Dpto. de Estadística e Investigación Operativa Aplicadas y Calidad

Predicción de Resultados Metalúrgicos en Flotación de Minerales mediante Análisis Multivariante y Aprendizaje Automático

Trabajo Fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

Autor: George Carlos Mogollón Gonzales

Tutor: Alberto José Ferrer Riquelme

Curso académico 2017-2018

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Listado de Siglas

AuGra	Concentración de oro en el concentrado flotación
Aukinetic	Velocidad de flotación de oro en la flotación
AuRmax	Recuperación de oro en el concentrado flotación
CuGra	Concentración de cobre en el concentrado de flotación
Cukinetic	Velocidad de flotación de cobre en la flotación
CuRmax	Recuperación de cobre en el concentrado flotación
CV	Validación cruzada
IC	Índice de condicionamiento
ICP	Plasma de acoplamiento inductivo
LV	Variables latentes
MassRo	Recuperación de espumas en el concentrado flotación
MLR	Regresión lineal múltiple
MSE	Error cuadrático medio
NIPALS	Iteración no lineal de mínimos cuadrados parciales
PCA	Análisis de componentes principales
PLS	Proyección en estructuras latentes
Recup.	Recuperación de concentrado
RF	Bosques aleatorios
SPE	Error cuadrado de predicción
SSQ	Suma de cuadrados
SVM	Máquinas de soporte vectorial

Resumen

En un proyecto metalúrgico, un problema a resolver es la cuantificación de la recuperación, la concentración y la velocidad cinética del mineral valioso que se espera obtener del material que será procesado. Uno de los métodos para realizar la cuantificación es llevar a cabo el proceso metalúrgico de flotación, el cual consiste en la separación del material valioso del no valioso en un laboratorio metalúrgico. Debido a que el modelado del complejo proceso de flotación es complicado, es esencial poder identificar variables con mayor poder explicativo para obtener un modelo lo más parsimonioso posible. En el presente trabajo de fin de máster se han utilizado metodologías de análisis multivariante y técnicas de aprendizaje automático para predecir los resultados metalúrgicos, las cuales podrían ser utilizadas, posteriormente, en la predicción de todo el yacimiento geológico. Se han logrado resultados satisfactorios en la predicción de la concentración de cobre y oro en el concentrado de flotación, demostrando la robustez de las técnicas aplicadas como herramientas de predicción.

Palabras clave: análisis multivariante, aprendizaje automático, flotación de minerales, modelado.

Abstract

In a metallurgical project, a problem to solve is the quantification of the recovery, the degree and the kinetic rate of the valuable mineral that is expected to be obtained from the material that will be processed. One of the methods to perform the quantification is to carry out the metallurgical froth flotation process, which consists of the separation of the valuable material from gangue minerals, in a metallurgical laboratory. Because the modelling of the complex flotation process is complicated, it is essential to identify variables with greater explanatory power to obtain a model as parsimonious as possible. In the present master's thesis, a methodology of multivariate analysis and machine learning techniques is applied and used to predict the metallurgical performance, which could later be used for the evaluation all the geological deposit. The model achieves satisfactory results in the prediction of the copper and gold grade in the flotation concentrate, demonstrating the robustness of the applied techniques as prediction tools.

Keywords: Multivariate analysis, machine learning, mineral froth flotation, modelling

Capítulo 1

1 Introducción

La producción de concentrado de mineral consiste en extraer materias primas de yacimientos geológicos. Para lograr este objetivo, se llevan a cabo diversas etapas de procesamiento del mineral. Uno de los procesos de mayor importancia es la separación de material valioso del no valioso mediante el proceso de flotación de minerales aprovechando propiedades físico-químicas de las partículas (Wills & Napier-Munn, 2006). En el yacimiento geológico, el cual se divide en bloques cúbicos, variables como la concentración de cobre, por ejemplo, son conocidas por cada bloque cúbico. Sin embargo, el valor de la concentración no es tan importante como determinar el comportamiento de cada bloque cúbico en su procesamiento de flotación.

El modelado matemático de flotación permitirá la optimización en la extracción del material valioso, incluyendo el hecho de encontrar oportunidades de mejora en el diseño de las celdas de flotación para evaluar el diagrama de flujo óptimo para el circuito de flotación (Wills & Napier-Munn, 2006). El modelado matemático se realiza sobre muestras representativas, las cuales se extraen tratando de abarcar todo el yacimiento geológico. Las pruebas de flotación sobre estas muestras son costosas, de tipo destructivo y toman mucho tiempo para ser realizadas, provocando la escasez de la información para realizar el modelado.

El problema en la predicción de las respuestas metalúrgicas de un proceso complejo como el de la cinética de flotación de minerales es que muchas de las fórmulas teóricas convencionales de procesamiento de minerales no son

aplicables. Según Wills et al. (2006), existen tres tipos de modelos matemáticos que pueden ser aplicados: teóricos, empíricos y fenomenológicos. Los modelos empíricos están basados en datos y son los que se utilizan en la presente tesis. Los modelos teóricos están basados en principios científicos y requieren mucho conocimiento del proceso. Sin embargo, debido a la complejidad de los procesos, como el de flotación, estos son difíciles de obtener. Por último, los modelos fenomenológicos son una combinación de modelos teóricos y empíricos, que tratan de dar una explicación científica a aquellos parámetros que son importantes en los modelos empíricos.

Para el desarrollo de modelos empíricos que involucran muchas variables, como es el caso del proceso de flotación de minerales, existen diversas técnicas. Por una parte están las técnicas multivariantes de proyección sobre variables latentes que, aprovechando la alta correlación de las variables, construyen nuevas variables latentes que son combinaciones lineales de las variables originales (MacGregor & Kourtl, 1995). Por otro lado, también pueden utilizarse técnicas de aprendizaje automático pertenecientes al campo de la inteligencia artificial y que tienen por objetivo crear sistemas que “aprendan” automáticamente, entendiendo por aprender el hecho de ser capaces de identificar patrones complicados en los datos. Los algoritmos que aprenden pueden predecir comportamientos futuros, mientras que, el término automático se refiere a que los algoritmos buscados se mejoran recursivamente sin intervención de ningún experto para conseguir una mejor representación del proceso (Chollet & Allaire, 2017). Ambos enfoques permiten encontrar relaciones estadísticas entre las variables predictoras y las variables respuestas, con el objetivo de predecir el comportamiento de las respuestas en función de las predictoras.

En el presente trabajo, se analiza un proyecto metalúrgico que se desarrolla sobre un yacimiento de cobre y oro, en Perú, de donde se extrajeron muestras

geológicas del yacimiento y, sobre cada una de ellas, se realizaron caracterizaciones químicas para luego ser sometidas al proceso de flotación.

Los objetivos de este trabajo fin de master son:

1. Obtener un modelo mediante análisis multivariante y técnicas de aprendizaje automático para predecir las siguientes respuestas:
    a. Recuperación del mineral valioso
    b. Calidad del mineral valioso recuperado
    c. Constante cinética de minerales
2. Comparar los resultados obtenidos de las técnicas de análisis multivariante y de aprendizaje automático.
3. Proponer un modelo que permita extrapolar los resultados metalúrgicos de muestras al yacimiento geológico.
4. Implementar una metodología para detectar datos anómalos en una base de datos geoquímicos.

El documento consta de cinco capítulos. En el primer capítulo, se describe un bloque introductorio, los objetivos que se pretender abordar y la justificación del trabajo. El segundo capítulo corresponde al marco teórico del análisis multivariante y del aprendizaje automático, acompañados del lenguaje de programación utilizado para el análisis estadístico, siendo R y Aspen ProMV los softwares utilizados. En el tercer capítulo se detallan las características del proyecto metalúrgico, el ambiente donde se desarrolla, de dónde se obtuvieron las muestras y el experimento del proceso de flotación de minerales. En el cuarto capítulo se propone una metodología para predecir los resultados metalúrgicos basados en análisis multivariante y aprendizaje automático. Finalmente, en el quinto capítulo se presentan las conclusiones.

Capítulo 2

2 Marco Conceptual

En la actualidad, grandes cantidades de datos se registran en muchos proyectos de ingeniería. Un objetivo implícito de obtener muchos datos es confirmar lo que sabemos del proceso o lo que podríamos aprender del mismo. Otro objetivo muy importante es la capacidad de realizar predicciones utilizando herramientas que puedan manejar grandes cantidades de datos.

2.1 Datos a Enfrentar

En la industria moderna, en la que se recoge una gran cantidad de datos, nos enfrentamos a muchos tipos de datos con diferentes características:

- Alta dimensionalidad de los datos, es decir que se miden muchas variables.
- Colinealidad, las variables X (predictoras) y/o las variables Y (respuestas) no son independientes.
- No existe una relación causa-efecto en la mayoría de las relaciones encontradas.
- Datos con información faltante, debido a fallos en los sensores o a valores fuera de los límites de detección.
- Bajo ratio señal-ruido, debido a la inestabilidad del proceso y a que muchos de los datos no son informativos (no proceden de ningún diseño experimental).

Registrar datos no es tarea fácil, generando costes que muchas veces pueden ser elevados y, en ocasiones, obligando a destruir las muestras.

Una base de datos de entrada puede representarse por una matriz, la cual llamaremos **X**, con una cantidad de individuos u observaciones *N*, y con un número de variables *K*. Así mismo, una base de datos de respuestas que acompaña a la base de datos de entrada, la llamaremos **Y**, con la misma cantidad de individuos y una cantidad de respuestas *M*. La Figura 1 muestra la representación de este esquema.

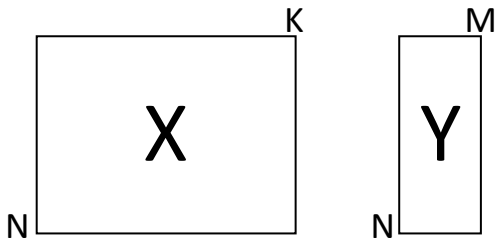

Figura 1. Estructura de matriz de datos

2.2 Regresión Lineal

En la regresión lineal múltiple, la solución por mínimos cuadrados está dada por: $Y = XB + \epsilon$. Donde: $B = (X^T X)^{-1} X^T Y$ (Dunn, 2018). Para el caso en que se tenga una sola variable respuesta, M = 1, por ejemplo, y considerando una base de datos grande y compleja, por lo general, las columnas X presentan algún tipo de correlación (colinealidad entre variables), pudiendo aparecer problemas a la hora de calcular $(X^T X)^{-1}$, para obtener los coeficientes de regresión. Si la colinealidad es exacta, la matriz $\boldsymbol{X^T X}$ es singular (determinante =0) y no existe la inversa; sí la colinealidad no es exacta, la matriz ($\boldsymbol{X^T X}$) es casi singular (determinante ~ 0), los valores de la inversa de la matriz ($\boldsymbol{X^T X}$) son muy grandes, generando problemas en la significación estadística de los coeficientes de regresión (puesto que las varianzas de los estimadores mínimocuadráticos son muy altas). Para solucionar el problema tendríamos que eliminar variables lo cual conllevaría a otro problema, el de perder información y dificultar la comprensión de la estructura de relación entre **X** e **Y**.

En la detección de la colinealidad, el factor de inflación de la varianza (VIF), indica de qué manera la varianza es “inflada” o incrementada debido a la colinealidad. Cuanto más grande sea este valor, superior a 10, habrá mayor correlación entre las variables. Por otro lado, el índice de condicionamiento (CI), el cual está definido por: $\text{CI} = \sqrt{\lambda_{max}/\lambda_{min}}$, donde λ, es el valor propio de la matriz de datos **X**, diagnostica la colinealidad multivariante. Si el IC es superior a 30, existe un grave problema de colinealidad (Gujarati & Porter, 2009).

2.3 Variables Latentes

Conceptualmente, una variable latente puede ser descrita como una variable que no es medida, no es observable directamente. Matemáticamente puede ser expresada como una combinación lineal de las variables primitivas u originales (x_1, x_2, … x_n) con sus respectivos pesos (p_1, p_2, … p_n). Geométricamente se puede definir como la recta que pasa por el origen y que logra capturar la máxima variabilidad de la proyección de los puntos sobre dicha recta (Dunn, 2018).

2.4 Análisis de Componentes Principales

El análisis de componentes principales (PCA) es una técnica multivariante que analiza una tabla de datos en la que las observaciones se describen mediante diversas variables intercorrelacionadas. Su objetivo es extraer información importante de la tabla, para representarla como un conjunto de nuevas variables ortogonales llamadas componentes principales, mostrando patrones de semejanzas entre las observaciones y las variables como puntos en gráficos de dispersión (Abdi & Williams, 2010).

Los principales objetivos al realizar un PCA son los siguientes (Wold, Esbensen, & Geladi, 1987):

- De un conjunto de datos, obtener la información más importante.
- Comprimir el tamaño de una tabla de datos, pero manteniendo la información más importante.

- Simplificar o reducir la información de una tabla de datos.
- Estudiar la estructura de las variables y las observaciones.

La Figura 2 muestra la estructura de un modelo PCA donde existen cuatro partes importantes: la matriz de datos (**X**), la matriz de *scores* (**T**), la matriz transpuesta de *loadings* (**P^t**) y la matriz de datos residuales (**E**), que tiene la misma estructura que la matriz **X**. Matemáticamente se puede expresar la descomposición de la matriz **X** como: $X = T \cdot P^T + E$.

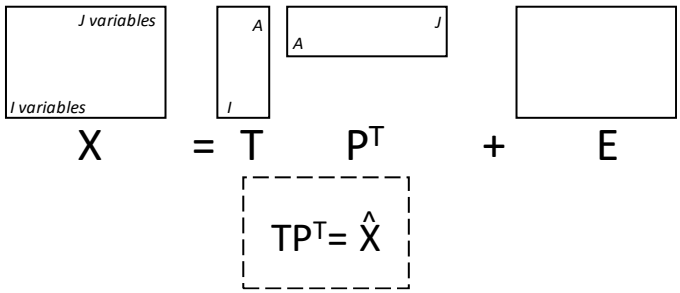

Figura 2. Estructura del PCA

La Figura 3 hace referencia a un conjunto de datos pre-procesados (centrados y escalados a varianza unitaria) con tres dimensiones, donde la recta de color naranja representa la recta que pasa por el origen y que recoge la máxima variabilidad de la proyección de los puntos sobre la recta con el mínimo error residual. A esta primera línea se le denomina primera componente; la distancia desde el origen hasta la proyección de cada punto es el *score* o variable latente (t_1) y la dirección de la recta es el *loading* (p_1). El número uno del subíndice, hace referencia a la variable latente asociada y se cumple que $\| \vec{p} \| = 1$. Una vez encontrada la primera recta, una segunda recta perpendicular a la dirección de la primera puede ser añadida, la cual recoge la máxima variabilidad restante, denominada segunda componente.

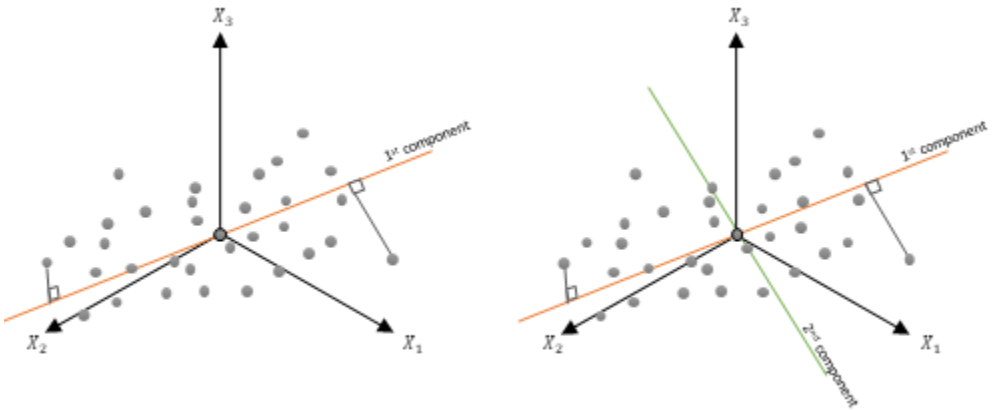

Figura 3. Interpretación geométrica del PCA

2.5 Regresión con Componentes Principales

La regresión con componentes principales (PCR) es una alternativa de la regresión lineal múltiple (MLR), donde la idea principal es reemplazar las variables primitivas por las componentes principales determinadas a partir del PCA. Al realizar PCR se comprimen las variables primitivas **X** aprovechando la correlación, obteniéndose *scores* (T) que son ortogonales, es decir, cada *score* explica una fuente de variabilidad distinta a las otras (Dunn, 2018).

La Figura 4 muestra que la regresión PCR es un proceso de dos pasos: el primero, donde se obtienen los *scores*, $T = X \cdot P$ a partir de un modelo PCA, y el segundo paso, donde se predice la variable respuesta a partir de las componentes principales (*scores*) calculadas en el primer paso, $\hat{y} = T \cdot b$, donde $b = (T^T T)^{-1} T^T y$.

