

Heber Hernández Guerra<sup>1</sup>, Cinthia Barboza Bravo<sup>2</sup> y Elisabete Alberdi Celaya<sup>3</sup>

<sup>1</sup> Universidad Santo Tomas, Av. Ejército Libertador 146, Santiago, Chile ([hhernandez9@santotomas.cl](mailto:hhernandez9@santotomas.cl))

<sup>2</sup> Universidad Nacional de Cajamarca, Av. Atahualpa Km. 3, Cajamarca, Perú ([cbarbozab17\\_1@unc.edu.pe](mailto:cbarbozab17_1@unc.edu.pe))

<sup>3</sup> Universidad del País Vasco, 48013 Bilbao, España ([elisabete.alberdi@ehu.es](mailto:elisabete.alberdi@ehu.es))

## Resumen

La definición de dominios geológicos es una de las primeras etapas en el proceso de modelado espacial y estimación de recursos minerales. La muestra resultante de un proceso exploratorio minero es dividida en zonas donde su ley mineral presenta homogeneidad o variación mínima, así mismo se diferencian unas de otras por sus rasgos geológicos y espaciales. La importancia de la definición de estos dominios radica en que la calidad de las técnicas de estimación y por consiguiente la correcta cuantificación del recurso mineral, mejorará en zonas geoestadísticamente estacionarias y geológicamente diferenciables.

En el presente estudio se buscan definir dominios geológicos y geoestadísticos de estimación para una ley mineral de oro, mediante un enfoque no tradicional basado en el algoritmo de agrupamiento k-prototipos. Este algoritmo está fundamentado en el paradigma k-medias de aprendizaje automático no supervisado, pero exento a la restricción de uso único en datos numéricos. Esto último es especialmente conveniente, al permitir incorporar variables categóricas como son los atributos geológicos en un agrupamiento automático mixto.

El caso de estudio corresponde a un depósito hidrotermal de oro de alta sulfuración, situado en la zona sur de Perú, donde se definen dominios de estimación a partir de un registro histórico de datos recuperados desde 148 sondajes de exploración. Las características involucradas en el estudio son; Oro (Au), plata (Ag), cobre (Cu), Arsénico (As), Mercurio (Hg), Azufre (S), alteración hidrotermal (A), coordenadas espaciales este (X), norte (Y) y elevación (Z). Los resultados obtenidos demostraron que el agrupamiento con k-prototipos es un enfoque eficiente y puede ser utilizado como una alternativa o complemento sobre la metodología tradicional. Se definieron dominios homogéneos, con sentido espacial, geológico y estacionariedad suficiente para el empleo de geoestadística lineal como técnica de estimación. Su uso reduce notablemente el tiempo en comparación a

la metodología tradicional, sin embargo, no deja exenta la supervisión y toma de decisiones por parte de un experto.

## 1. Introducción

En el campo de la estimación de recursos minerales, es una práctica común identificar volúmenes espacialmente coherentes, estadísticamente similares y geológicamente distintos de otros volúmenes a su alrededor. Estos se denominan dominios de estimación en la literatura geoestadística (Rossi y Deutsch, 2013), y conllevan a una mejora en el rendimiento de las técnicas de estimación. Los aspectos geológicos del depósito suelen ser las pautas fundamentales para la definición de dominios de estimación. Atributos como la alteración, mineralización y aspectos litológicos deben ser considerados (Emery y Ortiz, 2004). Glacken & Snowden (2001) afirman que un dominio geológico representa un área o volumen dentro del cual las características de la mineralización son más similares que fuera del dominio. Rossi y Deutsch (2013), son más específicos y definen los dominios geológicos como el equivalente a zonas geoestadísticamente estacionarias. En el contexto de la modelación de recursos minerales, el concepto de estacionariedad está estrechamente relacionado con la homogeneidad de los cuerpos geológicos y, simplificando la definición de Journel y Huijbregts (1978), se puede asumir que un fenómeno es estacionario cuando muestra valores esperados constantes, covarianza y estructuras de autocorrelación en cualquier ubicación dada del área de estudio. Los dominios de estimación que no se adhieran al principio de estacionariedad, pueden conducir a un sesgo significativo en las leyes minerales y, por lo tanto, a estimaciones erróneas (Sterk et al., 2019). En este artículo, el concepto de dominio geoestadístico se asocia directamente con el dominio geológico de estimación. Por lo que no habrá distinción en el término y de manera simplificada se hablará de dominios de estimación.

La metodología tradicional de definición de dominios de estimación para recursos minerales se sustenta de un estudio combinado entre geología y estadística., en

el cual predomina el entendimiento geológico de un experto y por consiguiente la intervención humana. A nivel general los pasos que sigue esta metodología son:

- Selección de atributos geológicos que controlan la ley mineral.
- Análisis estadístico de la ley mineral por cada categoría de cada atributo geológico.
- Combinación de categorías de cada atributo geológico por similitud estadística y contacto blando espacial, esto último también es conocido como definición de unidades geológicas (alteración, litología, mineralización, etc.).
- Combinación de atributos geológicos (unidades) por similitud estadística y contacto blando espacial.
- Validación de los dominios de estimación a nivel geológico, estadístico y geoestadístico.

Los aspectos más críticos en esta metodología son:

- Que es lenta, dado que todos los trabajos deben ser realizados de forma manual y estar siendo chequeados por un experto en la geología del depósito mineral.
- Que, en caso de modificaciones futuras de los dominios, el tiempo a invertir será similar al de comenzar a definir los dominios desde el principio.
- Que es subjetiva, ya que, de un experto a otro, se manifiestan distintos criterios e interpretaciones. Los códigos para reportar recursos minerales tienen como propósito definir los estándares mínimos (Stephenson, 2001), y esto abre la posibilidad de emplear múltiples métodos siempre y cuando puedan ser respaldados.
- Que no necesariamente el comportamiento de la ley mineral será homogéneo, se presentará estacionaridad y su estructura espacial será interpretable. Esto último al depender los dominios estrechamente de atributos geológicos que no necesariamente se agruparan con las características antes señaladas.

En búsqueda de una alternativa en la definición de dominios de estimación, el cumplimiento de los principios de estimación geoestadística, la reducción en el factor tiempo que se critica a la metodología tradicional y la disminución en la subjetividad, a través de un proceso con menor intervención., es que en este

estudio se evalúa la implementación de una metodología basada en aprendizaje automático no supervisado, mediante el uso de un algoritmo de agrupamiento multivariable.

Los algoritmos de agrupamiento han existido desde la década de 1960, cuando Sokal y Sneath (1965) presentaron la técnica jerárquica aglomerativa para trabajar en el campo de la taxonomía, y MacQueen (1967) introdujo el algoritmo k-medias. Este enfoque puede ser especialmente apropiado para la definición de dominios de estimación, dado que divide los datos en grupos basados en las relaciones entre las variables que son relevantes para el problema (Moreira et al., 2020). El agrupamiento automático es un enfoque para analizar datos espaciales a un nivel más alto de abstracción al agrupar de acuerdo con su similitud en grupos significativos (Kisilevich et al. 2009). Se organiza una colección de elementos de datos en grupos, de modo que los elementos dentro de un grupo sean más similares entre sí que a los elementos de los otros grupos. La agrupación en grupos se realiza generalmente cuando no hay información disponible sobre la pertenencia de elementos de datos a clases predefinidas. Por esta razón, se considera tradicionalmente como parte del aprendizaje no supervisado. Existen una gran variedad de enfoques de agrupamiento para diferentes aplicaciones y tamaño de datos (Gan et al. 2007). Algunos de estos métodos incluyen agrupación jerárquica, agrupación particional, agrupación de modelos de mezcla, agrupación basada en redes neuronales, agrupación difusa y agrupación de gráficos (Gan et al. 2007; Everitt et al. 2011; Abu-Jamous et al. 2015).

