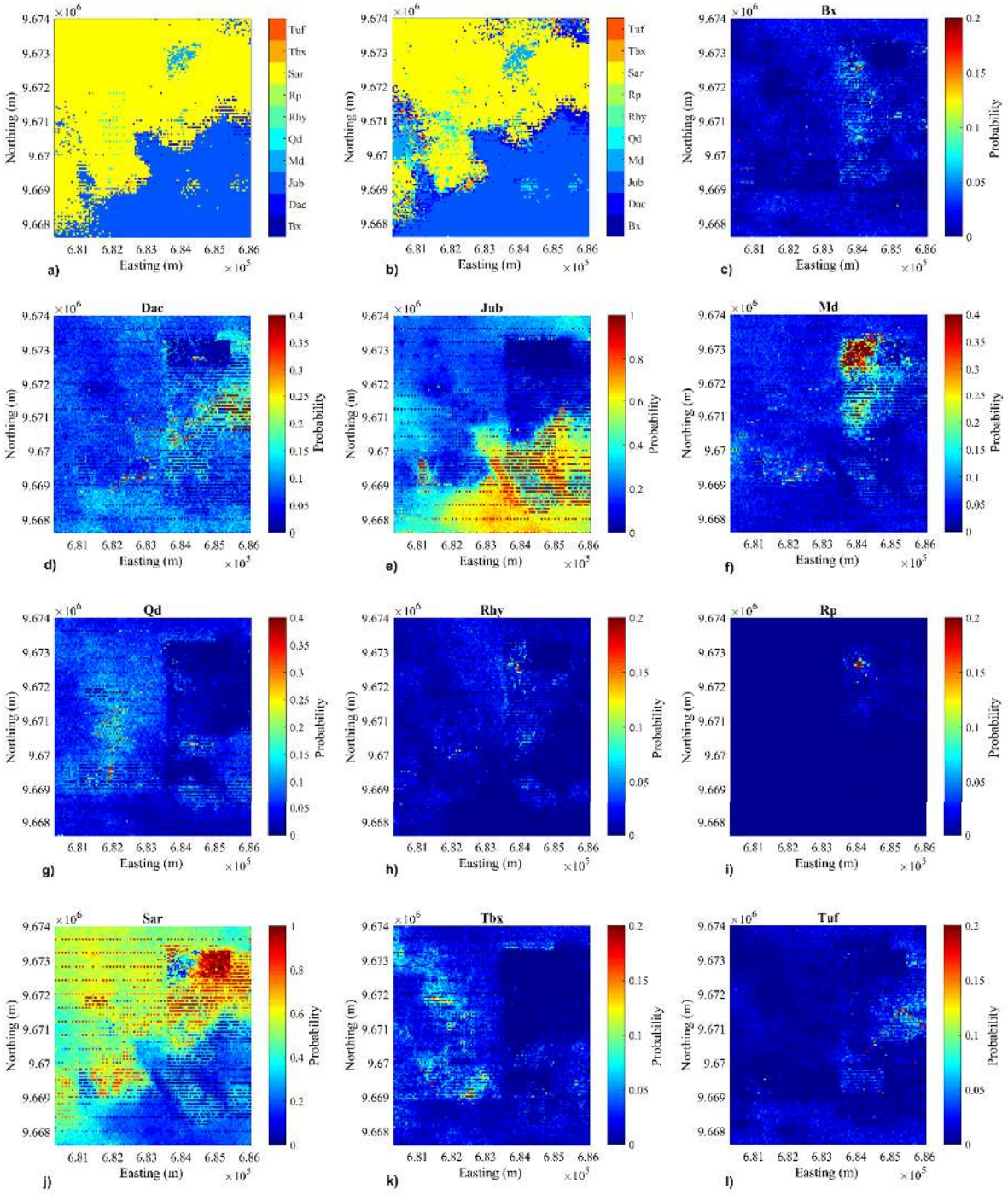

Figure A.11. a) Predictive lithological map on a grid with mesh 50m × 50m (14,964 nodes) covering the domain of interest, obtained by considering the most probable lithology on a set of 300 realizations filtering the nugget effect component; b) predictive map on the same grid obtained by using Soares' algorithm; c) to l) probability maps of Bx, Dac, Jub, Md, Qd, Rhy, Rp, Sar, Tbx and Tuf calculated with the same 300 realizations as in a)