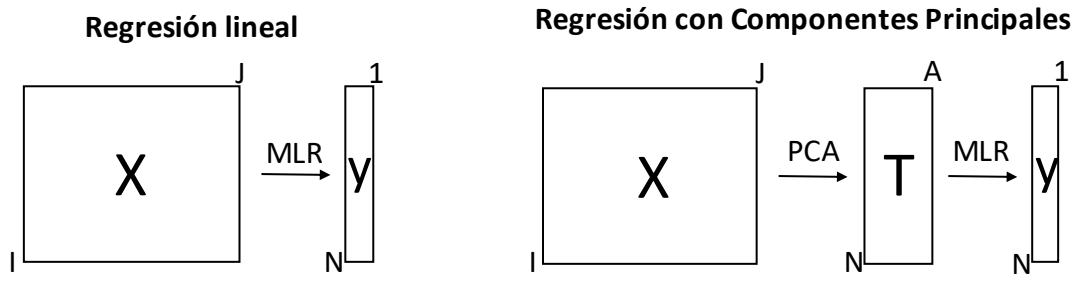

Figura 4. Estructura del PCR

Se tiene que tener presente que en el algoritmo del PCR, el cálculo de las variables latentes se realiza usando sólo la información de las variables explicativas, y no se utiliza información de la variable respuesta.

Por otro lado, el PCR presenta algunos inconvenientes. Por ejemplo, sí hay múltiples variables respuestas, se tendrá que realizar una regresión para cada variable respuesta. Adicionalmente, en el primer paso, al obtener las componentes principales, aquellas con autovalores pequeños pueden ser descartadas, a pesar de que pueden ser significativas para explicar la variable respuesta. Para evitar esto se tendrían que extraer muchas, o todas, las componentes principales, teniendo presente que este procedimiento hace difícil la interpretación de la variable respuesta a partir de las componentes.

2.6 Regresión en Mínimos Cuadrados Parciales

La regresión en mínimos cuadrados parciales (PLS) es un método que relaciona dos bloques de datos, matriz **X** y matriz **Y**. Deriva su capacidad para analizar datos con muchas variables con ruido, colinealidad e incluso incompletas tanto en el bloque **X** como en el **Y** (Wold, Sjöström, & Eriksson, 2011).

PLS es ampliamente utilizado para la predicción de varias variables respuestas. Desde el punto de vista predictivo, un modelo PLS es semejante a la regresión con componentes principales (PCR). Sin embargo, PLS introduce ventajas como la de no tener que crear tantos modelos PCR como número de variables respuestas. Además, PLS es más eficiente que el PCR ya que realiza el modelo en un solo paso. Semejante al PCA, en el PLS se extrae componentes, pero usando los bloques de datos **X** e **Y**. A menudo PLS requiere menos componentes que PCR para lograr el mismo nivel de varianza explicada, haciendo del modelo PLS un modelo compacto (Dunn, 2018).

En PLS hay dos objetivos simultáneos a cumplir:

1. Explicar las fuentes de variación en matriz **X** y la matriz **Y**.
2. Explicar la relación entre las fuentes de variación de la matriz **X** y la matriz **Y**.

Ambos objetivos se tratan de conseguir mediante la maximización de la covarianza entre las fuentes de variación de **X** e **Y**.

La Figura 5 hace referencia a un conjunto de datos **X** e **Y** pre-procesados (centrados en el origen y escalados a varianza unitaria) con tres dimensiones, donde la línea de color negra punteada representa la recta que pasa por el origen en ambos espacios **X** e **Y**, calculado para aproximar bien la nube de puntos en el espacio **X** e **Y**, con el objetivo de que a su vez, provean una buena correlación entre las proyecciones de cada observación ($\mathbf{t}_1$ y $\mathbf{u}_1$): a esta primera línea se denominará primera componente latente, donde la distancia desde el origen hasta la proyección de cada punto son los *scores* ($\mathbf{t}_1$ y $\mathbf{u}_1$) y los vectores unitarios de las direcciones de estas rectas ($\mathbf{w}_1$ y $\mathbf{c}_1$) son los *weightings*. El número uno del subíndice, hace referencia a la componente latente asociada. La segunda componente PLS es ortogonal a la primera sólo en el caso del espacio **X**, sin embargo, en el espacio Y puede o no poder serlo. Estas líneas tienen direcciones ($\mathbf{w}_2$ y $\mathbf{c}_2$) con proyecciones ($\mathbf{t}_2$ y $\mathbf{u}_2$).

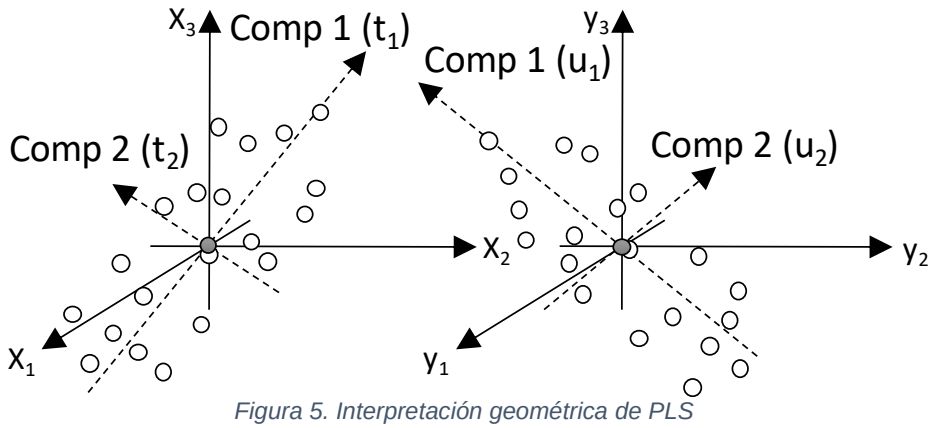

Figura 5. Interpretación geométrica de PLS

Las proyecciones (*scores*) de cada observación ($\mathbf{t}_1$ y $\mathbf{u}_1$) en ambos espacios **X** e **Y** están correlacionadas a través de la relación interna, donde se cumple que $u_{i1} = t_{i1} + h_i$, siendo h_i un residual. La Figura 6 muestra la relación entre las coordenadas de los *scores* ($\mathbf{t}_2$ y $\mathbf{u}_2$) usualmente es menor que en la primera componente.

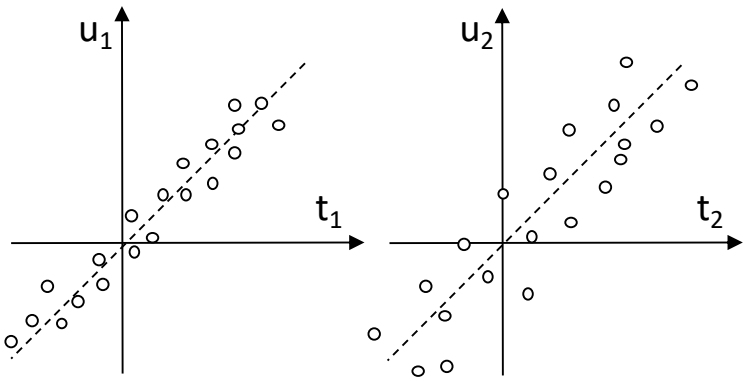

Figura 6. Relación interna del PLS

La Figura 7 muestra la estructura del PLS.

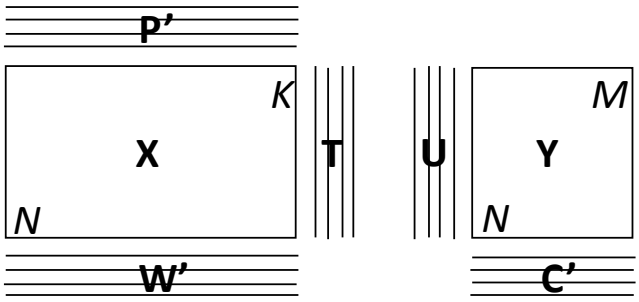

Figura 7. Estructura del PLS

Donde:

- K = número de variables **X**
- M = número de variables **Y**
- N = número de observaciones
- A = número de componentes PLS
- T = matriz de *scores*-**X** con columnas t_1, … t_A
- W* = matriz de *loadings*-**X** con columnas w^*_1, … w^*_A
- W = matriz de *weights*-**X** con columnas w_1, … w_A
- U = matriz de *scores*-**Y** con columnas u_1, … u_A
- C = matriz de *weights* -**Y** con columnas c_1, … c_A

Se cumple que $T = X \cdot W^*$ y la descomposición de la matriz **X** se logra mediante $X = T \cdot P^T + E$, donde **E** es la matriz residual del espacio **X**. Así mismo, la descomposición de la matriz **Y** se logra con $Y = U \cdot C^T + F$, donde **F** es la matriz residual del espacio **Y**. Además, $U = T \cdot m + H$, entonces $Y = T \cdot mC^T + G$.

Considerando un número de componentes a, uno de los objetivos del PLS es lograr maximizar $cov(t_a, u_a) = r(t_a, u_a) \cdot \sigma(t_a) \cdot \sigma(u_a)$

En el caso de PCA, la proyección de **X** es una aproximación óptima de **X**, mientras que, en el PLS, la proyección de **X** aproxima bien **X** e **Y**, y trata de maximizar la correlación entre **X** e **Y**.

PLS permite predecir simultáneamente un conjunto de variables dependientes a partir de un conjunto de variables predictores (Höskuldsson, 1988). Uno de los algoritmos más usados para ajustar el modelo PLS es el algoritmo NIPALS (*Nonlinear Iterative Partial Least Squares*).

La ventaja de los coeficientes de regresión PLS radica en la facilidad de interpretación del modelo, donde la influencia de cada variable viene indicada por la magnitud y signo de los coeficientes, que están centrados y escalados.

Las variables más importantes del modelo de regresión PLS con *a* componentes, pueden ser identificadas con el parámetro de importancia de la variable en la proyección, llamado VIP (*Variable Importance in the Projection*) introducido por Wold.

En el caso del análisis multivariante con muchas variables predictoras y muchas variables respuestas, el VIP es utilizado para seleccionar aquellas variables predictoras con mayor importancia respecto a las demás. Para realizar la poda del modelo, un criterio de selección de variables predictoras consiste en remover variables predictoras con valores de VIP menores a 1. Dado que el promedio de los VIP de las variables predictoras es igual a 1. Utilizando este punto de corte, se ajusta a un nuevo modelo con menor ruido y con mayor parsimonia. Este punto corte no está estadísticamente justificado.

Un método muy usado para seleccionar las variables estadísticamente significativas es mediante el cálculo de los intervalos de confianza (IC) *Jackknife* para los coeficientes de regresión PLS obtenidos, que proporcionan información acerca de la variabilidad en la estimación de los coeficientes de regresión (Ismartini, Sunaryo, & Setiawan, 2010). *Jackknife* es una técnica estadística de re-muestreo por validación cruzada muy versátil introducida por Quenouille

(1949). El proceso consiste en dejar una o un grupo de varias observaciones fuera, calcular los coeficientes del modelo con las observaciones restantes, y repetir este procedimiento tantas veces como observaciones (o grupos de observaciones) se tenga, y por cada componente creado. Al final del proceso se dispone de un conjunto de valores para cada uno de los coeficientes de regresión, a partir de los cuáles se calculan los intervalos de confianza *Jackknife*. La Figura 8 muestra un ejemplo del método cuando solo se elimina una observación cada vez (LOO, *Leave-one-out*).

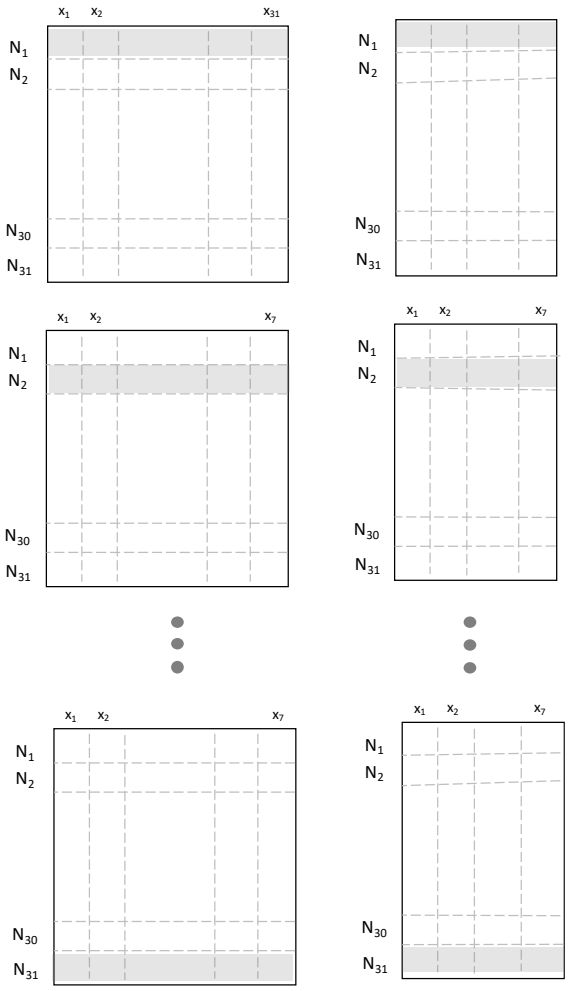

Figura 8. Método LOO Leave-one-out

La validación cruzada (CV) se utiliza también para ayudar a determinar un número adecuado de componentes en el modelo PLS. Uno de los métodos más usados es el LGO (*Leave-group-out*), que consiste en dividir las observaciones en grupos, excluir uno de los grupos, calcular un modelo con los restantes grupos

de observaciones, calcular el error de predicción del grupo excluido, obteniendo el PRESS (*Prediction Residual Error Sum of Squares*), $PRESS = \sum_i \sum_m (y - \hat{y})^2$, y repetir este procedimiento tantas veces como grupos se hayan formado, y por cada componente creado. La Figura 9 muestra el esquema de la validación cruzada.

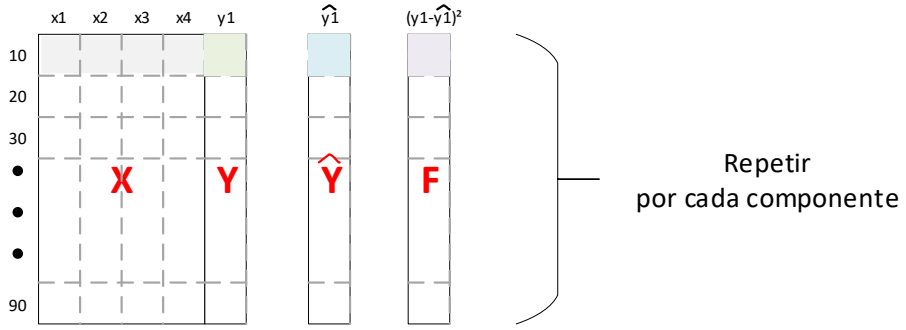

Figura 9. Estructura de la validación cruzada

La bondad de predicción Q^2, mide la potencia de predicción de nuevos datos del modelo PLS con *A* componentes, $Q^2 = 1 - (PRESS - SCT)$, y la bondad de ajuste R^2, evalúa la bondad de ajuste del modelo con *A* componentes a los datos existentes, $R^2 = 1 - (SCR/SCT)$. Se cumple de manera general $R^2 > Q^2$.

En el estudio de la detección de observaciones extremas se presentan dos estadísticos. El primero, la T² de Hotelling, definida como $T^2 = \sum_{a=1}^{a=A} \left(\frac{t_{i,a}}{S_a}\right)^2$, que geométricamente mide la distancia al cuadrado del centro del hiperplano a la proyección de una observación sobre dicho hiperplano. El segundo, SPE (*Squared Prediction Error*), $SPE_i = \sum_{i=1}^{i=N} {e_j}^2$, definida geométricamente como la distancia euclídea al cuadrado de una observación al hiperplano. A menudo se calculan percentiles al 95% para estos estadísticos.

2.7 Árbol de Decisión

Los árboles de decisión (*Decision Trees*) son algoritmos clásicos del aprendizaje automático introducidos por Morgan y Sonquist, (1963). Su atractivo radica en la simplicidad de la estructura resultante la cual es bastante fácil de ver, entender y, lo más importante, explicar. La Figura 10 muestra la estructura tradicional, que

empieza con un único nodo raíz que se divide en múltiples ramas, lo que lleva a nodos; cada uno de estos nodos se pueden dividir o terminar como un nodo de hoja. Asociado a cada nodo, una pregunta determina qué rama seguir. Los nodos de hoja o nodos terminales contienen las decisiones (Williams, 2011).

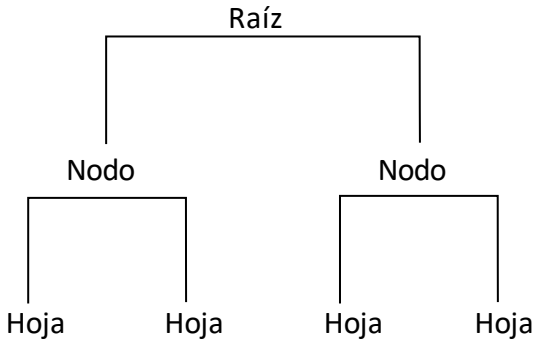

Figura 10. Estructura del árbol de decisión

El problema encontrado con los árboles de decisión es que se pueden volver excesivamente complejos. Además, al utilizar todos los datos en el análisis estadístico, suele producirse un sobreajuste de los datos. Por tanto, se considera que es complicado que un solo árbol de decisión pueda predecir bien datos a futuro.

2.8 Bosque Aleatorio

Los bosques aleatorios (*Random Forest*) son una combinación de predictores de árboles de modo que cada árbol depende de los valores de un vector aleatorio muestreado independientemente y con la misma distribución para todos los árboles en el bosque (Breiman, 2001).

La técnica del bosque aleatorio es similar al árbol de regresión, con la diferencia de que mejora la precisión mediante la incorporación de aleatoriedad en la construcción de cada clasificador individual.