Uno de los algoritmos más populares y difundidos en el agrupamiento automático es k-medias, el cual corresponde a un método iterativo numérico, no supervisado, no determinista, que es simple y muy rápido, por lo que en muchas aplicaciones prácticas el método ha demostrado ser una forma muy efectiva que puede producir buenos resultados de agrupamiento (Na et al. 2010). Oliver y Willingham (2016) utilizaron la agrupación de k-medias para identificar dominios geológicos en un depósito mineral de hierro en Queensland, Australia, con base en datos de análisis por laboratorio ( $Fe, SiO_2, Al_2O_3, P, TiO_2, TiO_2$  y  $LOI$ ). Rajabinasab y Asghari (2019) utilizaron k-medias, para definir dominios geo-metalúrgicos en un depósito de hierro en el noreste de Irán, utilizando datos de análisis por laboratorio ( $Fe, FeO, S, susceptibilidad magnética, coordenadas X, Y, Z$ ).

Moreira et al. (2020) utilizaron k-medias para definir dominios geológicos de estimación en un depósito de fosfato-titanio, utilizando principalmente datos de análisis por laboratorio ( $P_2O_5, Fe_2O_3, MgO, CaO, Al_2O_3, SiO_2, TiO_2, MnO, Na_2O, K_2O, BaO$  y  $NbO_5$ ).

Sin embargo, k-medias optimiza una función de coste definida en la medida de distancia euclidiana entre puntos de datos y medias. Minimizar la función de coste mediante el cálculo de medias limita su uso a datos numéricos (Huang, 1998). Esta limitación afecta a variables categóricas geológicas que controlan la ley mineral, no pudiendo aportar información en el agrupamiento de forma directa. Dado este problema, se pueden transformar los datos categóricos a valores numéricos discretos, asumiendo un conocimiento absoluto de las asociaciones de las categorías de cada atributo geológico, lo cual es cuestionable. Lo segundo es realizar el proceso inverso, transformar las variables numéricas continuas de las leyes minerales, a discretas, asumiendo una pérdida importante de información, para luego emplear un algoritmo de agrupamiento para datos categóricos como k-modas (Huang, 1998). Sin embargo, como es de esperar, ambos arreglos pueden deteriorar la calidad de la información.

En la literatura asociada a recursos minerales, no se han propuesto métodos de agrupamiento para tratar con datos mixtos, sin embargo, Huang (1998) propuso el primer algoritmo que está basado en una combinatoria de k-medias y k-modas, al cual llamo k-prototipos, el que en este caso posibilita agrupar de manera conjunta leyes minerales y atributos geológicos. El algoritmo agrupa objetos con atributos numéricos y categóricos de una manera similar a k-medias. La medida de similitud de objetos se deriva de atributos mixtos. Cuando se aplica a datos numéricos, el algoritmo es idéntico a k-medias. Al emplear algoritmos de aprendizaje no supervisado, los datos no están etiquetados, por lo que no se conoce la respuesta correcta a priori. En este caso los algoritmos de agrupamiento no jerárquicos requieren ser inicializados indicando el número de grupos (dominios en este caso) como parámetro de entrada. Para la selección del número óptimo de grupos, se emplea el método del codo, discutido por primera vez por Thorndike (1953), utilizando una función de coste mixta (variables numéricas y categóricas).

## 2. Metodología

El algoritmo k-prototipos destaca por permitir usar tanto datos numéricos como categóricos. Dado un conjunto de datos  $X$  que contiene  $n$  objetos, descritos por  $m_r$  atributos numéricos y  $m_c$  atributos categóricos, el objetivo de los k-prototipos es encontrar  $k$  agrupaciones donde se minimiza la siguiente función objetivo:

$$J = \sum_{l=1}^k \sum_{i=1}^n p_{il} d(x_i, Q_l), \quad (1)$$

Donde  $p_{il}$  es un elemento de la matriz  $P_{n \times k}$  que satisface  $0 \leq p_{il} \leq 1$  y  $\sum_{l=1}^k p_{il} = 1$ .  $p_{il} \in \{0, 1\}$  por lo que es una variable binaria que indica la pertenencia del objeto de datos  $x_i$  en el clúster  $l$ ,  $Q_l$  es el centroide o prototipo del clúster  $l$ , y  $d(x_i, Q_l)$  es la medida de disimilitud que se define de la siguiente manera:

$$d(x_i, Q_l) = \sum_{j=1}^{m_r} (x_{ij}^r - q_{lj}^r)^2 + \gamma_l \sum_{j=1}^{m_c} \delta(x_{ij}^c - q_{lj}^c), \quad (2)$$

Aquí  $x_{ij}^r$  representa los valores de los atributos numéricos y  $x_{ij}^c$  representa los valores de los atributos categóricos para cada objeto de datos  $i$ ;  $q_{lj}^r$  es la media del atributo numérico  $j^{th}$  en el grupo  $l$ , y  $q_{lj}^c$  es el valor más común (moda) para los  $j^{th}$  atributos categóricos en el grupo  $l$ ;  $\gamma_l$  es una ponderación de atributos categóricos el grupo  $l$ . La función  $\delta$  que se define para los atributos categóricos es:

$$\delta(x_{ij}^c, q_{lj}^c) = \begin{cases} 0, & x_{ij}^c \neq q_{lj}^c \\ 1, & x_{ij}^c = q_{lj}^c \end{cases} \quad (3)$$

El algoritmo principal del método k-prototipos consiste en calcular la distancia entre los objetos de datos y los centros de cada grupo o prototipos ( $Q$ ) utilizando la ecuación (2). El método finaliza cuando el centro del grupo actualizado coincide con el centro del grupo de la iteración anterior.

Los resultados experimentales mostrados por el método de k-prototipos han funcionado muy bien en conjuntos de datos pequeños y medianos, sin embargo, no logra tratar conjuntos de datos mixtos a gran escala (de millones de instancias), debido al alto costo computacional que requiere un total de  $n \times k$  cálculos de distancia en cada iteración (Vattani, 2011).

En cuanto al flujo de trabajo general, es que se puede resumir de la siguiente manera: Imputación de datos faltantes en la muestra (sondajes), selección de características a emplear en el agrupamiento, análisis exploratorio de datos (AED), estandarización de las características del tipo continuas; aplicación del algoritmo k-prototipos en 9 escenarios  $k$ , selección del

k óptimo y finalmente validación de los dominios de estimación propuestos.

La imputación corresponde al proceso de completar los valores faltantes en los sondeos mineros. La falta de información se puede deber a distintos motivos tales como; no recuperación de muestra en la operación, pérdida de la muestra, contaminación de la muestra, inviabilidad para el análisis de la muestra, presupuesto, entre otros. Un ejemplo de esto último es el caso del cobre soluble, el cual en ácido no puede ser ensayado a menos que el cobre total exceda un umbral de importancia. La imputación es necesaria particularmente en este estudio para la ejecución de los procesos posteriores en el flujo de trabajo. En este caso se utiliza el enfoque de imputación por vecinos más cercanos, el cual utiliza un promedio ponderado de los  $k$  vecinos más cercanos a nivel univariado, usando una métrica del inverso de su distancia euclidiana (Troyanskaya et al., 2011).

$$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}, \quad (4)$$

Donde  $x_i$  es el valor de la característica con valor conocido e  $y_i$  es el punto desconocido de la misma característica.  $k$  es el número de vecinos a utilizar.

Posteriormente se deben seleccionar las características que realmente aportan información en el agrupamiento, es decir, ha de estudiar aquellas características que poseen mayor dependencia con la característica objetivo, en este caso el oro (Au). Para esto se emplea la prueba F de regresión lineal univariante, la cual estima el grado de dependencia lineal entre dos variables aleatorias; primeramente, se realiza el cálculo de la correlación cruzada entre la variable objetivo (Au) y cada uno de los regresores (Ag, Cu, As, Hg y S). La alteración hidrotermal (ALTE), no entra en esta etapa, ya que es del tipo categórico y su participación en el agrupamiento será obligatorio, esto último se explicará con más detalle en el AED. Luego la correlación cruzada entre cada regresor se convierte en una puntuación F.

$$F = \frac{\frac{SCR}{k}}{\frac{SCE}{n-k-1}}, \quad (5)$$

Donde  $SCR$  es la suma de cuadrados de la regresión,  $SCE$  es la suma de cuadrados de los residuos,  $n$  es el número total de observaciones en la muestra y el  $k$  grado de libertad.