4. Conclusions

This study presents an approach for predicting lithology using geochemical concentrations, which relies on the combined use of geostatistical simulation, spatial component filtering based on coregionalization analysis and a predictive decision trees model with the CHAID classifier. The proposed approach has been applied to an exploration area in southern Ecuador, where lithology has been predicted from the geochemical information of surface samples.

The results show that the benefit of accounting for the spatial uncertainty and spatial variability of multivariate data for regionalized classification is twofold. On the one hand, the prediction of the lithology at an unsampled location is substantially improved by simulating the geochemical concentrations at this location, applying a classifier previously fitted on training data as many times as there are simulated scenarios, and retaining the most probable lithology as the prediction. This strategy increases the accuracy score by 15 percentage points (from 53.63% to 69.25%) with respect to the traditional approach consisting in applying the classifier to a prediction of the geochemical concentrations (average simulated concentrations), as seen in Table A.8 for the first experiment. Even more, the accuracy score is 10 percentage points above the one (58.75%) obtained by using the geochemical concentrations at the target point, assuming that these concentrations have been measured.

On the other hand, the decomposition of the joint spatial variability into a set of factors associated with different scales of variations and the filtering of the short-scale factors (in particular, the nugget effect corresponding to noise and measurement errors) improves the fitting of the classifier on the training data and the accuracy score at unsampled locations. When considering the most probable lithology among the simulated scenarios at the testing data subset, the latter score increases from 69.25% (simulation without filtering) to 73.13% (simulation with filtering) (Table A.8, first experiment).

Having a better-trained and more accurate classifier leads to the generation of a predictive lithological map and can be used as an exploratory analysis tool for the discoveries of geological processes.

A last word on the shift of the predictive model towards the most abundant lithologies, in the detriment of scarce lithologies: this tendency may be unavoidable when looking for a classifier that aims at minimizing the errors. However, having applied the classification to many scenarios not only provides the most probable lithology, but also an estimate of the probability of each lithology at each target location, so that the practitioner can be aware of how likely is the occurrence of each lithology. The probabilities can also be an input to classification methods that aim to reproduce the global lithological proportions, hence without an under-representation of the scare lithologies, at the price of a loss of local accuracy.

References

Adeli, A., Emery, X., & Dowd, P. (2018). Geological modelling and validation of geological interpretations via simulation and classification of quantitative covariates. *Minerals*, 8(1), 7.

Baldock, J.W. (1982). Geología del Ecuador. Boletín de Explicación del Mapa Geológico (1:1.000.000) de la República del Ecuador. Resource document. Ministerio de Recursos Naturales y Energéticos, Quito, 54 pp.

Barbosa, P., Oliveira, T., & Silva, J. (2010). Regionalized classification of multivariate geochemical data from Jacupiranga Alkaline Complex (Ribeira de Iguape Valley/Sao Paulo, Brazil). *Revista Brasileira de Geociencias*, 40(2), 212-219.

Barnett, R.M., Manchuk, J.G., & Deutsch, C.V. (2013). Projection pursuit multivariate transform. *Mathematical Geosciences*, 46(3), 337-359.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Carranza, E.J.M. (2009). *Geochemical Anomaly and Mineral Prospectivity Mapping in GIS*. Amsterdam: Elsevier.

Carrasco, P. (2010). Nugget effect, artificial or natural? *Journal of the Southern African Institute of Mining and Metallurgy*, 110(6), 299-305.

Castillo, P.I.C., Townley, B.K., Emery, X., Puig, A.F., & Deckart, K. (2015). Soil gas geochemical exploration in covered terrains of northern Chile: data processing techniques and interpretation of contrast anomalies. *Geochemistry: Exploration, Environment, Analysis*, 15(2-3), 222-233.