Los bosques aleatorios tratan de resolver el problema del sobreajuste de datos que se pueda presentar en los arboles de decisión, creando múltiples arboles de decisión, los cuales tendrán una mayor resistencia a las variables que tienen pequeña influencia en la variable objetivo.

En la selección del conjunto de datos para el entrenamiento del bosque aleatorio, la aleatoriedad se consigue mediante el concepto de *bagging* (*booststrap aggregation* o agregación *booststrap*), donde una misma observación pueden aparecer múltiples veces en el conjunto de datos. De las observaciones que no se han tenido en cuenta para el entrenamiento (construcción) del modelo, se vuelve a aplicar *bagging* para la validación del mismo. Para cada conjunto de datos de entrenamiento, se determina la elección de variables predictoras para un mismo punto de división de todos los árboles creados. Con frecuencia se crean 500 árboles de decisión en el bosque aleatorio. En el modelo resultante final, los nodos y/u hojas decisiones se unen para formar una decisión final, dando como resultado final, la importancia de cada variable en la formación del bosque aleatorio. En el modelo resultante final, los nodos y/u hojas decisiones se unen para formar una decisión final, dando como resultado final, la importancia de cada variable en la formación del bosque aleatorio.

Los bosques aleatorios permiten meter mediciones de importancia en las variables que intervienen en el modelo. Sí se elimina una variable, mide en cuanto incrementa o disminuye el error cuadrático medio (MSE) de predicción.

2.9 Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial o SVM, son métodos usados habitualmente en problemas de clasificación, llamados así debido a sus siglas en inglés *Support Vector Machine*. La Figura 11 muestra que la idea principal es minimizar el riesgo estructural (SRM), lo que se realiza encontrando el hiperplano óptimo, llamado vector de soporte, que logre maximizar el margen de separación entre las clases, mapeando las observaciones en un espacio mayor al inicial (Vapnik, 1995).

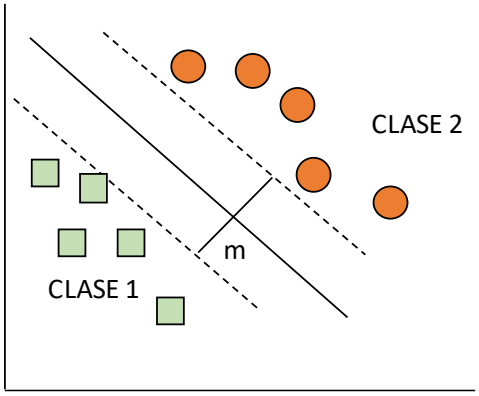

Figura 11. Estructura de SVM

En el caso en que las observaciones puedan ser linealmente separables, para el caso de dos clases C_1 y C_2, cada observación puede ser expresada: $(y_1, x_1), (y_2, x_2) \ldots (y_i, x_i)$ siendo *i* es el número de observaciones, donde $x_i \in R^N$, y las clases son etiquetadas como $y_i \in \{-1, 1\}$. Para encontrar el hiperplano óptimo, los puntos son mapeados en un espacio de una dimensión mayor. Sea $\boldsymbol{z} = \varphi(x)$ el vector en el espacio con un mapeo $\varphi(x)$ de R^N a un espacio de características $\boldsymbol{z}$. Se trata de encontrar el hiperplano denotado por: $w \cdot \boldsymbol{z} + b = 0$.

El hiperplano está definido por el par (w, b), y en la siguiente función es posible separar el punto (x).

$$f(x_i) = sign(w \cdot z_i + b) = \begin{cases} 1, & y_i = 1 \\ -1, & y_i = -1 \end{cases}$$

donde $w \in Z$ y $b \in R$. Entonces se puede resumir de la siguiente manera, el conjunto C es separable linealmente si existe el par (w, b) tal que se cumpla:

$$\begin{cases} (w \cdot z_i + b) \geq +1, & y_i = +1 \\ (w \cdot z_i + b) \leq -1, & y_i = -1 \end{cases}$$

La Figura 12 muestra el hiperplano determinado por el vector $\boldsymbol{w}$, no es único, existen infinitos. Sin embargo, es posible encontrar un hiperplano óptimo, en el cual, el margen entre las proyecciones de los puntos de dos diferentes clases es maximizado.

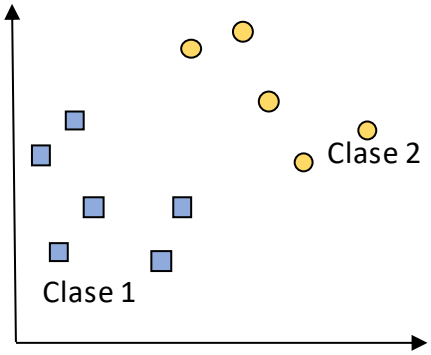

Figura 12. Separación de clases en SVM

Como pasa en los problemas reales, no es tan fácil que un hiperplano logre separar de manera lineal perfecta los datos. La Figura 13 muestra el caso en que las observaciones no puedan ser linealmente separables, introduciendo una variable denominada holgura, denotada por ξ, el procedimiento para este caso es similar al caso en que las observaciones son linealmente separables considerando que es posible generalizar la ecuación introduciendo esta variable de carácter no negativo, $\xi \geq 0$, de modo que la ecuación anterior pueda ser modificada a:

$$y_i(w \cdot z_i + b) \geq 1 - \xi_i$$

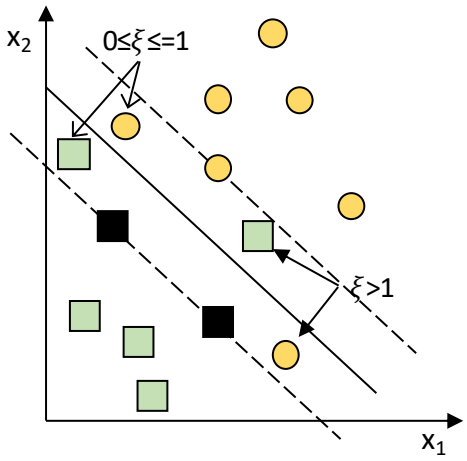

Figura 13. Datos no separables linealmente en SVM

En la función objetivo se añade un valor C que es un parámetro de regularización o penalización.

Ahora para encontrar el hiperplano óptimo se define como la solución a:

$$min_{w,b}\left\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{I}\xi_i\right\}$$

$$\text{Sujeto a: } \begin{cases} y_i(w \cdot z_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

La búsqueda del hiperplano óptimo es un problema de programación cuadrática, el cual puede ser resuelto mediante multiplicadores de Lagrange.

Las máquinas de soporte vectorial no sólo pueden ser aplicadas para clasificación, también para regresión. Las Máquinas de Soporte Vectorial para Regresión o SVMR, son llamadas así debido a sus siglas en inglés *Support Vector Machine for Regression*.

Para aplicar SVMR, una función de pérdida alternativa, llamada ε-*insensitive*, es introducida para incluir una medida de distancia. El objetivo es encontrar una función $f(x)$, que maximice la desviación de ε con respecto a los objetivos y_i, para todos los datos de entrenamiento, intentando que la función sea lo más plana posible, es decir, que busque el menor parámetro posible para w.

La Figura 14 muestra la variable ε, que es conocida como la anchura de la banda. Así mismo, la variable de la holgura ξ será introducida para permitir ciertos errores en el término de regularización C.

La función de pérdida más general, es descrita como:

$$|y - f(x)|_\varepsilon = max(0, |y_i - f(x)| - \varepsilon)$$

Se describe este problema como un caso de optimización convexa:

$$min_{w,b}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{I}(\xi_i + \widehat{\xi_\iota})\right)$$

$$sujeto\ a \begin{cases} (x_i \cdot w + b) - y_i \leq \varepsilon + \xi_i \ \ \forall i \\ y_i - (x_i \cdot w + b) \leq \varepsilon + \xi_i \ \ \forall i \\ \xi_i, \widehat{\xi_\iota} \geq 0 \ \ \forall i \end{cases}$$

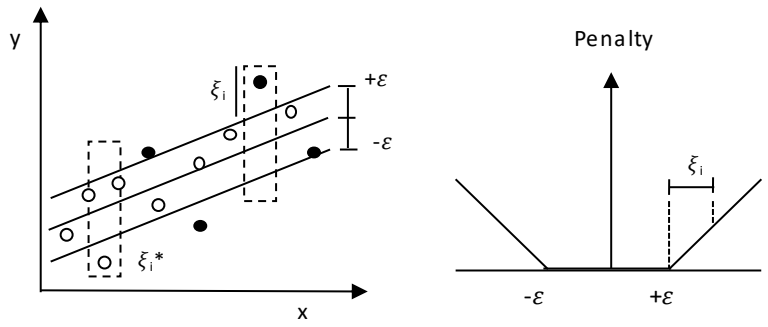

Figura 14. Influencia de ε y ξ en SVM

Sí la regresión lineal no se ajusta a los datos, un kernel de mapeo no lineal denotado K, se utiliza para mapear los datos a un espacio de características de dimensión mayor en donde se pueda predecir la regresión lineal.

Algunos de los Kernel más utilizados son:

- Polinomio: $k(x, y) = (\langle x, y \rangle + 1)^d, d = 1, 2, \ldots$
- Funciones de base radial: $k(x, y) = exp\{-|x - y|^2/\sigma^2\}$
- Redes neuronales de dos capas: $k(x, y) = tanh(b\langle x, y \rangle - c)$

2.10 Lenguaje de programación

Se considera que la programación de cualquier técnica estadística apoya al mejor entendimiento de la misma. En este caso, el lenguaje de programación a utilizar para aplicar las técnicas estadísticas es R[1]. R es un software libre (Matloff, 2011). La comunidad de personas que utilizan R es muy colaborativa, y las dudas en la programación pueden ser resueltas consultando en internet. Además, R contiene una gran cantidad de librerías estadísticas creadas por expertos; al ser de código abierto, las funciones y cálculos pueden ser examinadas.

2.10.1 Librería de Visualización

RGL es un sistema de representación 3D en tiempo real. Proporciona funciones de nivel medio a alto para gráficos interactivos en 3D (Adler , Murdoch , &

[1] "The R Project for Statistical Computing" https://cran.r-project.org/

Nenadic, 2018). Las muestras provienen de un depósito metalúrgico, su visualización en 3D es posible utilizando la librería en R llamada RGL.

2.10.2 Librería PCA

FactoMiner es un paquete de R dedicado al análisis multivariante de datos. La principal característica de este paquete es la posibilidad de tener en cuenta diferentes tipos de variables (cuantitativas o categóricas). A grandes rasgos, los métodos implementados en el paquete son conceptualmente similares a su objetivo principal, es decir, resumir y simplificar los datos reduciendo la dimensionalidad del conjunto de datos, dependiendo del tipo de datos disponibles. En el proyecto metalúrgico se realizaron pruebas de flotación de minerales, obteniéndose más de una variable respuesta. Este paquete fue utilizado para agrupar las variables respuestas de manera multivariante.

2.10.3 Librería PLS

En la sección 2.6 se describe que un método para determinar la regresión PLS es el algoritmo NIPALS, iterando ambos bloques de datos X e Y. La Figura 15 muestra un esquema de este algoritmo.

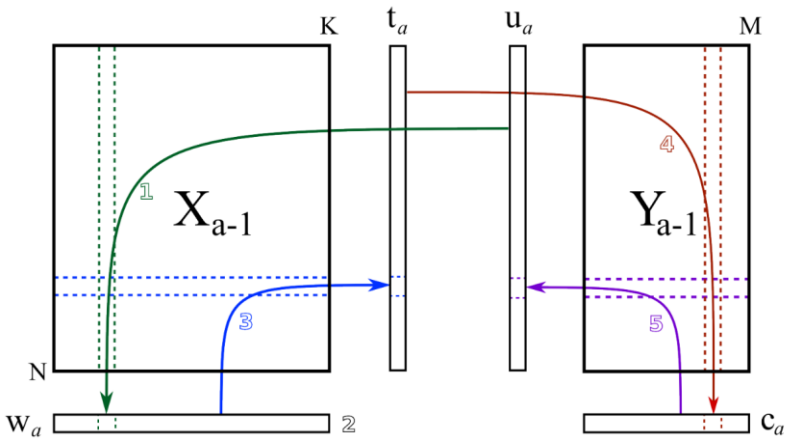

Figura 15. Algoritmo NIPALS para PLS

En esta tesis se ha programado en R el algoritmo básico propuesto por Wold para desarrollar el algoritmo NIPALS, efectuando los siguientes pasos:

1. Escalar y centrar matriz $\boldsymbol{X_o}$
2. Establecer u como la primera columna de Y
3. Hallar $w = X^t u/(u^t u)$

4. Convertir w a un vector unitario
5. Hallar $t = Xw$
6. Hallar $c = Y^t t/(t^t t)$
7. Convertir c a un vector unitario
8. Hallar $u = Yc/(c^t c)$
9. Repetir desde (3) hasta que converger
10. Calcular X-*loadings*: $p = X^t t/(t^t t)$
11. Calcular Y-*loadings*: $q = Y^t u/(u^t u)$
12. Hallar $b = u^t t/(t^t t)$
13. Deflactar $X_{res} = X - tp^t$, $Y_{res} = Y - btc^t$

Así mismo, los coeficientes de la regresión PLS por cada componente y los estadísticos la T^2 y SPEX fueron programados en R

Adicionalmente, dentro del mismo paquete R se encuentra la librería plsdepot que contiene diferentes métodos para el análisis PLS. La función plsreg1 realiza la regresión PLS para el caso de una variable respuesta. La función plsreg2 realiza la regresión PLS para el caso de más de una variable respuesta (Sanchez, 2012).

Estas funciones del paquete psdepot fueron utilizadas para realizar la regresión PLS y complementar los resultados con la programación realizada.

2.10.4 Librería PCR

La librería PLS implementa la regresión con componentes principales (PCR) y la regresión por mínimos cuadrados parciales (PLS), utilizando como función principal PCR para ajustar los modelos (Mevik, Wehrens, Liland, & Hiemstra, 2016).

2.10.5 Librería Árbol de Regresión

La librería rpart (*Recursive Partitioning and Regression Trees*), o partición recursiva para árboles de clasificación, regresión y supervivencia, ajusta modelos con árboles (Therneau, Atkinson, & Ripley, 2018).

2.10.6 Librería Bosque Aleatorio

Para generar el bosque aleatorio hay muchos métodos, los dos más conocidos son *boosting* (Shapire and Freud, 1998) y *bagging* (Breiman, 1996) para generar

árboles de regresión. La librería *randomForest* aparte de realizar el modelado con bosques aleatorios, opcionalmente proporciona dos informaciones adicionales útiles en la selección de variables importantes: una medida de importancia de variables predictoras y una medida de la estructura interna de los datos (Liaw & Wiener, 2002).

2.10.7 Librería Máquina de Soporte Vectorial

Kernlab es una librería extensible del método de aprendizaje automático basado en kernel en R. El algoritmo de aprendizaje basado en kernel más prominente es, sin lugar a duda, la máquina de soporte vectorial. La implementación de kernlab de la máquina de soporte vectorial está basada en los optimizadores encontrados en bsvm (Hsu and Lin, 2002b) y libsvm (Chang and Lin, 2001), los cuales incluyen una versión muy eficiente de la Optimización Secuencial de Minimización (SMO) (Karatzoglou, Smola, Hornik, & Zeileis, 2004).

El paquete e1071 contiene funciones para determinar la máquina de soporte vectorial. SVM se usa para entrenar una máquina de soporte vectorial. Puede ser usado para llevar a cabo un regresión general y también una clasificación (Meyer, y otros, 2017).

Capítulo 3

3 Proyecto Metalúrgico

El presente capítulo muestra la descripción del proyecto metalúrgico, el yacimiento mineral, las perforaciones geológicas, el modelo geológico y las pruebas metalúrgicas.

3.1 Yacimiento Mineral

Una definición de yacimiento mineral o depósito mineral sería: “parte de la corteza terrestre, en la cual, debido a procesos geológicos, ha habido una acumulación de materia prima mineral, la cual por sus características de cantidad, calidad y condiciones de depósito es rentable su explotación” (Servicio Geológico Mexicano, 2017). La acumulación de materia prima mineral, proveniente de la combinación del magma y los fluidos hidrotermales, es llamada veta mineral, la cual es almacenada en la roca caja y la roca volcánica. La Figura 16 muestra un esquema de este proceso. El yacimiento del presente proyecto pertenece a un pórfido cuprífero, es decir, un yacimiento geológico de gran extensión (400 millones de toneladas) con baja concentración de mineral de cobre [0.2% Cu - 1.5% Cu], que contiene asociaciones de oro en pequeñas cantidades [0.01 g/t Au - 0.3 g/t Au].

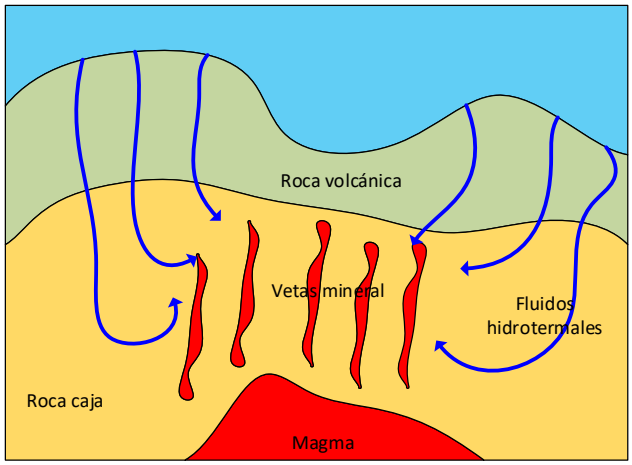

Figura 16. Yacimiento mineral

3.2 Perforaciones Geológicas

Las perforaciones otorgan información directa sobre la estructura del material debajo de la superficie. Los núcleos de perforación muestran la litología o tipo de roca del sitio de la perforación. El *logueo* o registro de la perforación proporciona información *in situ*, además de la información geológica directa (Wonik & Olea, 2007). La Figura 17 muestra un esquema de las perforaciones geológicas sobre un yacimiento geológico.