El análisis exploratorio de datos (AED) tiene como propósito describir a nivel estadístico y espacial cada una de las características que se encuentran en la muestra. Esta etapa por un lado complementa la selección de características estudiando las distintas correlaciones del Au, y por otro lado evalúa el grado de control que posee la alteración sobre los metales y contaminantes. Ha de entender que la única característica geológica y categórica en el proceso de agrupamiento, será la alteración, pero para justificar su participación ha de demostrar justamente la significancia de esta. Un requerimiento en el agrupamiento automático es que cada una de las características tenga la misma influencia. Esto último lo resuelve la estandarización de datos. En este caso se emplea la Puntuación Z.

$$Z = \frac{X-m}{s}, \quad (6)$$

donde  $Z$  es el valor estandarizado,  $X$  es el valor original,  $m$  es la media, y  $s$ , la desviación estándar de los valores originales. El resultado es una transformación en donde todas las características poseen el mismo valor de varianza.

Posteriormente el algoritmo k-prototipos es ejecutado en 9 escenarios distintos para  $k$ , en un rango de 2 a 10. Esto con motivo de seleccionar el  $k$  óptimo, que se condice con el número de agrupamientos en el que la inercia dentro del grupo es mínima y al mismo tiempo cada grupo se diferencia de manera considerable en su estadístico de valor medio.

Para la selección del número óptimo de grupos, se emplea el método del codo, el cual utiliza la función de costo de Huang (Huang, 1998), en diferentes escenarios de agrupamientos  $k$ . La ecuación de la función de costo para tipos de datos mixtos (numéricos y categóricos) es la siguiente:

$$Cost_l = \sum_{i=1}^k u_{il} \sum_{j=1}^{m_r} (x_{ij}^r - z_{ij}^r)^2 + \gamma_l \sum_{j=1}^{m_c} u_{il} \sum_{j=1}^{m_c} \delta(x_{ij}^c, z_{ij}^c), \quad (7)$$

O de forma simplificada:

$$Cost_l = Cost_l^r + Cost_l^c, \quad (8)$$

donde  $Cost_l^r$  denota el costo total de todas las variables numéricas para todos los objetos dentro del clúster  $l$ .  $Cost_l^c$  se minimiza mientras que  $z_{ij}$  se calcule con la siguiente ecuación:

$$z_{ij} = \frac{1}{n_l} \sum_{i=1}^n u_{il} \cdot x_{ij}; \text{ para } j = 1, 2, \dots, m, \quad (9)$$

donde  $n_l = \sum_{i=1}^n u_{il} \cdot x_{ij}$  es el número de objetos dentro del grupo  $l$ .

Para las variables categóricas,  $C_j$  es un conjunto de valores únicos en cada variable categórica  $j$  y  $p(q_{ij}^c \in C_j | l)$  es la probabilidad de que  $C_j$  este dentro del grupo  $l$ . Entonces, el costo se puede reescribir de la siguiente manera.

$$Cost_l^c = \gamma_l \sum_{j=1}^{m_c} n_l (1 - p(q_{ij}^c \in C_j | l)), \quad (10)$$

Donde  $n_l$  es el objeto dentro del grupo.

Finalmente, los grupos resultantes que son en definitiva los dominios de estimación deben ser validados. Coombes (2016) propone un enfoque de revisión de varianza simple, basado en cuadrículas o similar, como requisito mínimo para la validación de dominios de estimación, sin embargo, en este estudio se emplea de manera adicional el estimador del semivariograma (Matheron, 1963) para detectar dependencia espacial.

$$\gamma^*(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (Z(x_i) - Z(x_i + h))^2, \quad (11)$$

$\gamma^*(h)$  representa la semivarianza en función a un vector  $h$ , que separa una variable de esta en el espacio.

### 3. Caso de estudio

El conjunto de datos disponible es producto de 148 sondajes de exploración, realizados con fines de identificar, caracterizar y evaluar la zona de óxidos en un depósito hidrotermal de oro de alta sulfuración situado en la zona sur de Perú. El metal objetivo o de mayor interés y del cual se buscan definir los dominios de estimación, es el oro (Au). Este metal es naturalmente errático en su distribución, lo que quiere decir que a cortas distancias puede tener variaciones importantes, lo cual dificulta notablemente su modelación a través de técnicas geoestadísticas. Las características disponibles son 10; Oro (Au), plata (Ag), cobre (Cu), Arsénico (As), Mercurio (Hg), Azufre (S), alteración hidrotermal (A), coordenada este (X), norte (Y) y elevación (Z). La información está desplegada en 922 compositos de 10 metros de longitud, que corresponde al soporte o muestra de roca regularizada. Las coordenadas X, Y y Z, identifican al centroide de cada composito.

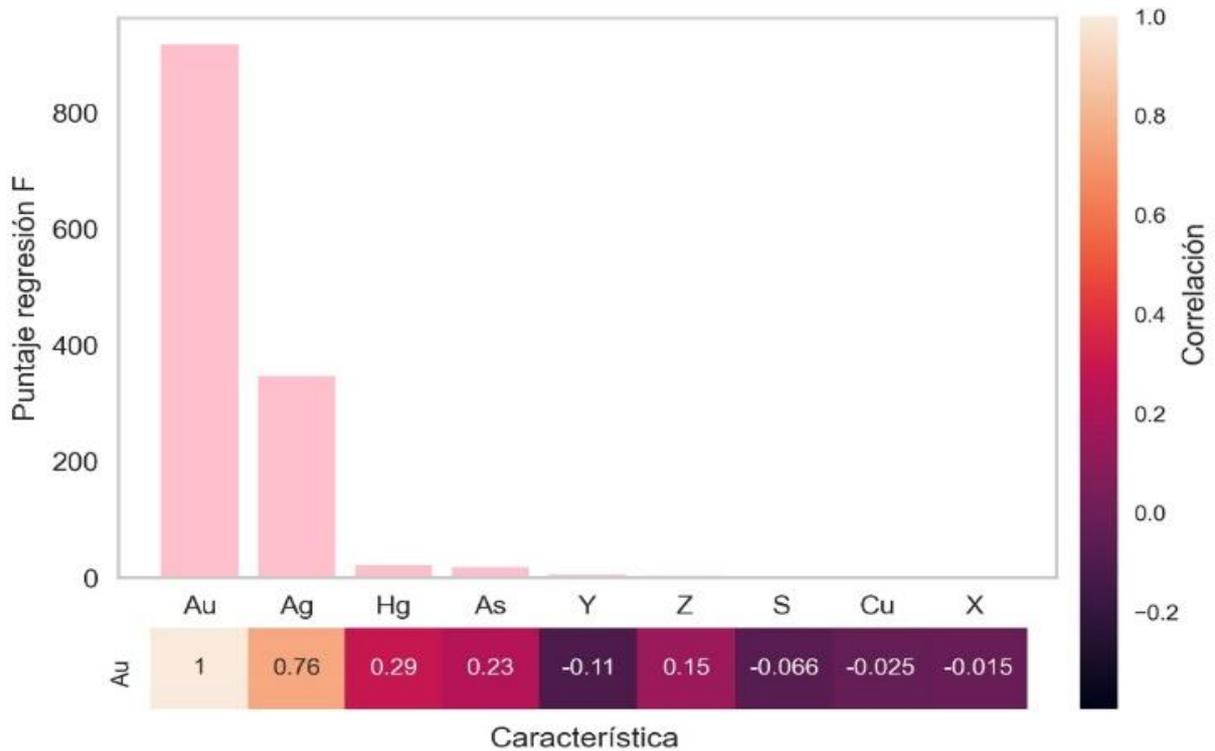
La definición de dominios de estimación, esta dada por un algoritmo de agrupamiento que se alimenta de características que aportan información al metal de interés principal (Au), es por ello por lo que es necesario seleccionar dichas características, y desestimar otras que no representen una mejora en el rendimiento del algoritmo o bien disminuyan su precisión. Inicialmente se detecta que, a excepción del Au y Ag, los demás elementos están sub-muestreados, es decir, no fueron enviados a análisis químico para conocer su concentración. Las técnicas estadísticas de selección de variables basadas en correlación cruzada requieren la misma cantidad de datos entre las características, por lo que es necesario un proceso de imputación. En este caso a través del método de vecinos más cercanos. El proceso comienza buscando en cada uno de los sondajes aquellos compositos que no posean información, posteriormente se detectan los cuatro compositos más cercanos, y a estos se le asigna un peso en función a su distancia al cuadrado. Aquellos compositos más cercanos tendrán una influencia mayor en el valor estimado del composito sin información. Como se observa en la *tabla 1.*, la imputación mantiene valores de media similares y un suavizamiento típico de la interpolación empleada.