Chilès, J.P., & Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Darsow, A., Schafmeister, M.T., & Hofmann, T. (2009). An ArcGIS approach to include tectonic structures in point data regionalisation. *Ground Water*, 47(4), 591-597.

Emery, X. (2008). A turning bands program for conditional co-simulation of cross-correlated Gaussian random fields. *Computers & Geosciences*, 34(12), 1850-1862.

Emery X. (2010). Iterative algorithms for fitting a linear model of coregionalization. *Computers & Geosciences*, 36(9), 1150-1160.

Emery, X., Arroyo, D., & Porcu, E. (2016). An improved spectral turning-bands algorithm for simulating stationary vector Gaussian random fields. *Stochastic Environmental Research and Risk Assessment*, 30(7), 1863-1873.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179-188.

Galli, A., Gerdil-Neuillet, F., & Dadou, C. (1984). Factorial kriging analysis: a substitute to spectral analysis of magnetic data. In: G. Verly, M. David, A.G. Journel, A. Maréchal (Eds.), *Geostatistics for Natural Resources Characterization* (pp. 543-557). Dordrecht: Reidel.

Goovaerts, P. (1992). Factorial kriging analysis: a useful tool for exploring the structure of multivariate spatial soil information. *Journal of Soil Science*, 43, 597-619.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford: Oxford University Press.

Goulard, M., & Voltz, M. (1992). Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, 24(3), 269-286.

Gringarten, E., & Deutsch, C.V. (2001). Teacher's aide: variogram interpretation and modelling. *Mathematical Geology*, 33(4), 507-534.

Grunsky, E.C. (2010). The interpretation of geochemical survey data. *Geochemistry: Exploration, Environment and Analysis*, 10, 27-74.

Grunsky, E.C., Corrigan, D., Mueller, U.A., & Bonham-Carter, G.F. (2012). Predictive geologic mapping using lake sediment geochemistry in the Melville Peninsula. Geological Survey of Canada, Open File 7171. http://dx.doi.org/10.4095/291901 (1 sheet).

Grunsky, E.C., Mueller, U.A., & Corrigan, D. (2014). A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: applications for predictive geological mapping. *Journal of Geochemical Exploration*, 141, 15-41.

Hassan, A.E., Bekhit, H.M., & Chapman, J.B. (2009). Using Markov Chain Monte Carlo to quantify parameter uncertainty and its effect on predictions of a groundwater flow model. *Environmental Modelling and Software*, 24(6), 749-763.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second ed.). New York: Springer.

Hofmann, T., Darsow, A., & Schafmeister, M.T. (2010). Importance of the nugget effect in variography on modeling zinc leaching from a contaminated site using simulated annealing. *Journal of Hydrology*, 389(1-2), 78-89.

Ibarguren, I., Lasarguren, A., Pérez, J.M., Muguerza, J., Arbelaitz, O., & Gurrutxaga, I. (2016). BFPART: Best-first PART. *Information Sciences*, 367-368, 927-952.

Jaquet, O. (1989). Factorial kriging analysis applied to geological data from petroleum exploration. *Mathematical Geology*, 21(7), 683-691.

Jenny, H. (1941). *Factors of Soil Formation: a System of Quantitative Pedology*. New York: McGraw-Hill.

Kass, G.V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29(2), 119.

Kuhn, S., Cracknell, M.J., & Reading, A.M. (2019). Lithological mapping in the Central African Copper Belt using random forests and clustering: strategies for optimised results. *Ore Geology Reviews*, 112, 103015.

Larocque, G., Dutilleul, P., Pelletier, B., & Fyles, J.W. (2006). Conditional Gaussian co-simulation of regionalized components of soil variation. *Geoderma*, 134, 1-16.

Leuangthong, O., & Deutsch, C.V. (2003). Stepwise conditional transformation for simulation of multiple variables. *Mathematical Geology*, 35(2), 155-173.

Liu, Y., Carranza, E.J.M., Zhou, K.F., & Xia, Q.L. (2019). Compositional balance analysis: an elegant method of geochemical pattern recognition and anomaly mapping for mineral exploration. *Natural Resources Research*, 28, 1269-1283.