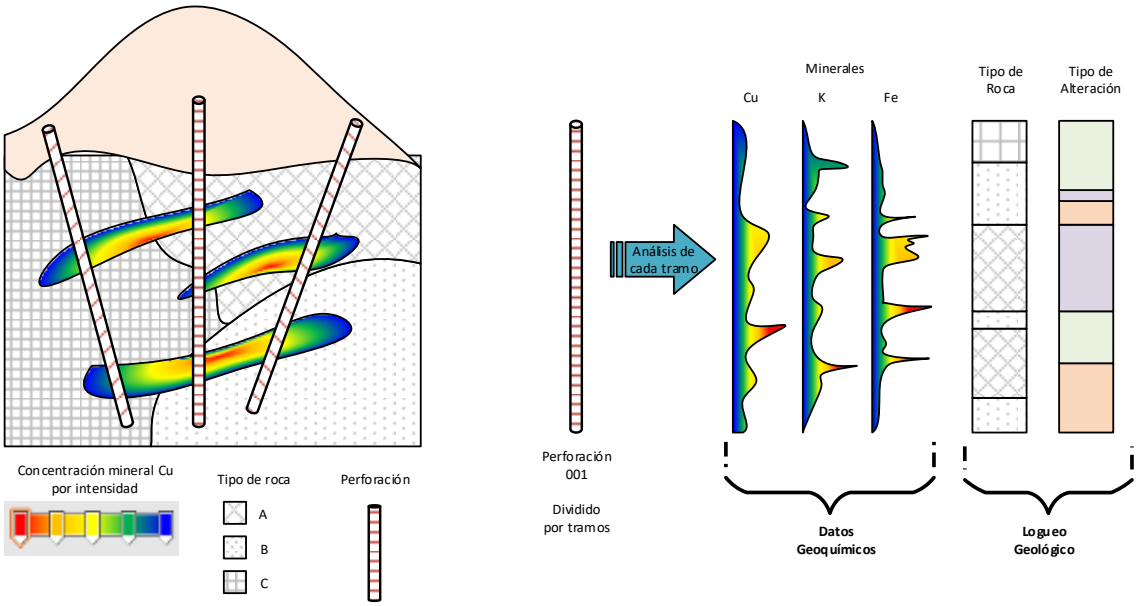

Figura 17. Perforaciones geológicas

El registro geoquímico de las perforaciones consiste en medir propiedades químicas de muestras que se obtienen dividendo la perforación en tramos de dos

metros a lo largo del taladro. La Figura 18 muestra los tramos de las perforaciones obtenidos una vez extraídas.

Figura 18. Tramos de la perforación

Más de 200 perforaciones geológicas diamantinas del tipo de núcleo HQ, con 63.5 mm de diámetro de núcleo, fueron realizadas sobre el yacimiento mineral del proyecto. Sin embargo, debido a factores económicos, sólo sesenta perforaciones geológicas fueron seleccionadas para realizar pruebas metalúrgicas.

3.3 Modelo Geológico

La formación de un yacimiento mineral es gobernada por procesos complejos. La estructura del yacimiento mineral es, en parte, determinista y, en parte, estocástica. El modelo de bloques se utiliza comúnmente para proporcionar una representación espacial de las variables geológicas, y proporcionar un formato útil para almacenar otros atributos importantes que incluyen la estimación de la concentración del mineral (Rossi & Deutsch, 2014). La Figura 19 muestra un boceto del modelo geológico representado por un modelo de bloques. El proyecto consta de cerca de 29000 bloques de dimensiones de 20 m x 20 m x 15 m con una densidad promedio de cada bloque de 2.4 m^3/t, con información adicional de la concentración en el bloque de cobre, oro, plata, hierro, calcio y estroncio. Sobre el modelo de bloques y a partir de la base de datos de las perforaciones, a través de técnicas de interpolación espacial como *kriging*, por ejemplo, se obtienen variables como la ley de cobre en el yacimiento geológico.

En el ámbito geológico, debido al alto coste y tiempo que supone, sólo se interpolan las variables más importantes. Por lo general, sólo las concentraciones de los minerales más importantes en el yacimiento geológico son incluidas al modelo de bloques.

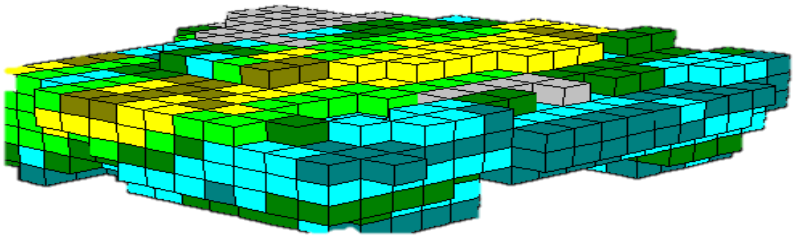

Figura 19. Modelo de bloques

3.4 Muestras Metalúrgicas

La heterogeneidad *in situ* de la mineralización da como resultado variabilidad en la concentración y parámetros del proceso. El enfoque geo-metalúrgico requiere muestras distribuidas espacialmente dentro de un yacimiento mineral para soportar los modelos de variabilidad. Un protocolo de muestreo bien diseñado y planificado puede respaldar la predicción de recursos, reduciendo la variabilidad y proporcionando resultados adecuados (Dominy, O'Connor, & Xie, 2016). Noventa muestras metalúrgicas con una longitud aproximada de ocho metros fueron seleccionadas de las perforaciones para metalurgia. La Figura 20 muestra una representación espacial de las muestras que fueron extraídas para el proyecto cumpliendo el protocolo de representatividad de muestras dentro de un yacimiento.

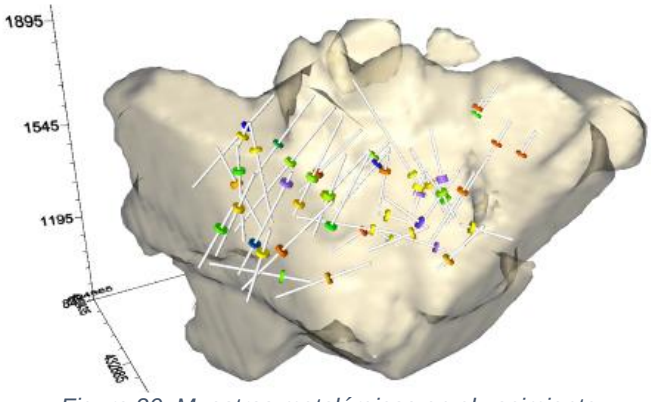

Figura 20. Muestras metalúrgicas en el yacimiento

3.5 Flotación de Minerales

El proceso de flotación es una interacción compleja entre material valioso, material no valioso, reactivos de flotación, burbujas de aire y el agua, donde modificando las propiedades fisicoquímicas superficiales de las partículas para convertirlas en hidrofóbicas e hidrofílicas se puede concentrar el material valioso (Bulatovic, 2007). La Figura 21 muestra un esquema del proceso de flotación, donde en la parte superior se recoge el concentrado, y en la parte inferior la parte no valiosa, también llamada relave o ganga, es mantenida dentro de la celda de flotación.

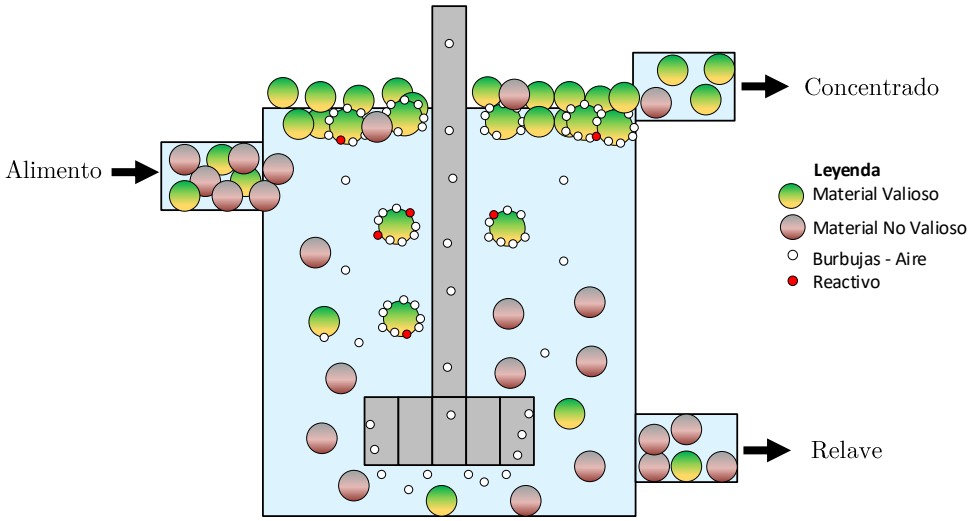

Figura 21. Flotación de minerales

El rendimiento del proceso de flotación varía significativamente dependiendo de las características físicas intrínsecas del material como la intensidad de oxidación en la superficie, grado de liberación, concentración de mineral, la mineralogía que acompaña a la parte valiosa, las cuales son costosas y difíciles de cuantificar. Estas características del mineral tienen una crucial importancia ya que al mezclarse en la solución, los parámetros de la pulpa como el índice de acidez, pH, o el índice de potencial de reducción-oxidación, Eh, afectan al rendimiento de la flotación, provocando que el proceso de flotación sea complejo de determinar por modelos matemáticos convencionales.

3.6 Cinética de Flotación de Minerales

La cinética de flotación puede ser definida como una medida de la eficiencia de flotación expresada a un tiempo determinado. La velocidad de flotación contabiliza la cantidad de partículas flotadas con respecto al tiempo y es la fuente más fiable para describir la cinética de flotación (Bulatovic, 2007).

La construcción de un buen modelo cinético depende de la calidad de ajuste de los datos al modelo; no existe un modelo cinético universal para el proceso de flotación.

En la actualidad existen muchos modelos cinéticos para estimar la recuperación en función del tiempo. Sin embargo, el modelo para los elementos cobre y oro que mejor se ajusta a este proyecto es el modelo cinético de primer orden de Klimpel, que es expresado de la siguiente manera:

$$R_t = R_{max}\left[1 - \frac{1}{k \cdot t}(1 - \exp(-k \cdot t)\right]$$

Donde:

R_t = Recuperación de un elemento al tiempo t

R_{max} = Recuperación máxima

k = Constante de velocidad de flotación

t = Tiempo de flotación

La diferencia entre el modelo clásico y el de Klimpel radica en que la representación de la constante de velocidad (k) se muestra como una distribución rectangular, en la cual la cantidad de partículas en flotación es constante durante un intervalo de tiempo limitado. En cambio en el modelo clásico se considera que la cantidad de partículas en flotación es constante en todo el tiempo (Gupta & Yan, 2016).

3.7 Pruebas Metalúrgicas de Flotación

Todas las pruebas metalúrgicas y análisis químicos fueron llevados a cabo en un laboratorio de procesamiento de minerales en Perú. Un total de noventa muestras distribuidas dentro del yacimiento fueron utilizadas en este proyecto. Un amplio rango de análisis ICP (*Inductively Coupled Plasma*) multielementos fueron realizados al total de las muestras. Sobre cada una de ellas se realizó una reducción de tamaño a 2.0 mm y se almacenaron en recipientes con nitrógeno en el interior para evitar su oxidación. Posteriormente, una porción de 1 kg fue extraída y molida en un molino de laboratorio hasta obtener un tamaño de partícula con P80 de 150 µm a un 33% de sólidos para realizar la prueba de flotación. La prueba de flotación fue llevada a cabo en una celda de flotación de laboratorio durante un tiempo total de 16 minutos, recogiendo una muestra de concentrados en distintos intervalos de tiempo a los minutos 1, 2, 4, 8 y 16, utilizando 60 g/t del reactivo de óxido de calcio (CaO) para modificar el pH de la pulpa, 30 g/t del reactivo AP-9950, que es un tionocarbamato utilizado para la colección de partículas valiosas, y 5 g/t de espumante utilizado para promover la formación de burbujas. La Tabla 1 condiciones de las pruebas metalúrgicas de flotación. Al finalizar la prueba metalúrgica de flotación, sobre los dos productos de la prueba (material valioso y el material no valioso), se determina la concentración mineral de los elementos de cobre y oro.

Tabla 1. Condiciones de la prueba de flotación

Etapa	Reactivos, g/t			Tiempo	pH
	CaO	AP-9950	Espumante	min	
Molienda	40	20	2.5	10	9
Acondicionamiento	20	10	2.5	3	9
Flotación	-	-	-	16	9

La Tabla 2 muestra un ejemplo del balance metalúrgico realizado para una de las 90 muestras, en el cual se indica que para un tiempo de 16 minutos de flotación se ha obtenido un concentrado de 12.3% de masa total, con una ley de 5.4% Cu y 0.24 g/t Au, con una recuperación de 93.8% Cu y 62% Au. Así mismo,

un 87.7% de masa total equivale a la parte no valiosa, llamada relave, donde se aprecia una mínima concentración de cobre y oro, 0.05% y 0.02 g/t, respectivamente.

Tabla 2. Balance metalúrgico

Producto	Tiempo	Peso	Análisis		Recuperación	
ID	min.	%	Au, g/t	Cu, %	Au, %	Cu, %
Conc.1	1	3.6	0.45	10.7	33.1	54.2
Conc.2	2	5.3	0.37	10.2	40.8	76.4
Conc.3	4	6.9	0.35	9.0	49.2	87.7
Conc.4	8	8.9	0.31	7.3	56.2	91.4
Conc.5	16	12.3	0.24	5.4	62.0	93.8
Relave	-	87.7	0.02	0.05	38.0	6.2
Cabeza calc.	-	100.0	0.05	0.7	100.0	100.0

La Figura 22 muestra las curvas de cinética de flotación de cobre y oro obtenida para la observación *5*. Asimismo, las líneas punteadas de color azul muestran la estimación del modelo cinético. Para el cobre, por ejemplo, se obtuvo una recuperación máxima (Rmax) de 88.0% Cu y una constante de velocidad cinética (k) de 2.4.

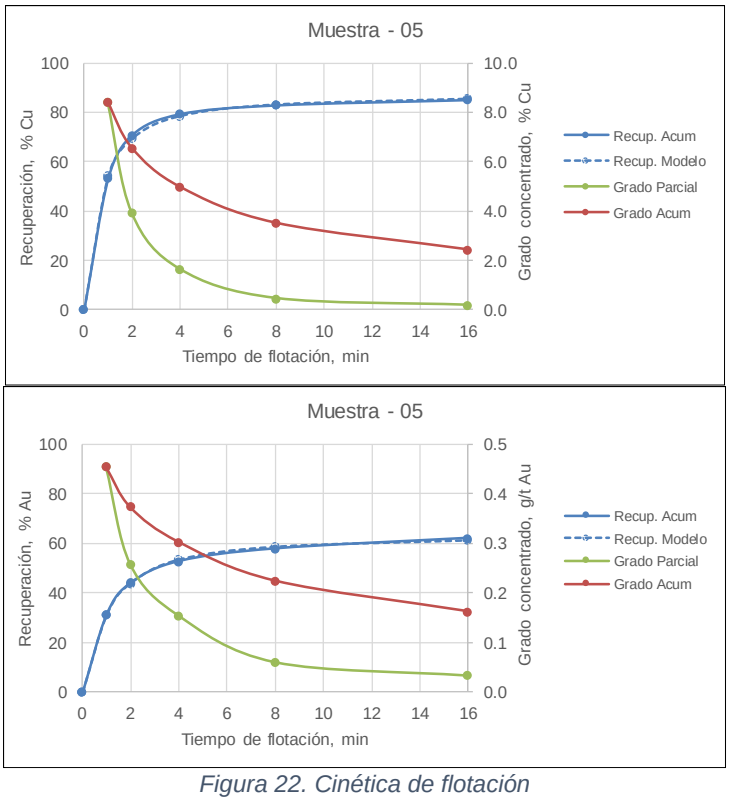

Figura 22. Cinética de flotación

3.8 Estimación del Coste del Proyecto

La estimación del coste del proyecto, que tiene por objetivo principal la predicción de respuestas metalúrgicas, se realiza estimando el coste de tres ítems. El primero es el coste de extracción de rocas cilíndricas del interior del yacimiento a través de perforaciones diamantinas, a un precio de extracción de 100€ por cada metro. En promedio cada perforación tiene 200 metros, con lo que se obtendría un coste de 20.000€ por cada perforación a realizar. Considerando que se han realizado 60 perforaciones, este primer ítem es el más costoso (1.200.000€). El segundo ítem es el coste de los análisis químicos ICP realizados sobre cada uno de los tramos de las perforaciones. Aproximadamente se toma una muestra representativa de intervalos de cada 2 metros por cada perforación, obteniendo 100 intervalos por cada perforación. Considerando 60 perforaciones, son 6000 análisis químicos a realizar que, a un precio unitario de 50€ por cada análisis ICP en un laboratorio químico, generan un coste de 300.000€. El último ítem es el coste de las pruebas metalúrgicas de flotación realizadas en un laboratorio sobre las noventa muestras, a un precio unitario de 500€ por cada prueba da un coste de 45.000€.