**Tabla 1.** Descripción estadística de características del tipo continua (medición en ppm) antes y después de la imputación.

Características	Originales			Imputados		
	N	$\mu$	$\sigma$	N	$\mu$	$\sigma$
Au	922	0.327	0.317	922	0.327	0.317
Ag	922	7.282	6.843	922	7.282	6.843
Cu	240	0.011	0.019	922	0.008	0.011
Hg	172	0.001	0.002	922	0.002	0.001
As	229	0.063	0.077	922	0.062	0.052
S	166	0.710	0.780	922	0.670	0.421

Aplicando el método de selección por regresión F, es que resulta que la característica Ag es notoriamente superior al resto en cuanto a la relación con el Au, siendo las demás poco correlacionales y no significativas para incorporarse en el agrupamiento. De esto último se entiende que, a nivel de variables numéricas, solo se utilizaran en el agrupamiento Au y Ag, desestimando el uso de las demás (Cu, Hg, As y S).

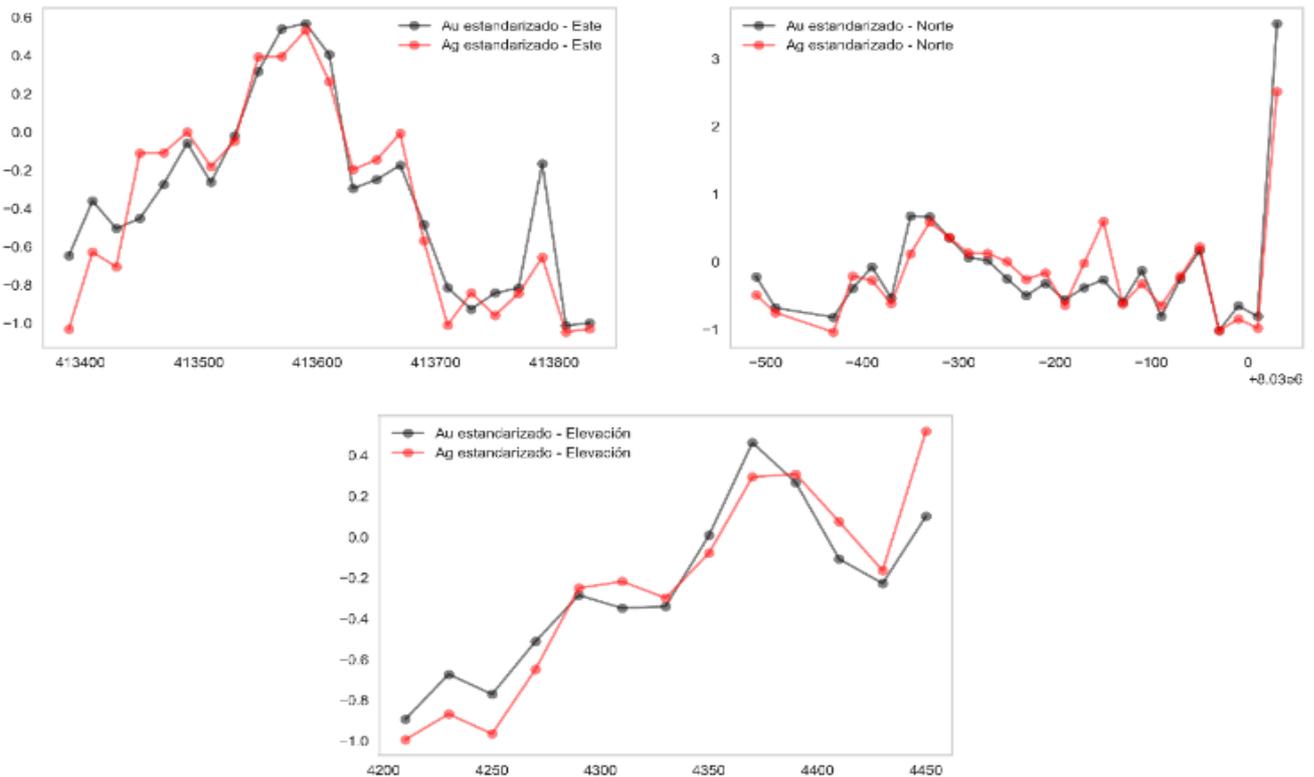
A diferencia de Fouedjio (2016), Rajabinasab y Asghari (2019), quienes incluyeron en sus análisis de agrupamiento las coordenadas X, Y, Z fundamentando



el hecho de que las ubicaciones geográficas de las muestras consideran la dependencia espacial de las variables, en este estudio no se comparte la misma apreciación, ya que por un lado resulta artificial el hecho de que variables de localización asuman una mayor influencia en el agrupamiento que la

la cual incorpora de forma implícita las coordenadas X, Y y Z en su naturaleza.

A nivel de estructura espacial, el Ag se mantiene en una relación directa con el Au como se puede apreciar en las gráficas de deriva de la *figura 2*.



característica principal, luego el Au es una variable regionalizada que posee su propia estructura espacial,

**Figura 2.** Derivas Au-Ag estandarizadas

Desde el punto de vista geológico, la zona mineral en estudio está compuesta por óxidos de hierro; hematitas,

jarosita y goethita. Estos óxidos usualmente en este tipo de yacimiento contienen oro en altas proporciones. En la zona mineral se presentan cinco tipos de alteraciones hidrotermales; argílico avanzado, argílico intermedio, silicificación, cobertura del suelo y vuggy. La alteración hidrotermal es un complejo proceso que involucra cambios mineralógicos y químicos, esta crea una especie de control en el yacimiento junto con otros atributos geológicos como pueden ser las estructuras y la litología.

La alteración argílica se define por la formación de minerales arcillosos producto de una intensa lixiviación ácida, un subtipo de esta alteración se define como "intermedio" por la presencia de montmorillonita, illita, clorita, arcillas del grupo caolín y sericita. La alteración "avanzada" es resultado del ataque ácido intenso, algunos minerales de esta alteración son dickita, caolinita, pirofilita, barita, alunita y diáspora. La silicificación es un proceso que ocurre cuando el fluido con sílice, encuentra zonas con menor temperatura, y la sílice precipita y se deposita en la roca. La cobertura del suelo se

depósitos de alta sulfuración, el agua subterránea y el dióxido de azufre condensado crean ácido, este ácido en interacción con la roca genera esta alteración residual.