Matheron, G. (1962). *Traité de Géostatistique Appliquée*. Paris: Technip.

McKay, G., & Harris, J.R. (2016). Comparison of the data-driven random forests model and a knowledge-driven method for mineral prospectivity mapping: a case study for gold deposits around the Huritz Group and Nueltin Suite, Nunavut, Canada. *Natural Resources Research*, 25(2), 125-143.

Merian, E., Anke, M., Ihnat, M., & Stoeppler, M. (2004). *Elements and Their Compounds in the Environment - Occurrence, Analysis and Biological Relevance*. New York: Wiley.

Mitchell, T.M. (1997). *Decision Tree Learning*. Singapore: WCB/McGraw-Hill Inc.

Olea, R.A. (1999). *Geostatistics for Engineers and Earth Scientists*. New York: Springer.

Quinlan, J.R. (1993). *C4.5, Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71, 804-818.

Salminen, R., Batista, M.J., Bidovec, M., Demetriades, A., De Vivo, B., De Vos, W., Duris, M., Gilucis, A., Gregorauskiene, V., Halamic, J., Heitzmann, P., Lima, A., Jordan, G., Klaver, G., Klein, P., Lis, J., Locutara, J., Marsina, K., Mazreku, A., O´ Connor, P.J., Olsson, S.À., Ottesen, R.-T., Plant, J.A., Reeder, S., Salpeeur, I., Sandström, H., Siewers, U., Steenfelt, A., & Travainen, T. (2005). *Geochemical Atlas of Europe*. Espoo: Geological Survey of Finland.

Sandjivy, L. (1984). The factorial kriging analysis of regionalized data - its application to geochemical prospecting. In: G. Verly, A.G. Journel, A. Maréchal (Eds.), *Geostatistics for Natural Resources Characterization* (pp. 559-571). Dordrecht: Reidel.

Simpson, E.H. (1949). Measurement of diversity. *Nature*, 163(4148), 688.

Soares, A. (1992). Geostatistical estimation of multi-phase structures. *Mathematical Geology*, 24(2), 148–160.

Stanley, C.R., & Sinclair, A.J. (1989). Comparison of probability plots and the gap statistic in the selection of thresholds for exploration geochemistry data. *Journal of Geochemical Exploration*, 32(1-3), 355–357.

Sun, T., Chen, F., Zhong, L.X., Liu, W.M., & Wang, Y. (2019). GIS-based mineral prospectivity mapping using machine learning methods: a case study from Tongling ore district, eastern China. *Ore Geology Reviews*, 109, 26-49.

Talebi, H., Mueller, U., Tolosana-Delgado, R., Grunsky, E.C., McKinley, J.M., & de Caritat, P. (2019). Surficial and deep earth material prediction from geochemical compositions. *Natural Resources Research*, 28, 869-891.

Tolosana-Delgado, R., Mueller, U., & van den Boogaart, K.G. (2019). Geostatistics for compositional data: an overview. *Mathematical Geosciences*, 51(4), 485-526.

van den Boogaart, K.G., Mueller, U., & Tolosana-Delgado, R. (2017). An affine equivariant multivariate normal score transform for compositional data. *Mathematical Geosciences*, 49(2), 231-251.

Wackernagel, H. (1988). Geostatistical techniques for interpreting multivariate spatial information. In: C.F. Chung, A.G. Fabbri R. Sinding-Larsen (Eds.), *Quantitative Analysis of Mineral and Energy Resources* (pp. 393-409). Dordrecht: Reidel.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Berlin: Springer.

Xiang, J., Xiao, K.Y., Carranza, E.J.M., Chen, J.P., & Li, S. (2020). 3D mineral prospectivity mapping with random forests: a case study of Tongling, Anhui, China. *Natural Resources Research*, 29(1), 395-414.

Zuo, R.G., & Xiong, Y.H. (2018). Big data analytics of identifying geochemical anomalies supported by machine learning methods. *Natural Resources Research*, 27(1), 5-13.