La Tabla 3 resume la estimación de coste del proyecto, con un coste total de más de un millón y medio de euros, donde el mayor coste está relacionado con las perforaciones.

Tabla 3. Estimación de coste

Ítem	Descripción	Cantidad	Coste unitario, €	Coste total, €
1	Perforaciones	60	20000	1200000
2	Análisis químicos de intervalos	6000	50	300000
3	Pruebas flotación a muestras	90	500	45000
	Total			1545000

Capítulo 4

4 Metodología Propuesta

Este capítulo describe la metodología propuesta aplicada en la construcción de modelos de predicción para las respuestas metalúrgicas del proyecto.

4.1 Datos de las Pruebas Metalúrgicas

Como se ha descrito en el Capítulo 3, para investigar la relación de las respuestas metalúrgicas con los análisis químicos sobre noventa muestras extraídas de un yacimiento geológico, se realizaron análisis químicos y luego fueron llevadas a un laboratorio metalúrgico donde se realizaron pruebas metalúrgicas de flotación, obteniéndose un bloque de variables predictoras **X** y un bloque de variables respuestas **Y**.

- Bloque **X** con ubicación espacial XYZ y datos geoquímicos ICP multielementos, 37 elementos:
    - Au (g/t), Ag (g/t), Cu (%), CuSS(%), CuCN(%), CuRes(%), Fe (%), S (%), C(%), arcillas (%), Al (%), Ca (%), K(%), Mg (%), Mn (%), Na (%), Pb (%), Ti (%), Zn (%), As (g/t), Ba (g/t), Be (g/t), Bi (g/t), Cd (g/t), Co (g/t), Cr (g/t), Mo (g/t), Ni (g/t), P (g/t), Sb (g/t), Sc (g/t), Sn (g/t), Sr (g/t), V (g/t), W (g/t), Y (g/t), Zr (g/t).
- Bloque **Y** con resultados de flotación:
    - Masa de concentrado recuperada (MassRo)
    - Recuperación máxima de Cu y Au (CuRmax & AuRmax)
    - Concentración Cu y Au (CuGra & AuGra) en el concentrado
    - Constante de velocidad cinética de Cu y Au (Cukinetic & Aukinetic)

La Tabla 4 muestra un fragmento del bloque **X**, mientras que la Tabla 5 muestra un fragmento del bloque **Y**, ambos para las 16 primeras muestras.

Tabla 4. Datos del bloque de variables X

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Sample	x	y	z	RQD	Ag	Au	Cu	CuSS	CuCN	CuRes	Fe	St	Ct	Arc	Al	Ca
2	ID-01	2805	4033	493	0.82	0.60	0.05	0.74	0.15	0.53	0.03	4.11	2.08	0.08	1.20	1.44	0.20
3	ID-02	2805	4042	467	0.05	0.60	0.06	0.25	0.02	0.04	0.18	3.82	2.67	0.04	1.15	1.22	0.20
4	ID-03	2805	4051	442	0.45	0.60	0.03	0.18	0.01	0.02	0.14	6.05	4.03	0.04	1.29	1.99	0.21
5	ID-04	2805	3962	462	0.18	0.60	0.03	0.49	0.13	0.33	0.02	5.18	4.66	0.03	1.20	1.82	0.04
6	ID-05	2805	3950	415	0.08	0.60	0.03	0.31	0.10	0.18	0.03	7.05	4.69	0.05	0.95	1.75	0.12
7	ID-06	2805	3854	434	0.34	0.60	0.01	0.23	0.10	0.11	0.01	6.30	4.75	0.03	1.20	1.99	0.05
8	ID-07	3109	3714	512	0.23	0.60	0.02	0.17	0.06	0.10	0.01	5.25	2.81	0.04	1.30	2.59	0.11
9	ID-08	3414	3980	574	0.00	0.60	0.03	0.48	0.13	0.31	0.03	4.66	2.69	0.06	1.15	2.18	0.15
10	ID-09	3414	3962	534	0.14	0.60	0.17	1.42	0.16	1.16	0.09	3.63	1.84	0.03	1.45	2.03	0.13
11	ID-10	3514	3726	531	0.24	0.60	0.05	0.29	0.08	0.10	0.08	4.29	0.64	0.05	1.39	2.08	0.10
12	ID-11	3514	3719	519	0.48	0.60	0.04	0.76	0.14	0.57	0.06	4.55	2.45	0.02	1.29	2.34	0.04
13	ID-12	3811	3950	680	0.23	1.20	0.20	0.29	0.11	0.01	0.13	4.54	0.15	0.06	1.05	1.63	0.21
14	ID-13	3710	4115	663	0.19	0.60	0.03	0.21	0.04	0.12	0.03	3.83	1.97	0.04	0.30	2.45	0.05
15	ID-14	3710	4111	633	0.13	3.00	0.31	1.19	0.17	0.78	0.21	7.90	7.41	0.04	0.34	1.66	0.06
16	ID-15	4430	4048	694	0.11	0.60	0.06	0.42	0.05	0.31	0.02	2.01	1.51	0.04	0.50	1.33	0.03
17	ID-16	4425	4047	672	0.00	0.60	0.06	0.82	0.10	0.68	0.03	3.06	2.51	0.02	0.30	1.65	0.04

Tabla 5. Datos del bloque de variables Y

	A	AG	AH	AI	AL	AM	AP	AQ
1	Sample	MassRo	AuGra	CuGra	AuRmax	CuRmax	Aukinetic	Cukinetic
2	ID-01	12.33	0.24	5.39	61.83	92.02	1.54	0.88
3	ID-02	10.33	0.49	2.14	78.13	91.71	1.00	0.62
4	ID-03	14.38	0.18	1.11	78.72	91.04	1.31	0.84
5	ID-04	31.65	0.13	1.36	89.33	88.52	0.44	0.48
6	ID-05	10.82	0.16	2.44	63.87	87.96	1.54	2.37
7	ID-06	24.42	0.03	0.74	62.19	81.43	1.09	1.34
8	ID-07	14.49	0.09	1.10	72.20	89.63	1.98	2.21
9	ID-08	13.39	0.27	3.17	75.59	84.79	3.15	0.99
10	ID-09	13.07	0.52	10.42	62.16	97.79	0.78	2.46
11	ID-10	12.77	0.15	1.20	43.37	51.62	0.67	2.65
12	ID-11	16.90	0.21	3.95	84.56	90.50	2.65	2.64
13	ID-12	16.01	0.59	0.40	55.61	25.08	0.46	0.33
14	ID-13	19.21	0.11	0.82	64.90	80.34	0.67	1.47
15	ID-14	14.66	1.17	6.96	66.98	90.60	1.43	1.99
16	ID-15	14.49	0.33	2.72	65.13	91.74	1.48	1.28
17	ID-16	14.49	0.33	5.36	65.13	95.63	1.48	1.61

La Figura 23 representa la vista espacial (ubicación en el yacimiento) de las muestras en estudio.

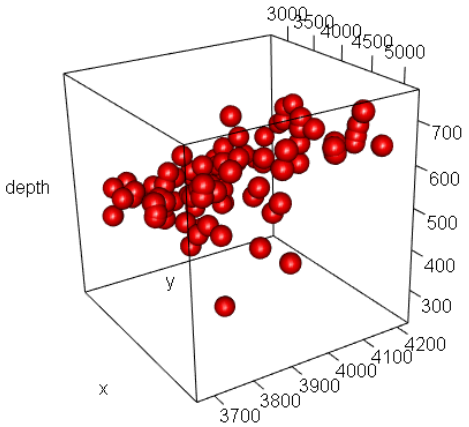

Figura 23. Muestras en el espacio

La Figura 24 representa la vista espacial de las muestras en estudio, mostrando la variabilidad del concentrado de cobre con una escala de colores.

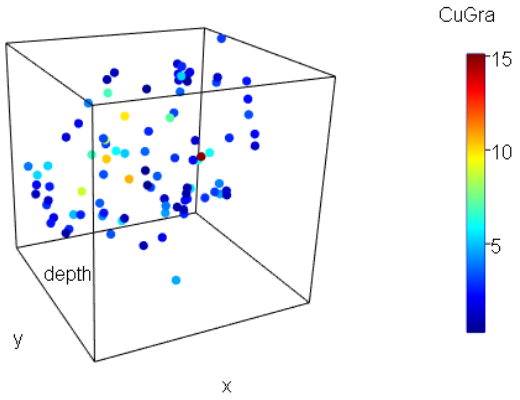

Figura 24.Muestras por concentración de cobre

En la tabla de datos de la matriz **X**, el análisis químico de cobre secuencial determina la cantidad de cobre que es soluble en ácido súlfurico y en cianuro, variables CuSS y CuCN respectivamente. En la industria del procesamiento mineral del cobre este análisis es muy importante, pues es un indicador de la cantidad de sulfuros de cobre que puede llegar a contener el mineral. La extracción de minerales de cobre por flotación es más eficiente con minerales que tienen sulfuros de cobre. La Figura 25 muestra la colinealidad presente entre las variables de la matriz **X**: Cu, CuSS y CuCN.

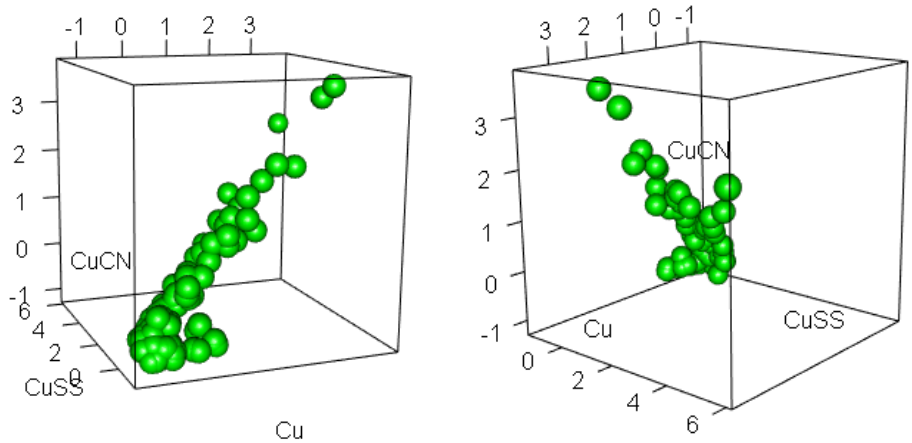

Figura 25. Colinealidad entre variables de la matriz X

Por otro lado, la Figura 26 muestra la existencia de correlación entre los contenidos de Fe y S en la matriz **X**. Esto es comprensible ya que, al tratarse de minerales sulfurosos, tienen presencia de pirita (FeS_2) en su estructura cristalina.

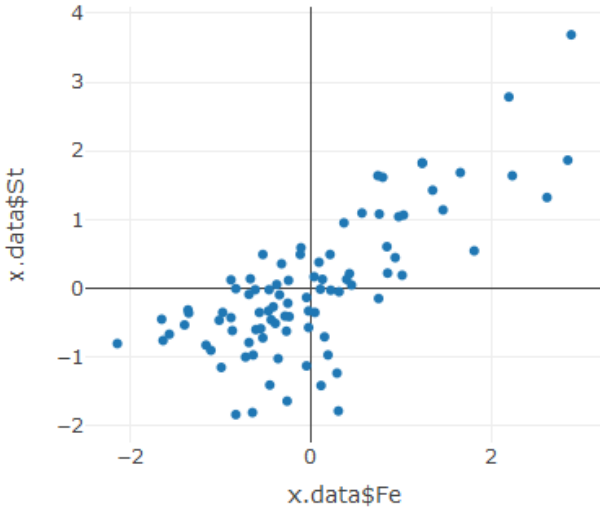

Figura 26. Relación entre Fe y S

Con el objetivo de detectar los problemas de colinealidad se calculó el índice de condicionamiento (CI), realizando un PCA al bloque de variables predictoras para obtener los valores propios, resultando un CI $= \sqrt{5.5505/0.00044} = 112.32$, el cual muestra un grave problema de colinealidad, CI superior a 30. Asimismo, la Tabla 6 muestra el índice de inflación de varianza (VIF) tras realizar la regresión lineal para predecir la concentración de cobre en el concentrado de flotación. Se observa que muchas variables predictoras como Cu, CuSS, CuRes, Fe, etc., tienen valores superiores a 10.

Tabla 6. VIF

Variables	VIF
Cu	1155.33
CuSS	92.93
CuCN	849.54
CuRes	156.41
Fe	18.4
St	16.64
Al	7.51
Mg	9.49
As	6.55
Bi	6.69
Ni	6.29
Sc	5.02
V	12.13

4.2 Análisis Explorado De Datos

4.2.1 Pre-procesamiento de Datos

Un paso anterior a la construcción de modelos es el pre-procesamiento de datos. El centrado se aplica sustrayendo el valor promedio de cada variable de la matriz de datos para facilitar la interpretación de los modelos. El escalado controla la importancia que se da inicialmente a cada variable. El más común es el escalado a varianza unitaria, que se consigue dividiendo cada variable entre su desviación típica. Es muy recomendable cuando las variables están medidas en diferentes unidades. El pre-procesamiento más habitual es el llamado auto-escalado, tras el cual el valor medio de cada variable (columna de la matriz de datos) será cero y su varianza será uno (Dunn, 2018).

En los datos del bloque X, las unidades de medida de las concentraciones de los elementos están en diferentes unidades (gramos por tonelada (g/t), porcentajes (%), etc.), por tanto, se aplicará auto-escalado utilizando las siguientes ecuaciones, $x_k = \frac{x_{k;datos} - promedio(x_{k;datos})}{desviación\ estándar(x_{k;centrado})}$, donde *k* es el indicador de variable.

4.2.2 Análisis Global

Para realizar el análisis exploratorio de los datos, se usa la matriz global resultante de añadir a la matriz **X** las columnas de la matriz **Y**, y se realiza un análisis de componentes principales (PCA). La Tabla 7 muestra los autovalores, la varianza y la varianza acumulada del PCA realizado sobre la matriz de datos global de las primeras 15 componentes extraídas, que alcanzan a explicar hasta un 85% de la variabilidad de los datos. La Figura 27 muestra el gráfico de la varianza explicada por cada componente.

Tabla 7. PCA del bloque global

Componente	Autovalor	Varianza	Varianza Acumulada
1	5.7	15.0	15.0
2	4.7	12.4	27.4
3	3.7	9.8	37.2
4	3.1	8.1	45.3
5	2.5	6.6	51.9
6	2.3	6.0	57.9
7	2.0	5.4	63.3
8	1.5	3.9	67.1
9	1.2	3.2	70.4
10	1.2	3.2	73.6
11	1.0	2.8	76.3
12	1.0	2.7	79.0
13	0.9	2.3	81.2
14	0.8	2.2	83.5
15	0.7	1.9	85.4

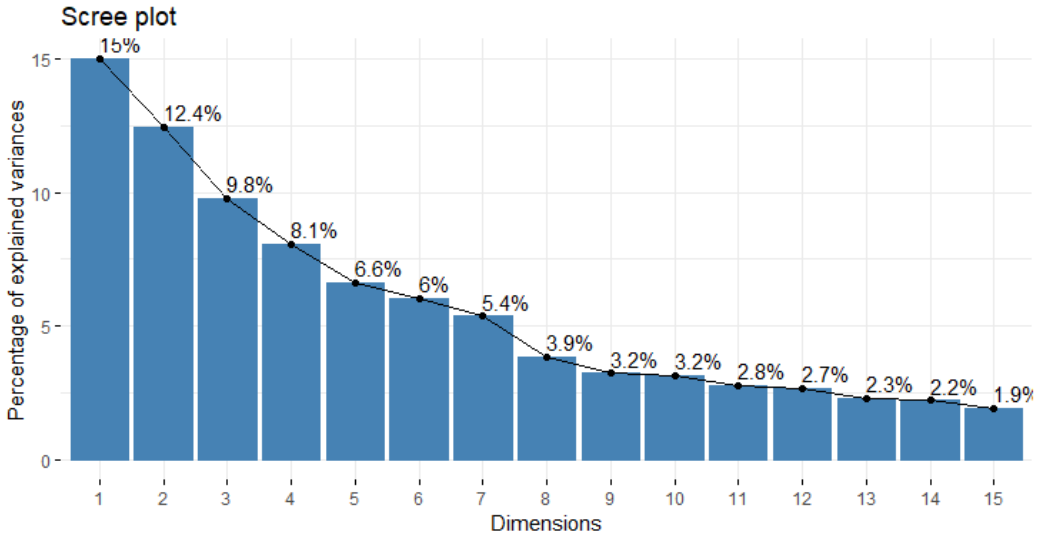

Figura 27. Porcentaje de varianza explicada del bloque global

La Figura 28 muestra la evolución de la bondad de ajuste (R^2) y de predicción (Q^2) obtenida hasta extraer 15 componentes. Se observa que el crecimiento de la capacidad de predicción es mínimo a partir de la componente 15, por tanto, considerando este comportamiento y sumado a la variabilidad explicada, se ha decidido extraer 15 componentes.