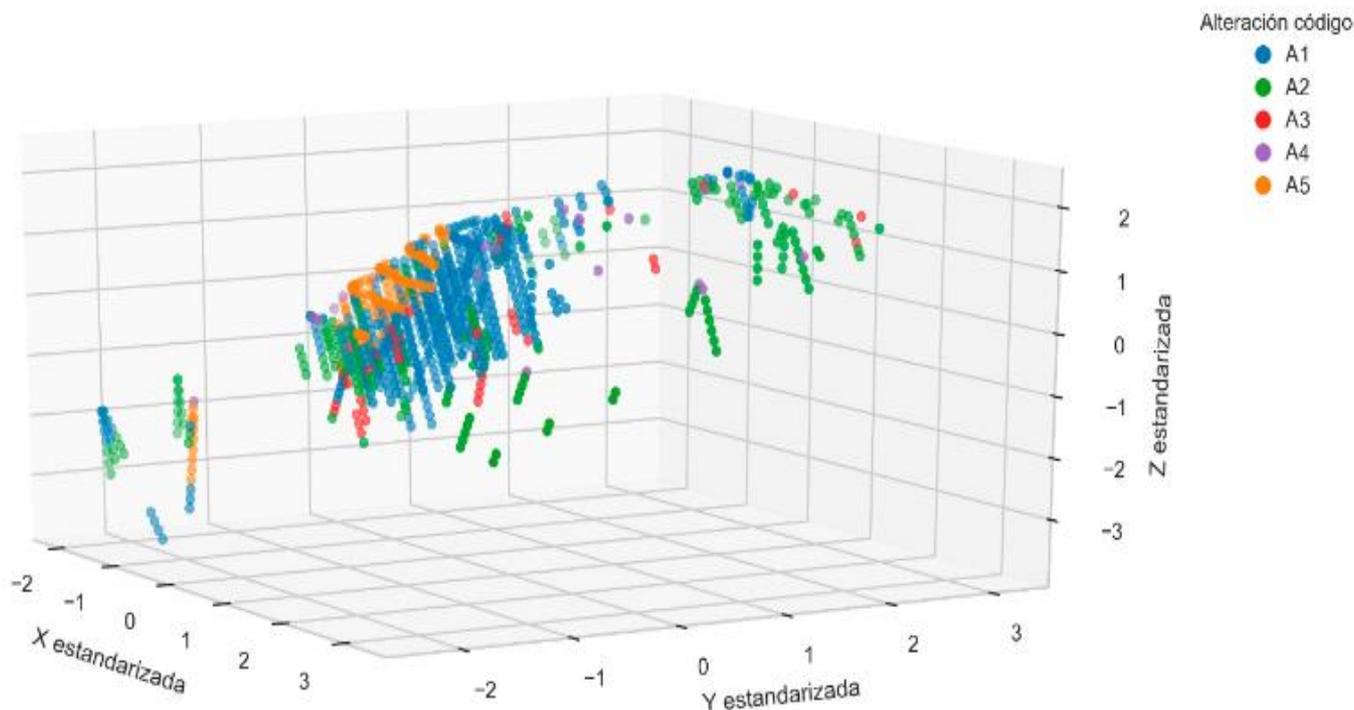
Las alteraciones predominantes en el depósito son del tipo argílica (A1 y A2), ocupando el 80% de la muestra, luego en abundancia sigue el vuggy (A5) con un 13% y

en menor cantidad la cobertura de suelo (A4) y la silicificación (A3), con un 4% y 3% respectivamente. La roca encajonante de la zona mineral de óxidos corresponde a alteraciones vuggy, argílico avanzado y argílico intermedio, la alteración vuggy es más favorable para que se deposite el oro, luego en el argílico avanzado y en menor grado se puede depositar en el argílico intermedio.

La alteración hidrotermal es la característica geológica que mantiene control sobre la ley mineral de Au y se incorpora en el agrupamiento como única variable categórica.

**Tabla 2.** Estadísticas desagrupadas de Au por tipo de alteración

**Figura 3.** Vista isométrica de alteraciones en sondajes de exploración



refiere a la vegetación, incluidos los cultivos y los residuos de cultivos en la superficie del suelo. Vuggy es el resultado de una intensa lixiviación ácida en

cód	descripción	N	$\mu$	$\sigma$	mín	máx
A1	Argílico avanzado	518	0.30	0.20	0.00	1.58

A2	Argílico intermedio	216	0.13	0.22	0.00	1.44
A3	Silicificación	41	0.32	0.21	0.03	0.87
A4	Cobertura del suelo	24	0.22	0.11	0.01	0.36

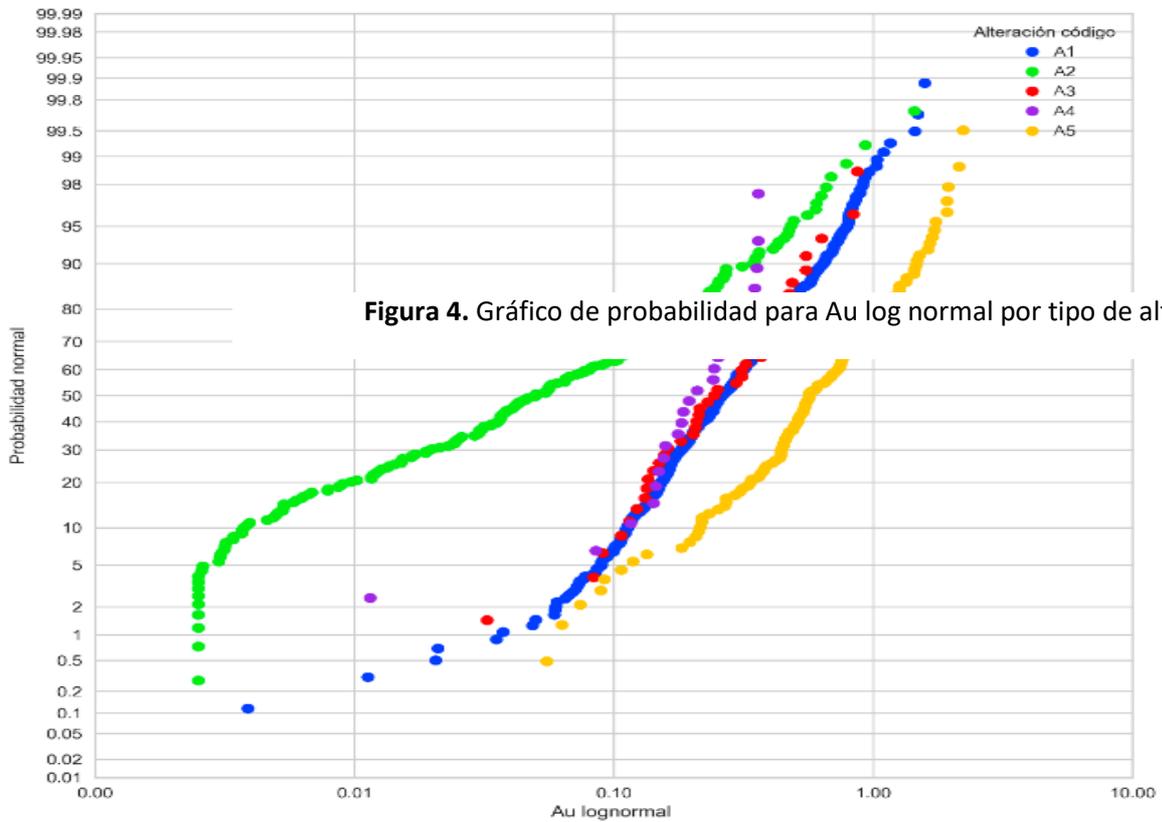


Figura 4. Gráfico de probabilidad para Au log normal por tipo de alteración

En la tabla 2. se presentan las estadísticas desagrupadas del Au, este proces

A5	Vuggy	123	0.61	0.47	0.06	2.22
----	-------	-----	------	------	------	------

A diferencia de A1 con A3, cada alteración mantiene una distribución y ley media diferenciable, lo cual justifica el hecho de que esta característica mantenga un control sobre el Au. Es posible en este caso el poder unificar estas alteraciones (A1 y A3), dado el tamaño que tiene

A3 y la similitud estadística considerable que estas presentan. Sin embargo, esta decisión es parte de la metodología tradicional, y no se modificará en vías de evaluar el algoritmo de agrupamiento automático.

o se lleva a cabo por la disposición espacial de los sondeos donde se presenta un agrupamiento considerable en la alteración argílica avanzada y vuggy. El método de desagrupamiento empleado es el de celdas, donde el óptimo tamaño es de 80 metros en los tres ejes formando un cubo. Ha de considerarse que las dimensiones en cada eje del espacio para la zona en estudio son de 450, 550 y 250 metros en X, Y y Z respectivamente.