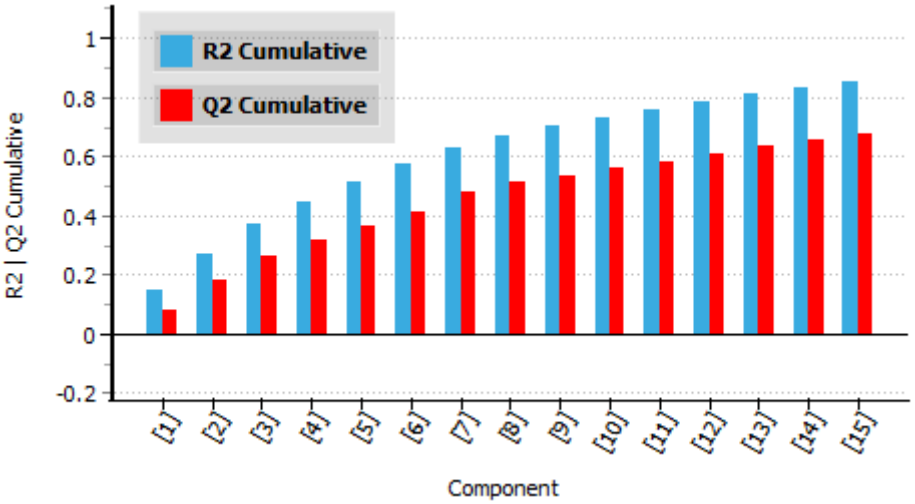

Figura 28. R^2 y Q^2 del PCA global

Usualmente los estadísticos, error cuadrado de predicción (SPE) y la T^2 de Hotelling (T^2) se usan para validar el modelo y detectar observaciones atípicas y extremas, contrastando si las observaciones toman valores para estos estadísticos muy por encima de su percentil 95%. Por tanto, es esperable que si no hay datos anómalos haya como mucho un 5% de valores ligeramente por encima de estos límites de los estadísticos.

La Figura 29 muestra el límite de control para el estadístico T^2 de Hotelling y es utilizado para determinar observaciones extremas, mientras que, la Figura 30 muestra el límite de control para el estadístico SPE y es utilizado para detectar observaciones atípicas, que rompan la estructura de correlación en los datos. De los gráficos anteriores, se observa que la observación *51* es la observación más extrema y la *54* la más atípica.

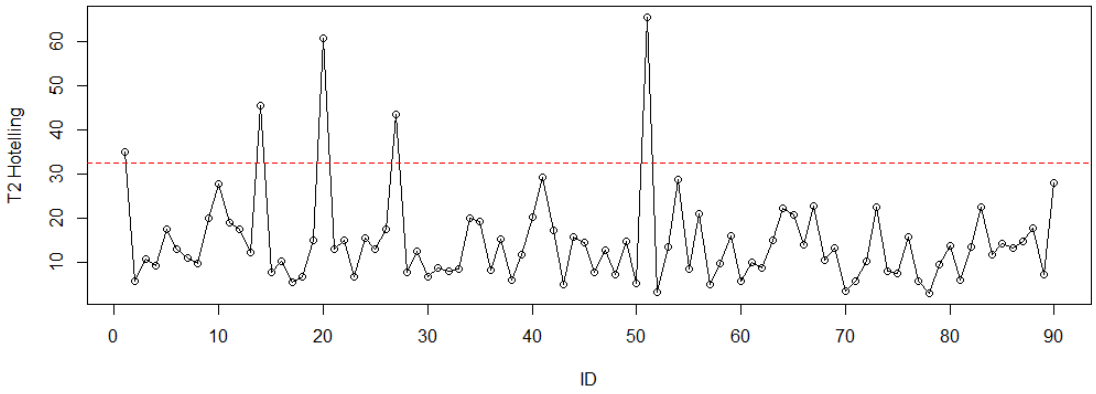

Figura 29. T^2 de Hotelling del PCA global

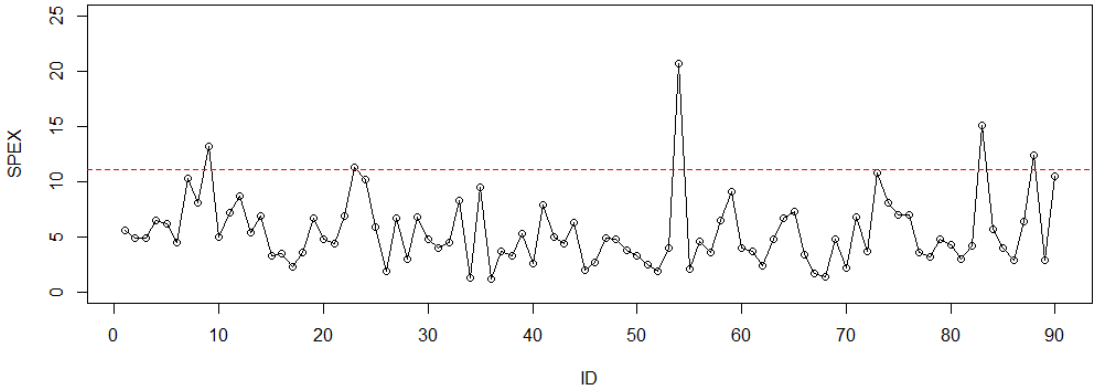

Figura 30. SPEX del PCA global

Con el objetivo de identificar la/s variable/s que pueden estar influyendo en la atipicidad de estas observaciones, se construyen los gráficos de contribución. La Figura 32 muestra que la observación *51* presenta altos valores de la concentración de arsénico y bismuto, mientras que, la Figura 32 muestra que la observación *54* presenta altos valores de la concentración de Ca. Consultados los expertos del proceso, se confirma la anomalía y se decide eliminar estas observaciones para los siguientes análisis.

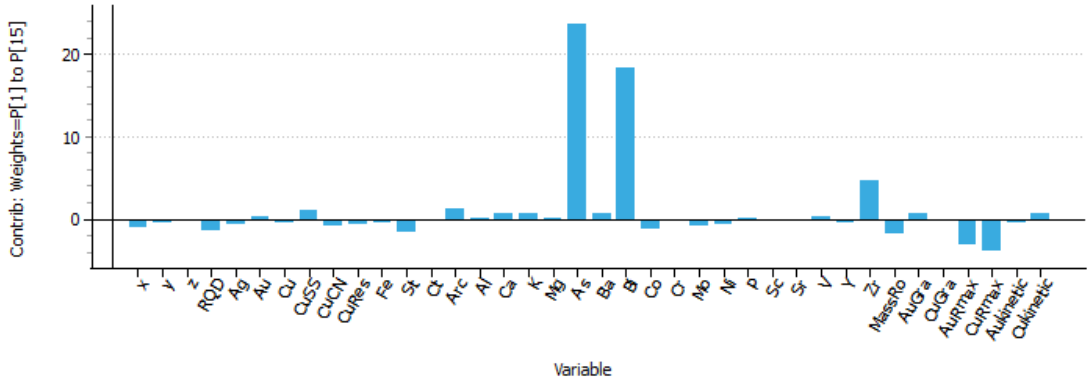

Figura 31 Contribución de la observación 51 en el PCA global

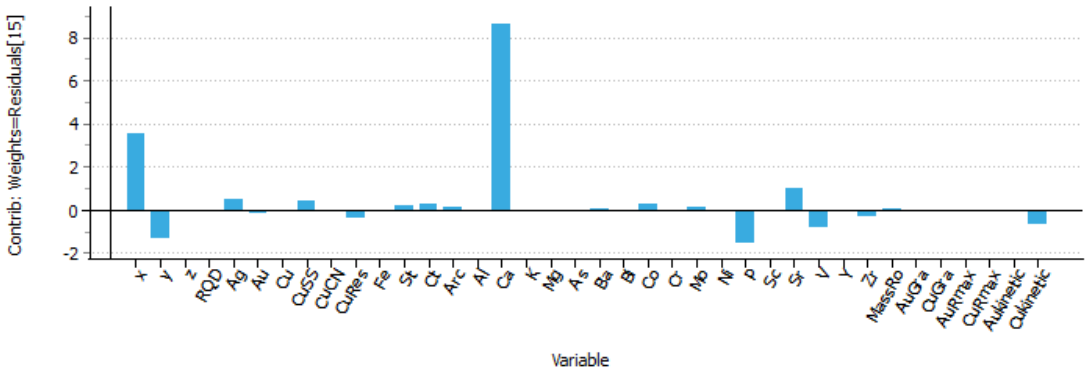

Figura 32. Contribución de la observación 54 en el PCA global

Una vez eliminadas las observaciones anómalas, se vuelve a realizar un nuevo análisis de componentes principales. Siguiendo con la misma metodología, la observación *1* es eliminada; en el gráfico de contribución realizado se observa que las variables causantes fueron los altos valores de la concentración de arsénico y bismuto. Posteriormente, la observación *20* es eliminada; se observa que las variables causantes fueron los altos valores de la concentración de cromo y níquel. Posteriormente, la observación *14* es eliminada; se observa que las variables causantes fueron los altos valores de la concentración de plata y oro. Posteriormente, la observación *73* es eliminada; se observa que la variable causante fueron los altos valores de la concentración de calcio.

La Figura 33 y la Figura 34 muestran que, una vez eliminadas las observaciones anómalas, las observaciones están básicamente dentro de los límites de control en los gráficos de T² de Hotelling y SPEX, respetando la tasa de falsas alarmas esperada.

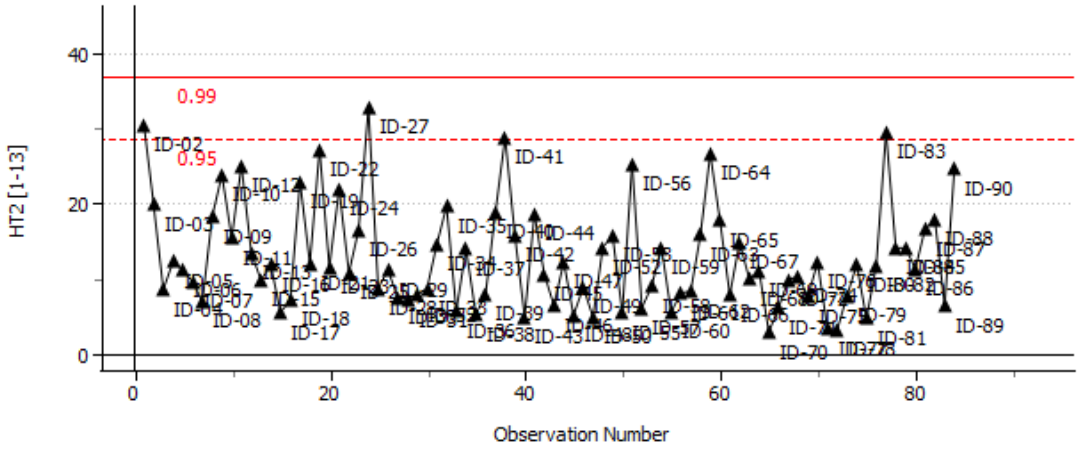

Figura 33. T² de Hotelling del PCA global sin atípicos

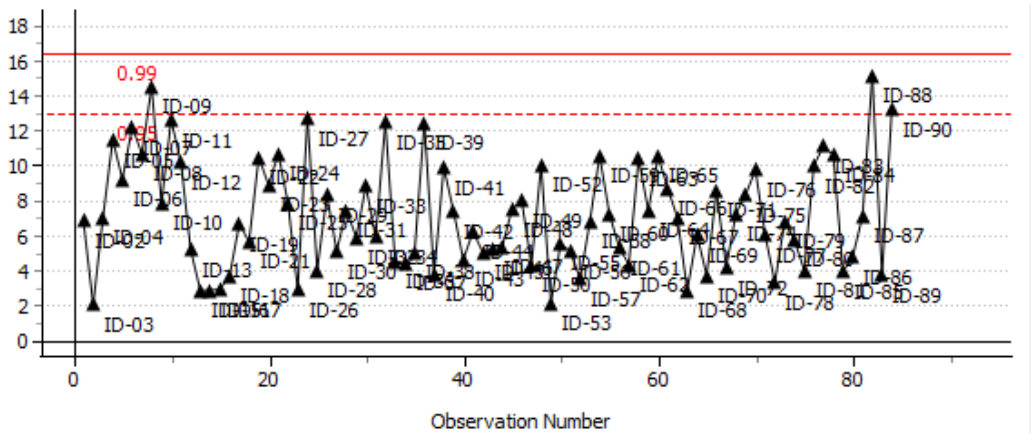

Figura 34. SPEX del PCA global sin atípicos

La construcción de diagrama de cajas es un método gráfico para mostrar información de datos de manera univariante (Tukey, 1977). Es un método alternativo comúnmente utilizado para mostrar los valores extremos inferiores y superiores. La Figura 35 muestra el diagrama de cajas del bloque global de datos, donde se observa que las observaciones atípicas vistas de manera univariante, como por ejemplo, las observaciones 1, 51, 20, 14, 73, son cazadas o detectadas de manera multivariante por el análisis exploratorio con PCA realizado.

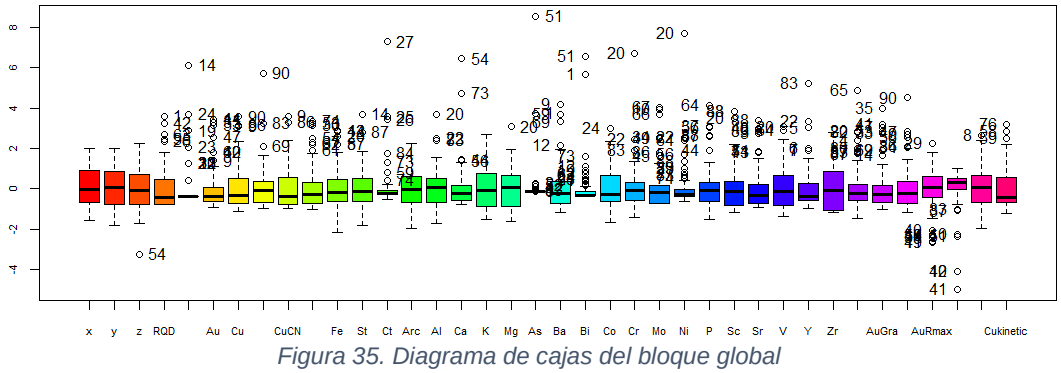

Figura 35. Diagrama de cajas del bloque global

En el anexo 7.1 se resumen los parámetros estadísticos clásicos de todas las variables realizado mediante análisis univariante. Se muestra, por ejemplo, la curtosis o apuntamiento, que es una medida que determina el grado de concentración de los datos cercanos a su media. La tabla del anexo señala en rojo, aquellas variables con curtosis muy elevadas (gran apuntamiento). Se comprueba que las variables calcio, arsénico y bismuto, que tienen curtosis

elevadas, corresponden a aquellas variables para las que se detectaron datos anómalos en el PCA global.

La Figura 36 muestra el *loading plot* de las dos primeras componentes, con las observaciones anómalas previamente eliminadas. Para interpretar este gráfico hay que recordar que variables en posiciones cercanas (pero lejos del origen) están correlacionadas positivamente y variables en posiciones opuestas están correlacionadas negativamente.

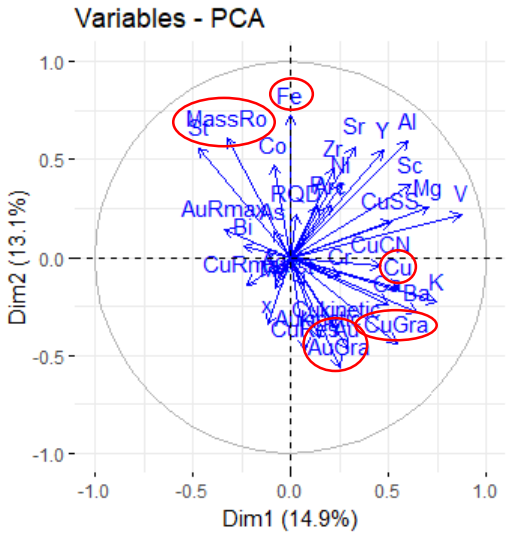

Figura 36. Loading plot del PCA global sin atípicos

Gracias al análisis de componentes principales se puede identificar estructuras de correlación entre variables. Por ejemplo, se puede intuir que existe una correlación positiva entre la variable Y “MassRo” y la X “St”, o entre la variable Y “AuGra” y la X “Au”. De la misma manera, se puede intuir correlación negativa entre las variables Y “MassRo” y “CuGra”. La Figura 37 y Figura 38 corroboran estas “intuiciones”.

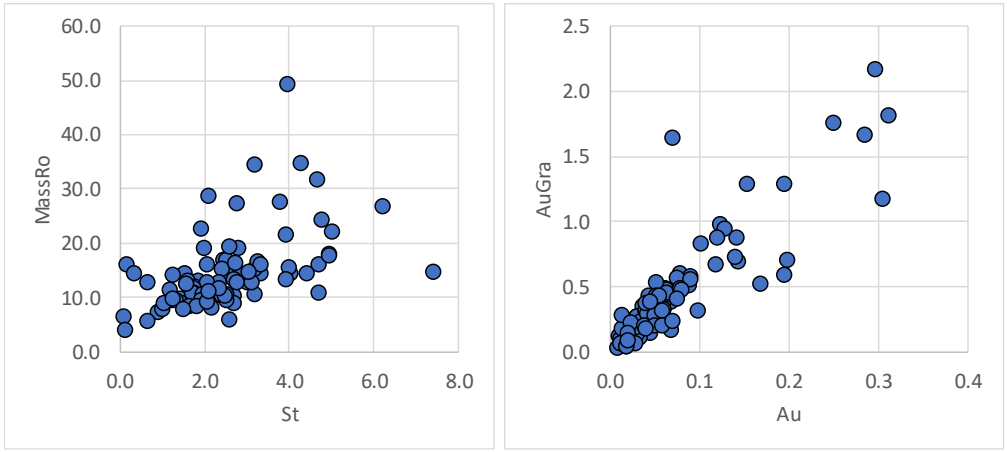

Figura 37. Estructuras de correlación positivas en el PCA global

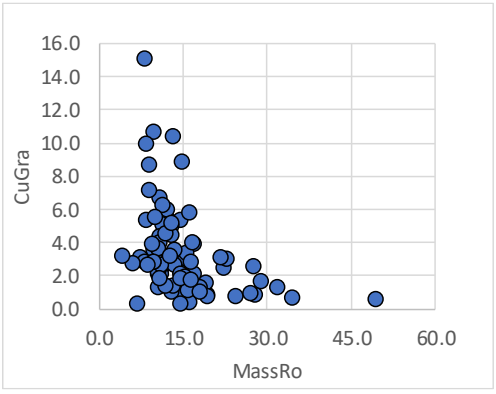

Figura 38. Estructura de correlación negativa en el PCA global

De aquí se puede sospechar que los modelos que intenten explicar, por ejemplo, la variable respuesta “MassRo”, estarán relacionados con las variables predictoras “St” y “CuGra”.