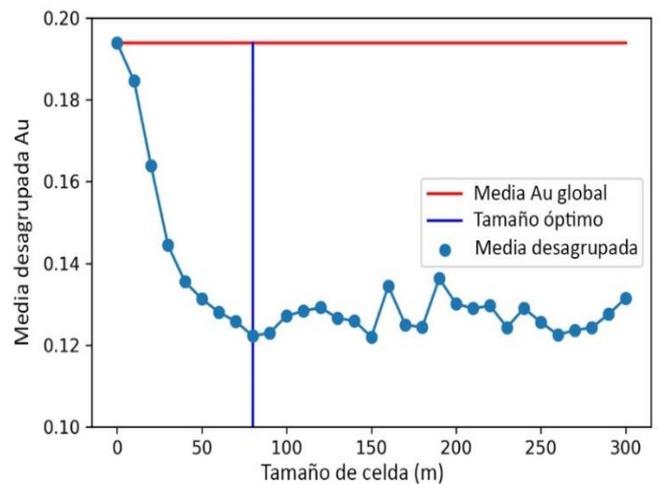
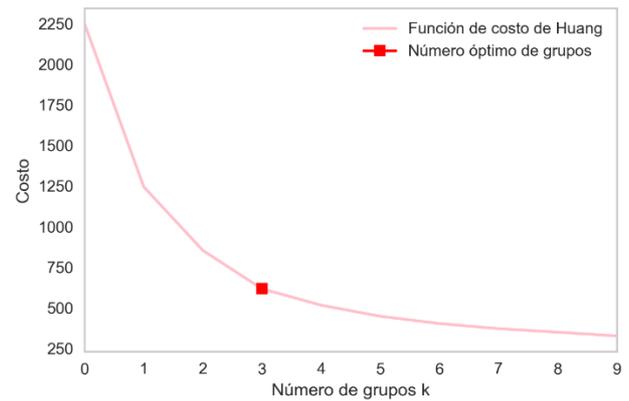


Figura 5. Media desagrupada de Au por tamaño de celda

Como se ha señalado anteriormente, el Au y Ag son características que poseen distinta escala y varianza, por lo que, para un correcto aporte de información de cada una de ellas en el agrupamiento, se estandarizan a



**Figura 6.** Gráfico del método del codo

El método del codo recomienda 3 dominios como se observa en la *figura 6*, donde se encuentra la mayor inflexión. Cada uno de estos dominios se diferencia claramente a nivel estadístico como espacial. El primer dominio (D1) es el de mayor tamaño y se compone de leyes de Au intermedias promediando 0.34 ppm y siendo compuesto mayormente por la alteración del tipo argílica avanzada (A1) ocupando un 81% del dominio. D2 concentra leyes de Au bajas en promedio 0.09 ppm y se compone mayormente de la alteración argílica intermedia (A2), la cual ocupa un 70% del dominio. Finalmente, D3 que es el dominio de menor tamaño, concentra las leyes altas, promediando 0.96 ppm y compuesto en su mayoría por la alteración Vuggy (A5), la cual ocupa el 55% del dominio. Las alteraciones A3 y A4 se presentan con una baja frecuencia en los dominios D1 y D2.

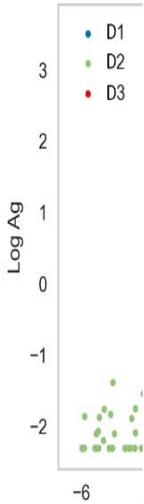
través de una transformación de puntuación z, la cual resulta en que las dos variables presenten media 0 y varianza 1.

El algoritmo k-prototipos es aplicado en 9 escenarios distintos, desde k=2 hasta k=10, donde k, es el número de dominios geológicos propuesto. Luego los k escenarios son evaluados con el método del codo utilizando la función de costo de Huang.

**Tabla 3.** Estadísticas sobre dominios y relación con característica geológica

Dominio	Alteración predominante	Cantidad	Media	Desviación
D1	A1	538	0.34	0.18
D2	A2	286	0.09	0.08
D3	A5	98	0.96	0.43

**Figura 8.** Gráfico de barras con alteraciones por dominio geológico



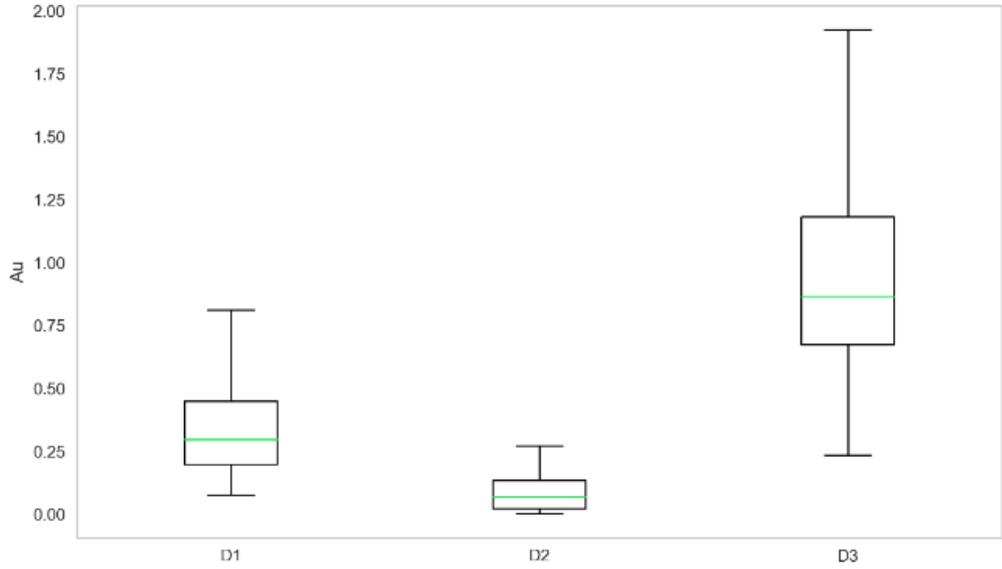
**Figura 7.**  
Diagrama de dispersión

Au log normal vs Ag log normal por dominios de

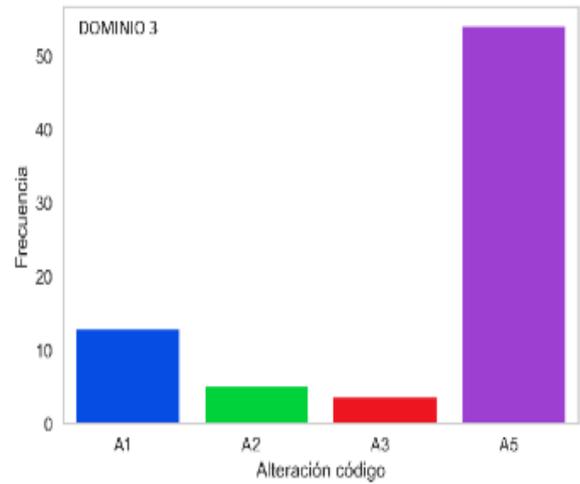
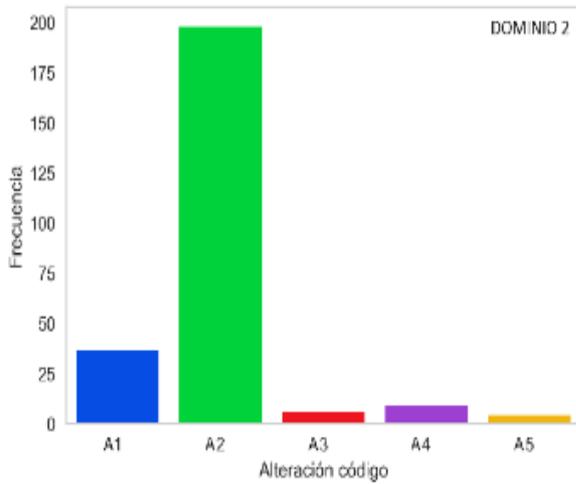
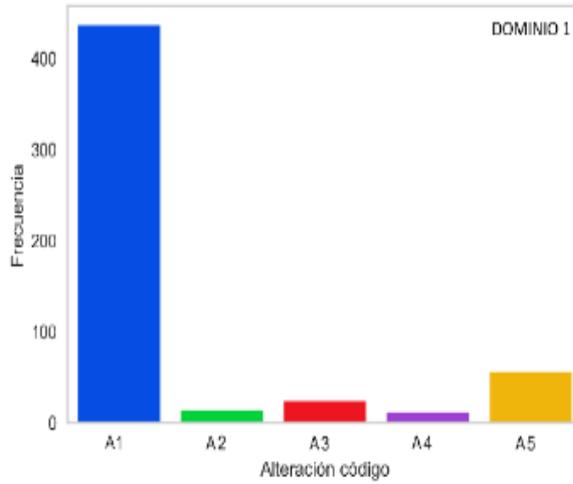
El resultado del agrupamiento automático mixto ha separado claramente los dominios en conjuntos de leyes

bajas, intermedias y altas de Au, cada cual con su

propia distribución. Además, ha respetado la diferenciación del tipo geológica a través de la alteración hidroter



mal.