La Figura 39 corrobora la intuición que no existe correlación entre las variables X “Cu” y “Fe”.

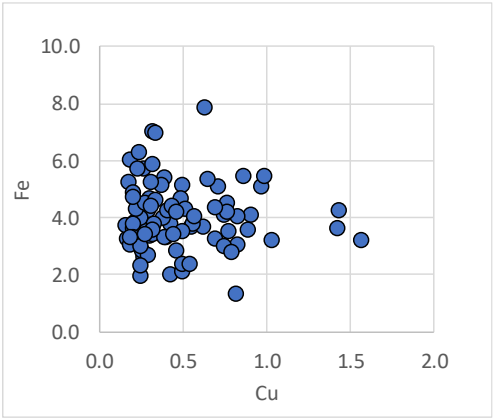

Figura 39. Estructura sin correlación en el PCA global

La Figura 40 (gráfico izquierdo) muestra, por ejemplo, el gradiente de la concentración de Al con dirección noreste. Las observaciones que estén en la dirección de este gradiente presentarán mayor concentración de Al; así la observación *83* presenta más Al que la observación *76*. Este gráfico es intuido del *loading plot* (Figura 36), ya que la variable Al está alejada del origen y en la dirección noreste. El mismo análisis puede ser aplicado a la concentración de Cu (ver Figura 40, gráfico derecho).

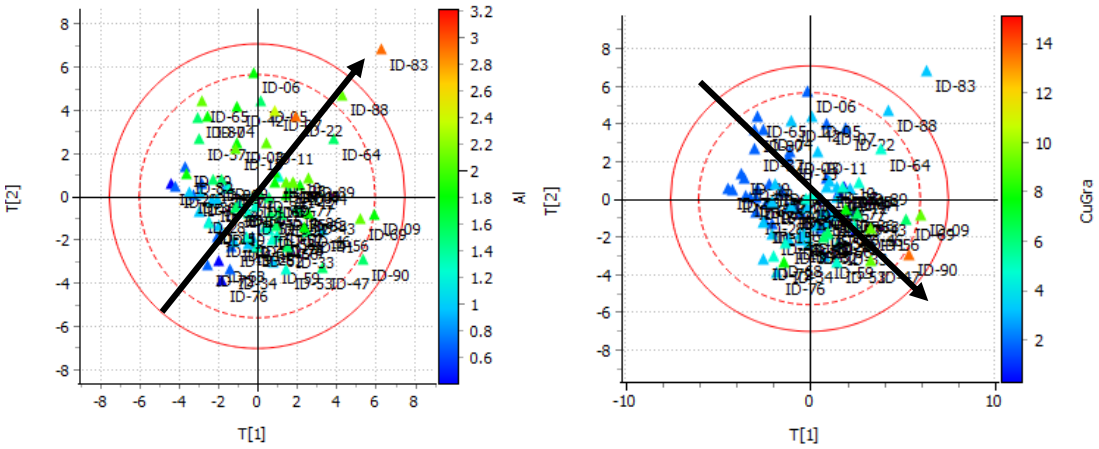

Figura 40. Gradiente en el PCA global

4.3 Predicción Global

4.3.1 PLS2

Una vez realizado el análisis exploratorio del modelo global y, tras eliminar las observaciones anómalas, se procede a realizar un modelo PLS con el grupo de la matriz **X**, como variables predictoras, y el grupo la matriz **Y**, como variables respuestas.

La Figura 41 muestra la evolución de la bondad de ajuste (R^2) y de predicción (Q^2) obtenida hasta extraer 5 componentes. Se observa que el crecimiento de la capacidad de predicción es mínimo a partir de la componente 5, por tanto, se ha decidido extraer cinco componentes.

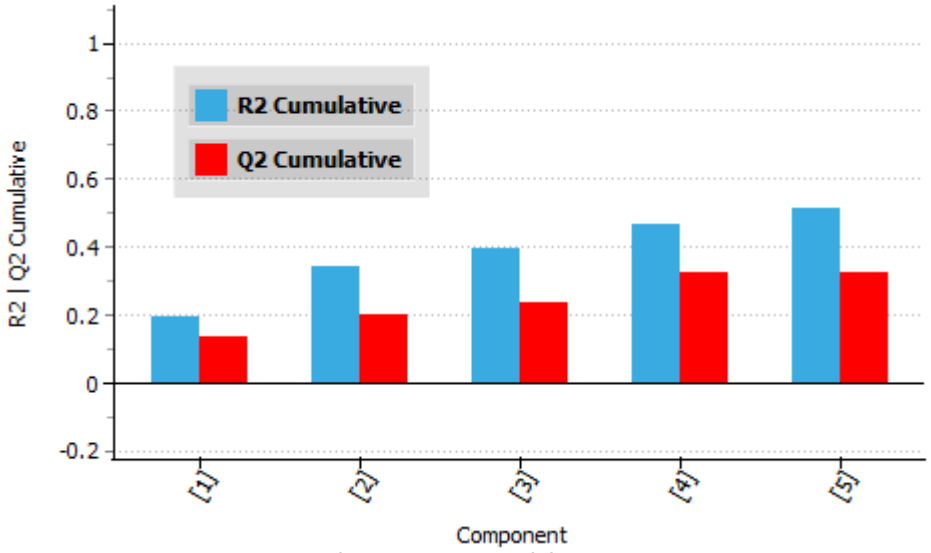

Figura 41. R² y Q² del PLS2

La Figura 42 muestra el gráfico *w*,c* de las dos primeras componentes. Las variables respuestas en posiciones cercanas (pero lejos del origen) a las variables predictoras están correlacionadas positivamente y variables respuestas en posiciones opuestas a las variables predictoras están correlacionadas negativamente.

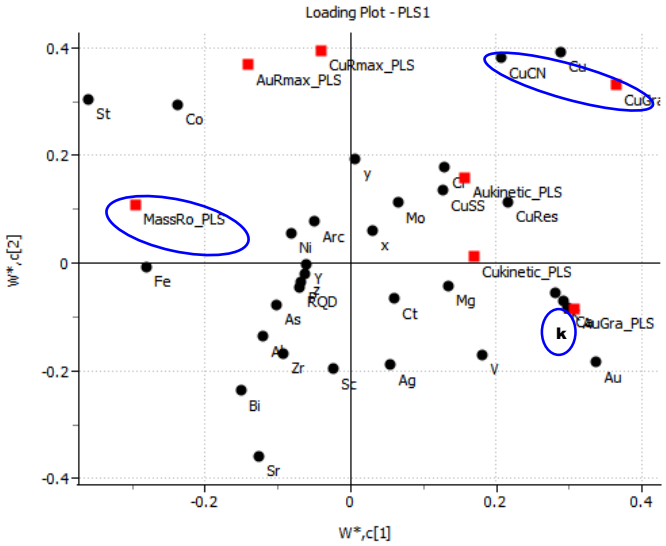

Figura 42. Gráfico w, c del PLS2*

La Figura 43 muestra los límites de control (al 95% y 99%), del modelo PLS con cinco componentes, para el estadístico T² de Hotelling, mientras que, la Figura 44 los muestra para el estadístico SPE. De los gráficos anteriores, las observaciones *41* y *27* podrían mostrar algún comportamiento anómalo, sin

embargo, el comportamiento no es demasiado exagerado, por tanto, no serán apartadas del análisis futuro.

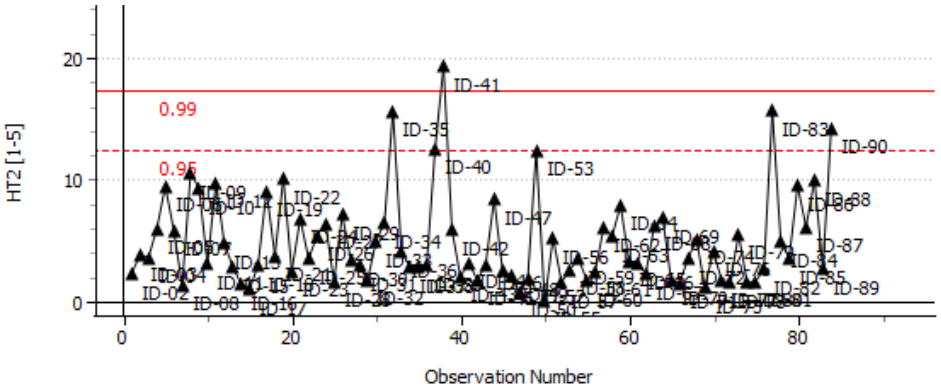

Figura 43. T² de Hotelling del PLS2

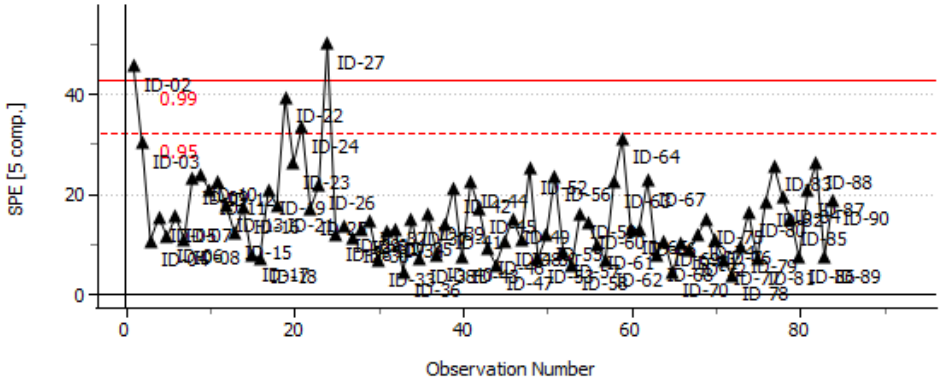

Figura 44. SPEX del PLS2

La Figura 45 corrobora la identificación de estructuras de correlación entre variables, intuidas de la Figura 42. Por ejemplo, se puede intuir que existe una correlación positiva entre la variable Y “MassRo” y la X “K”, o una correlación negativa entre la variable Y “CuGra” y la X “CuCN”.

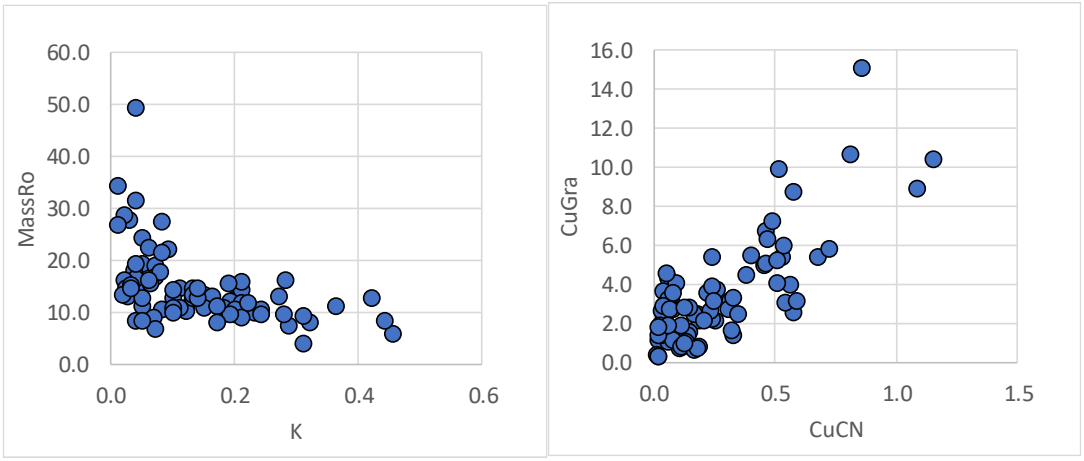

Figura 45. Estructuras de correlación positiva del PLS2

La Figura 46 (gráfico izquierdo) muestra, el gradiente de la variable Y “AuRmax” con dirección noroeste. Las observaciones que estén en la dirección de este gradiente presentarán mayor recuperación de Au; así la observación *29* presenta mayor recuperación de Au que la observación *41*. Este gráfico es intuido del *loading plot* (Figura 42), ya que la variable Y “AuRmax” está alejada del origen y en la dirección noroeste. El mismo análisis puede ser aplicado a la variable Y “AuGra” (ver Figura 46, gráfico derecho).

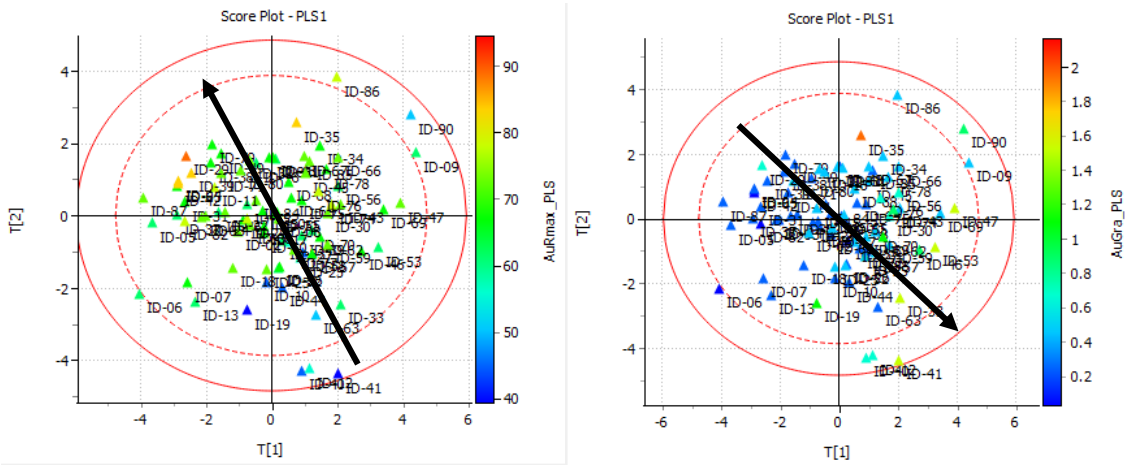

Figura 46. Gradiente del PLS2

Debido a la gran cantidad variables predictoras que existen en el modelo inicial, y a que no todas tienen por qué estar relacionadas con las variables respuesta, es conveniente tratar de simplificar el modelo PLS eliminando los predictores no relacionados con las respuestas. La Figura 47 muestra el proceso de depuración del modelo PLS, mediante el cálculo del VIP para cada variable, el cual es un indicador de la importancia de cada variable en la predicción de al menos una de las variables respuesta. Este proceso de depuración es un proceso secuencial que consiste en eliminar aquellas variables que tengan un valor VIP por debajo de uno, reestimar el modelo resultante, calcular los nuevos VIP, y repetir el procedimiento hasta que no haya predictores con VIP<1.

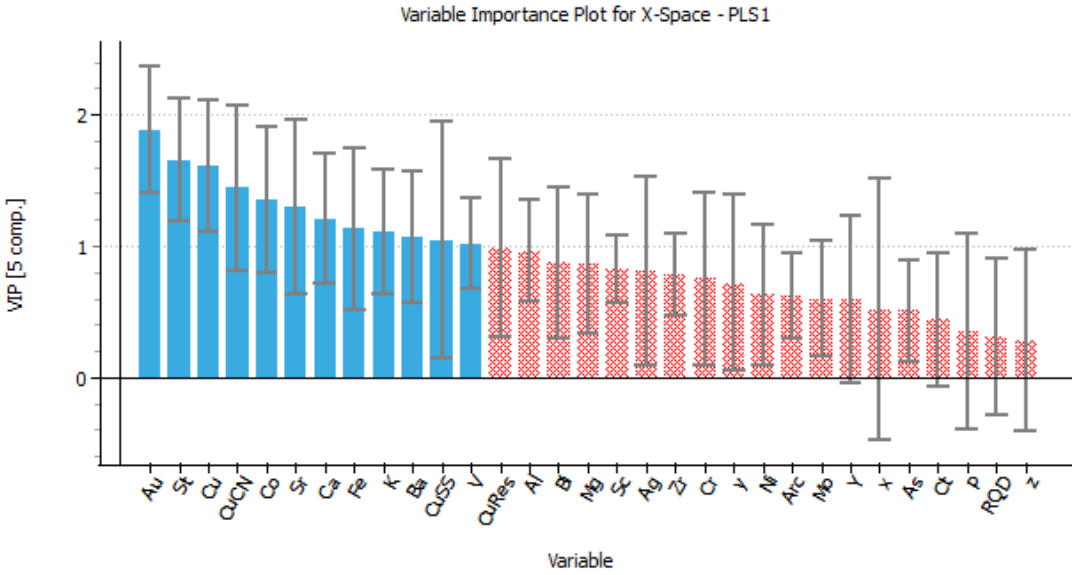

Figura 47. VIP del PLS2

La Figura 48 muestra los intervalos de confianza de los coeficientes de regresión del modelo PLS con cinco componentes. Se confirma que, el intervalo de confianza *Jackknife* de las variables a eliminar con el criterio de VIP contienen el valor cero (indica que los coeficientes no estadísticamente significativos al 5%).