estimación

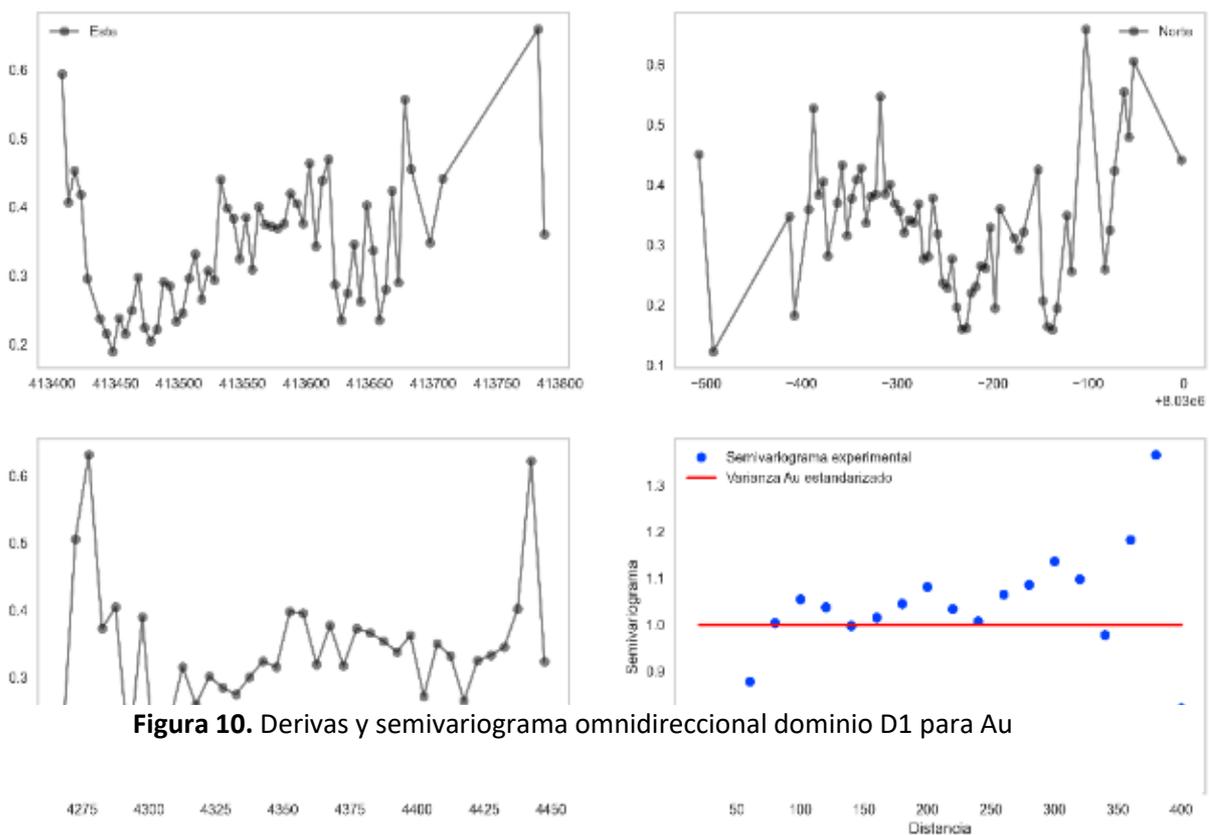
El dominio D2 presenta derivas con varianzas constantes en sus tres ejes, lo cual es suficiente para avanzar en un modelo geoestadístico. En cuanto al semivariograma omnidireccional, este no es tan fácil de

**Figura 9.** Diagrama de cajas por dominio de estimación

En cuanto al análisis espacial, D1 presenta una continuidad entre 100 y 150 metros, la cual puede ser modelable con una única estructura del tipo esférica. Este dominio es el más robusto en información, y presenta una estacionaridad local suficiente para el empleo del modelo geoestadístico.

modelar como D1, sin embargo, con dos estructuras anidadas (gaussiana + esférica), podría tomar una buena forma.

Finalmente, D3 es el dominio más complejo, ya que es el menos denso en información y de mayor varianza, esto último denota una mayor variabilidad en el comportamiento del oro, haciendo honor a su naturaleza errática. Este dominio no podría ser modelado y estimado con técnicas geoestadísticas,



**Figura 10.** Derivas y semivariograma omnidireccional dominio D1 para Au

predominante en D3, y en caso de haber utilizado la metodología tradicional, tampoco sería posible el uso del modelo geoestadístico.

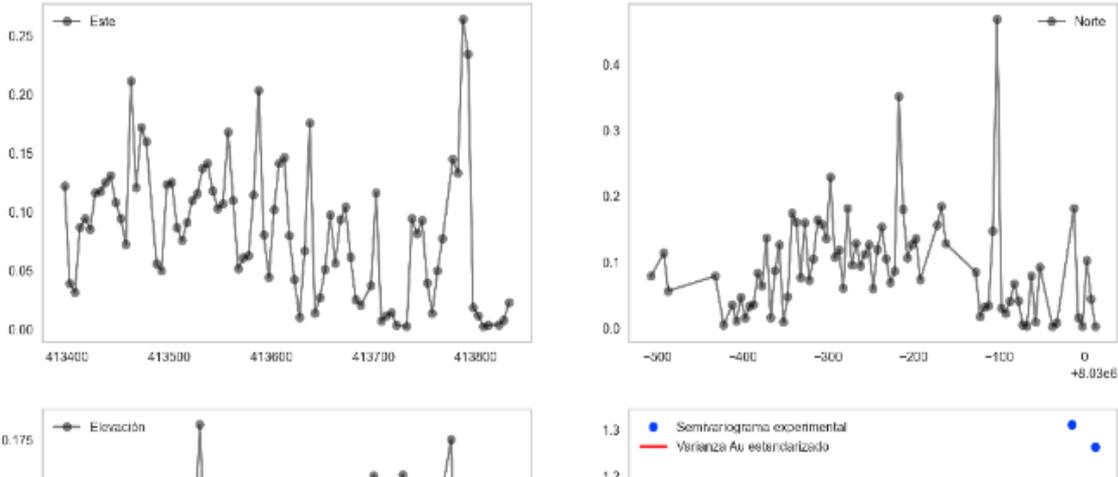


Figura 11. Derivas y semivariograma omnidireccional dominio D2 para Au

debe  
ndo  
utiliza  
rse  
otro  
méto  
do de  
estim  
ación

que no tenga como requerimiento la

#### 4. Conclusiones

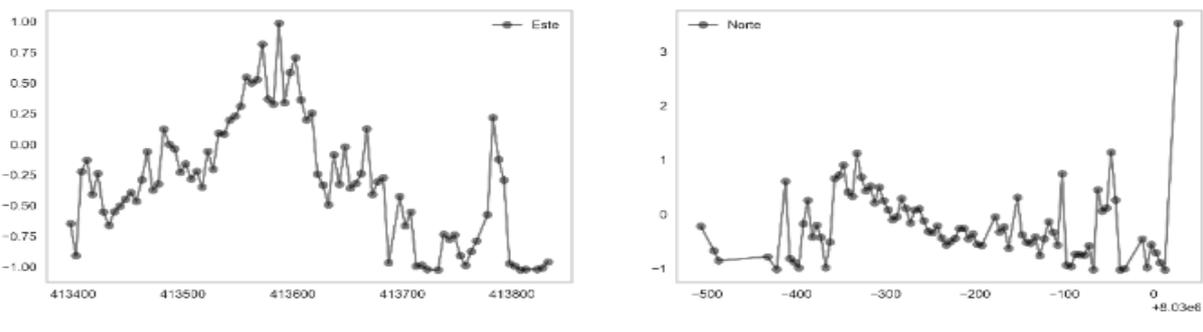


Figura 12. Derivas y semivariograma omnidireccional dominio D3 para Au

estacionaridad. Para justificar esta decisión, se ha realizado el mismo análisis de derivas y de semivariograma sobre la alteración vuggy, la cual es

s, ha demostrado que es posible definir dominios geológicos de estimación, mezclando variables

Los resultados de la implementación del algoritmo de agrupamiento k-prototipo

numéricas y categóricas en un proceso automático mixto. Para esto, inicialmente es necesario que exista una correcta selección de características, es decir, aquellas del tipo numéricas deberán mantener una alta correlación ya sea positiva o negativa y compartir una estructura espacial relativamente similar. En cuanto a las variables categóricas, estas deben ser atributos geológicos que posean control sobre la ley mineral, y entreguen la coherencia geológica requerida en la definición de cada dominio. En segundo lugar, es fundamental la preparación de las características seleccionadas, por parte del tipo numérica es necesaria su transformación para que todas aporten la misma varianza. En cuanto a las variables del tipo categóricas, es requerida una codificación binaria, se ha comprobado que si este preproceso no es realizado, el aporte de la característica geológica pierde peso.

El enfoque propuesto en este estudio no reemplaza la metodología tradicional, ya que en este se asume un detalle mayor debido a la supervisión del equipo geomodelador, sin embargo el enfoque automático se presenta como una alternativa comparativa o complemento más eficiente en cuestión de tiempo, homogeneidad dentro del dominio y heterogeneidad entre dominios geológicos.

## Referencias

1. Abu-Jamous, B., Fa, R., Nandi, A.K. 2015. Integrative cluster analysis in bioinformatics. John Wiley & Sons, p.143-154.
2. Coombes, J. (2016. I'd Like to be OK with MIK, UC?: A Critique of Mineral Resource Estimation Techniques. Coombes Capability, p. 162-177.
3. Dominy, S.C., Stephenson, P.R., Annels, A.E. 2001. Classification and reporting of mineral resources for high-nugget effect gold vein deposits. *Exploration and Mining Geology*, v. 10, p. 215-233.
4. Emery, X., Ortiz, J. 2004. Defining Geological Units by Grade Domaining. Department of Mining Engineering, University of Chile, p. 1-13.
5. Everitt, B. S., Landau, S., Leese, M., Stahl, D. 2011. Cluster analysis. John Wiley, p. 71-110.
6. Fouedjio, F. 2016. A hierarchical clustering method for multivariate geostatistical data, *Spatial Statistics*. Science Direct, v. 18, p. 333-351.
7. Gan, G., Ma, C., Wu, J. 2007. Data Clustering: Theory, Algorithms, and Applications. American Statistical Association, p. 43-52.
8. Glacken, I. & D. Snowden. 2001. Mineral Resource Estimation. The Australasian Institute of Mining and Metallurgy: Melbourne, p. 189-198.
9. Huang, Z. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, v. 2 , p. 283-304.
10. Journel, A.G., Huijbregts, C.J. 1978. Mining geostatistics. Academic press, v.600, p. 236-293.
11. Kisilevich, S., Mansmann, F., Nanni, M., Rinzivillo, S. 2009. In *Data mining and knowledge discovery handbook*. Springer, p. 855- 874.
12. MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* , v. 1, p. 281-297.
13. Matheron, G. 1963. Principles of geostatistics. *Economic geology*, v. 58, p. 1246-1266.
14. Moreira, G.D.C., Coimbra, J.F., Marques, D.M. 2020. Defining geologic domains using cluster analysis and indicator correlograms: a phosphate-titanium case study. *Applied Earth Science*, v. 129, p. 176-190.
15. Na, S., Xumin, L., Yong, G. 2010. Research on k-means clustering algorithm: an improved k-means clustering algorithm. In: 2010 third international symposium on intelligent information technology and security informatics, p. 239-244.
16. Oliver, S., Willingham, D. 2016. Maximum orebody value through the automation of resource model development using machine learning. In *The third AusIMM international geometallurgy conference*, Perth, Australia, p. 295-301.
17. Rajabinasab, B., Asghari, O. 2019. Geometallurgical domaining by cluster analysis: Iron ore deposit case study. *Natural Resources Research*, v. 28, p. 665-684.
18. Rossi, M.E., Deutsch, C.V. 2013. Mineral resource estimation. Springer Science & Business Media, p. 241-265.
19. Sokal, R.R., Camin, J.H., Rohlf, F.J., Sneath, P.H.A. 1973. Numerical taxonomy: some points of view. *Systematic Zoology*, v. 14, p. 237-243.

20. Sterk, R., de Jong, K., Partington, G., Kerkvliet, S., Van de Ven, M. 2019. Domaining in Mineral Resource Estimation: A Stock-Take of 2019 Common Practice, p. 1-13.
21. Thorndike, R.L. 1953. Who belongs in the family?. Psychometrika, v. 18, p. 267–276.
22. Troyanskaya, O., Cantor, M., Sherlock, G., Hastie, T., Tibshirani, R., Botstein, D., Altman, R. 2001. Missing Value Estimation Methods for DNA Microarrays. Bioinformatics, v. 17, p. 520-525.
23. Vattani, A. 2011. K-means requires exponentially many iterations even in the plane. Discrete & Computational Geometry, v. 45, p. 596-616.

**Nombre autor:** Heber Hernández Guerra

**Cargo:** Académico

**Empresa:** Universidad Santo tomas

**Correo electrónico:** [hhernandez9@santotomas.cl](mailto:hhernandez9@santotomas.cl)

**Celular:** 9-99210201

**Dirección:** Av. Ejército Libertador 146, Santiago, Chile

**Perfil profesional:**

Ingeniero civil de minas, master en minería, fundador y director general de Nube Minera e Ingeniería SPA, ha sido director académico y de carrera en Universidad de Aconcagua, docente en cátedras relativas a Evaluación de Yacimientos en distintas instituciones de educación superior, investigador en temas relacionados al estudio de variables espaciales, donde ha publicado en importantes revistas científicas tales como Minerals Journal MDPI y Sustainability Journal MDPI. Actualmente es académico en Universidad Santo Tomas donde además imparte las cátedras y laboratorios de geoestadística y planificación minera.

**Nombre coautor (1):** Cinthia Gladbet Lizzet Barboza Bravo.

**Cargo:** Estudiante ingeniería de minas.

**Empresa:** Universidad Nacional de Cajamarca, Perú.

**Correo electrónico:** [cbarbozab17\\_1@unc.edu.pe](mailto:cbarbozab17_1@unc.edu.pe)

**Celular:** +51 950 073165

**Dirección:** Av. Atahualpa Km. 3, Cajamarca, Perú

**Perfil profesional:**

Estudiante de X ciclo de Ingeniería de Minas en la Universidad Nacional de Cajamarca, dentro del décimo superior, ha sido becaria de Beca Permanencia 2021 por PRONABEC, becaria del programa Cantera de Talentos para la Minería por el Instituto de Ingenieros

de Minas del Perú, coordinadora en el área de eventos y comunicaciones en la Red de Ingenieras de Minas del Perú. Ha sido asistente en el área de geoestadística en Nube Minera e Ingeniería SPA.

Actualmente trabaja como practicante pre profesional en la Dirección Regional de Energía y Minas Cajamarca.

**Nombre coautor (2):** Elisabete Alberdi Celaya

**Cargo:** Académica del Departamento de Matemática Aplicada

**Empresa:** Universidad del País Vasco, España.

**Correo electrónico:** [elisabete.alberdi@ehu.eus](mailto:elisabete.alberdi@ehu.eus)

**Celular:** +34-946-017-790

**Dirección:** 48013 Bilbao, España

**Perfil profesional:**

Licenciada en matemáticas, master en ingeniería computacional y sistemas inteligentes y doctora en ciencias matemáticas de la Universidad del País Vasco/Euskal Herriko Unibertsitateal. Trabaja actualmente como académico de planta en el Departamento de Matemática Aplicada de la Universidad del País Vasco y está dedicada a la investigación en Matemática Aplicada donde posee más de 50 publicaciones en revistas científicas.