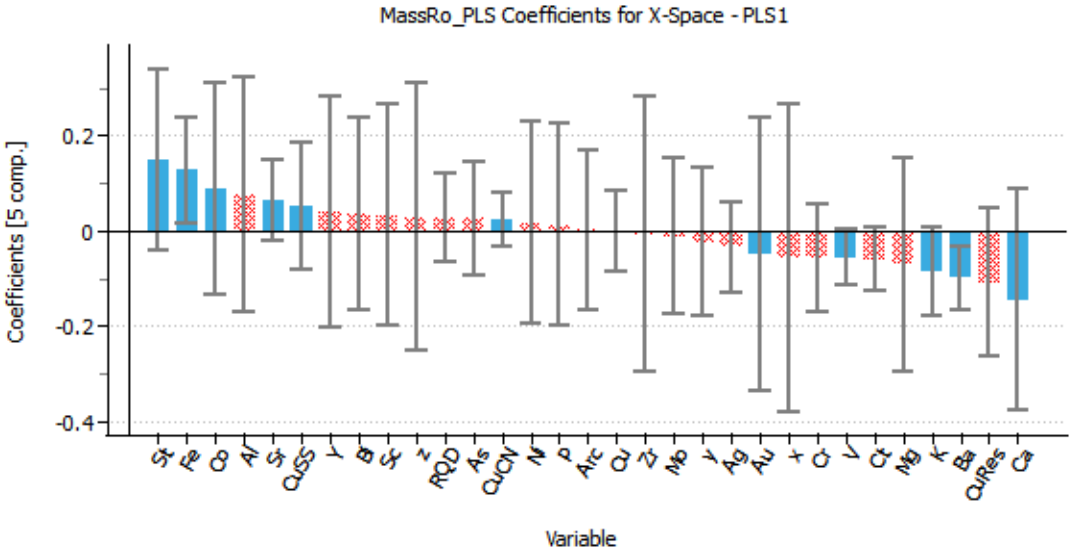

Figura 48. Intervalo de confianza Jackknife de los coeficientes del PLS2

La Figura 49 muestra los límites de control (al 95% y 99%) para el estadístico T^2 de Hotelling y la Figura 50 para el estadístico SPE, aplicados al modelo PLS, con cinco componentes, después de la primera depuración de variables, mediante el criterio del VIP. De los gráficos, las observaciones *35* y *64* podrían mostrar algún comportamiento anómalo.

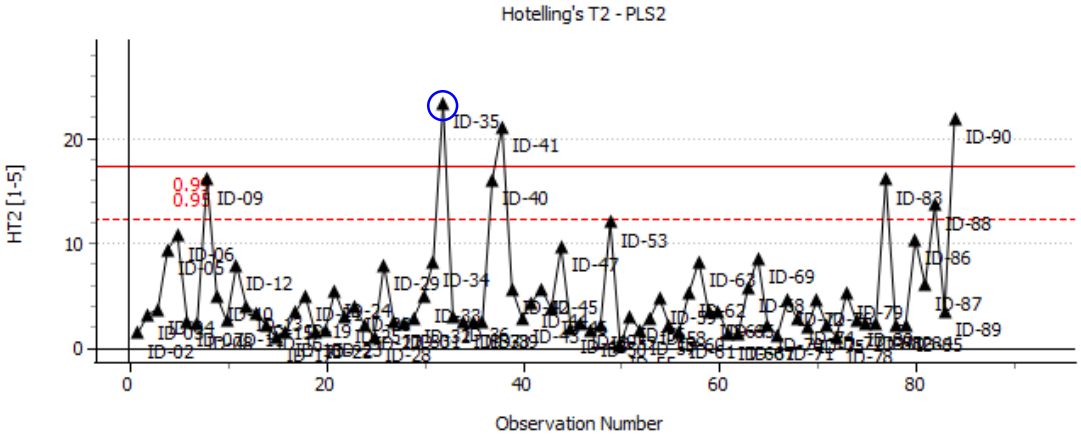

Figura 49. T² de Hotelling PLS2

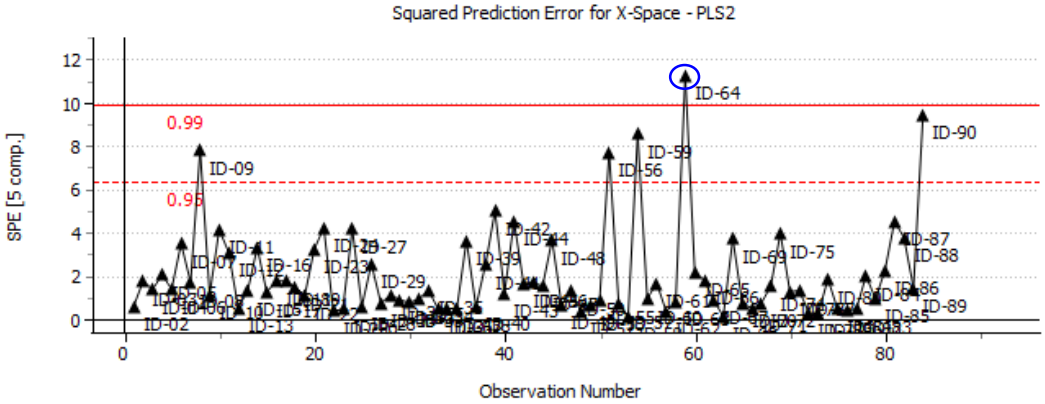

Figura 50. SPEX PLS2

Con el objetivo de identificar la/s variables que puede/n ser responsables de las anomalías se calculan los gráficos de contribución. La Figura 51 muestra que la observación *35* presenta altos valores en la concentración de oro, mientras que la Figura 52 muestra que la observación *64* presenta altos valores en la concentración de Ca y V, y bajos valores de K. Consultados con los expertos se decide eliminarlas del estudio.

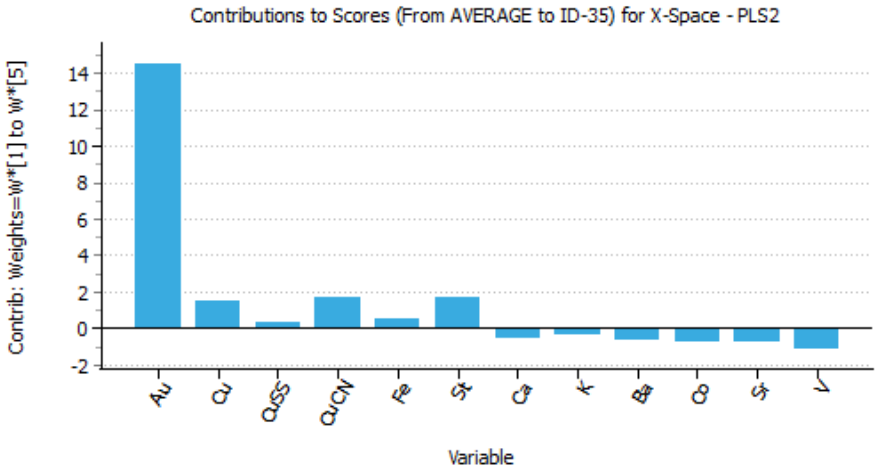

Figura 51 Gráfico de contribución de la observación 35 del PLS2

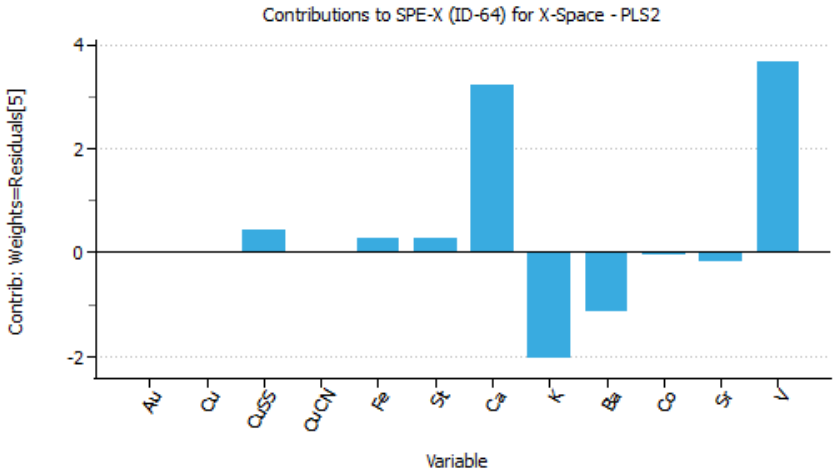

Figura 52. Gráfico de contribución de la observación 64 del PLS2

Siguiendo con el proceso secuencial de depuración del modelo PLS. La Figura 53 muestra el cálculo del VIP para cada variable. Se eliminan aquellas variables que tengan un valor VIP por debajo de uno y se vuelve a reestimar el modelo resultante.

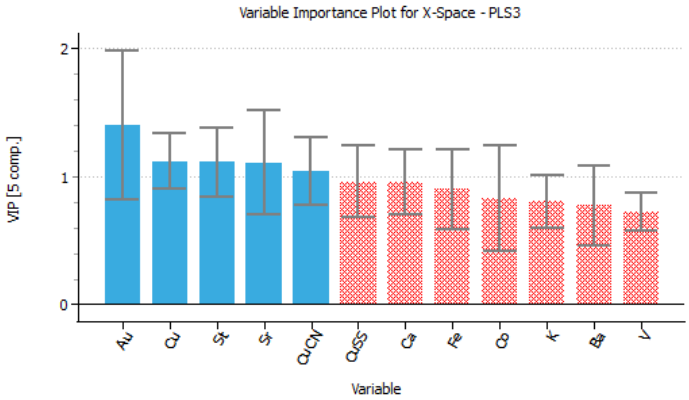

Figura 53. VIP del PLS2

La Figura 54 muestra los intervalos de confianza de cada uno de los coeficientes de regresión del nuevo modelo PLS. Sí se utilizara el criterio de eliminación de variables según el VIP (menores que uno), las variables Fe y Ca serían eliminadas. Sin embargo, el intervalo de confianza *Jackknife* para las variables Fe y Ca no contienen el valor cero. Las variables Fe y Ca, ayudan a predecir al menos una de todas las variables respuestas. La Figura 54 muestra que variables Fe y Ca aportan de manera significativa en la predicción de la variable respuesta “MassRo”. Por tanto, las variables Fe y Ca no serán eliminadas del modelo.

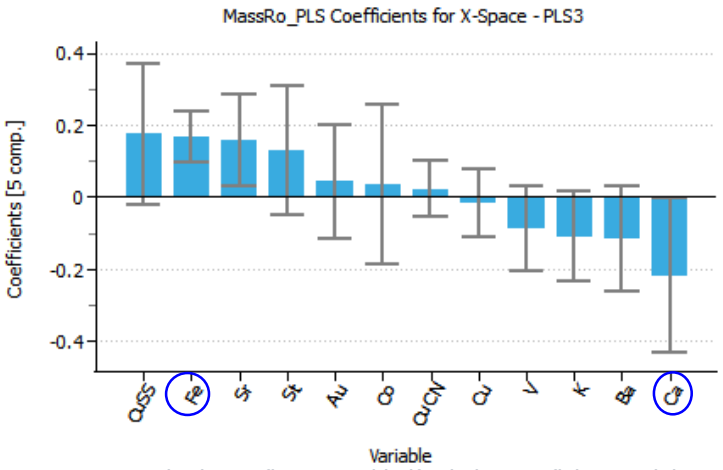

Figura 54. Intervalo de confianza Jackknife de los coeficientes del PLS2

Este proceso iterativo de depuración es realizado hasta encontrar el último modelo PLS, con tres componentes. Se valida el modelo con los límites de control (al 95% y 99%) para el estadístico T^2 de Hotelling y para el estadístico SPE observando que no hay valores atípicos. La Tabla 8 muestra los coeficientes de regresión del modelo PLS .

Tabla 8. Coeficientes de regresión del modelo PLS2

Variable	MassRo	AuGra	CuGra	AuRmax	CuRmax	Aukinetic	Cukinetic
Au	-0.06	0.70	0.13	-0.06	-0.16	-0.10	-0.20
Cu	0.03	0.04	0.41	0.10	0.12	0.06	0.00
CuCN	0.05	-0.04	0.37	0.12	0.13	0.06	0.00
Fe	0.20	0.05	-0.18	0.12	-0.04	-0.13	-0.21
St	0.30	-0.12	-0.10	0.22	0.04	-0.12	-0.23
Ca	-0.24	0.23	0.11	-0.18	-0.06	0.07	0.13

La Figura 55 muestra el gráfico *w*,c* de las dos primeras componentes del último modelo PLS. Se puede "intuir" que la primera componente explicará las variables respuestas "MassRo", "AuGra" y "CuGra" (por encontrarse lejos del origen), por otro lado, la segunda componente podría explicar "CuGra".

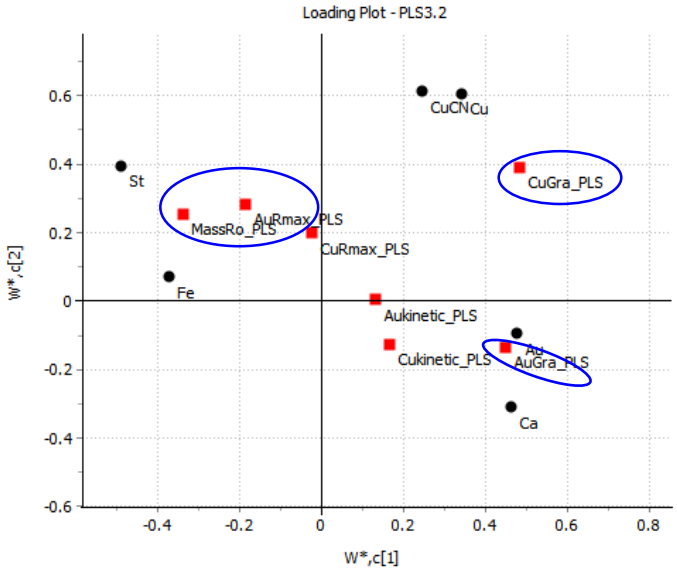

Figura 55. Gráfico w, c del modelo PLS2*

La Figura 56 confirma la intuición de la Figura 55. La primera componente explica mejor las variables respuestas “MassRo”, “AuGra” y “CuGra”. La segunda componente explica mejor “CuGra” y “CuRmax”.

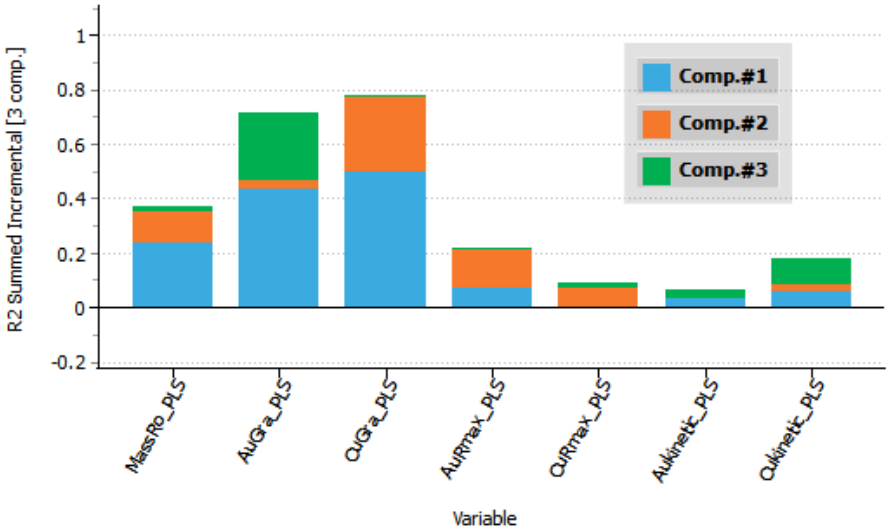

Figura 56. Resumen de componente del modelo PLS2

En el presente modelo PLS2, se obliga al modelo a predecir las siete variables respuestas. Sin embargo, hay variables predictoras que ayudan mucho a predecir cierta variable respuesta, pero otras no. Aunque el modelo PLS2 permite una interpretación global de la compleja estructura de relaciones entre todos los predictores y todas las respuestas (lo que puede ser muy útil para comprender el proceso en estudio), si el objetivo es predecir una respuesta, es mejor construir

un modelo PLS para esa respuesta. Esto, además, permitirá comparar el modelo PLS con otras técnicas de aprendizaje automático que solo pueden predecir una respuesta por modelo. La variable respuesta elegida es el CuGra, una de las mejores explicadas por el PLS2.

4.4 Predicción Individual

4.4.1 Predicción de CuGra

Todos los modelos que se describen en este apartado se han realizado con los mismos datos de entrenamiento y validación que han sido previamente depurados en el análisis exploratorio de datos.

Los modelos son entrenados con el 70% de los datos seleccionados aleatoriamente, dejando el 30% restante de los datos para la validación.

4.4.1.1 Regresión en Mínimos Cuadrados

La Figura 57 muestra la capacidad de predicción Q^2 y coeficiente de ajuste de datos R^2, para el modelo PLS. A partir de la cuarta componente, el Q^2 decrece, por lo que se han seleccionado tres componentes para este análisis.

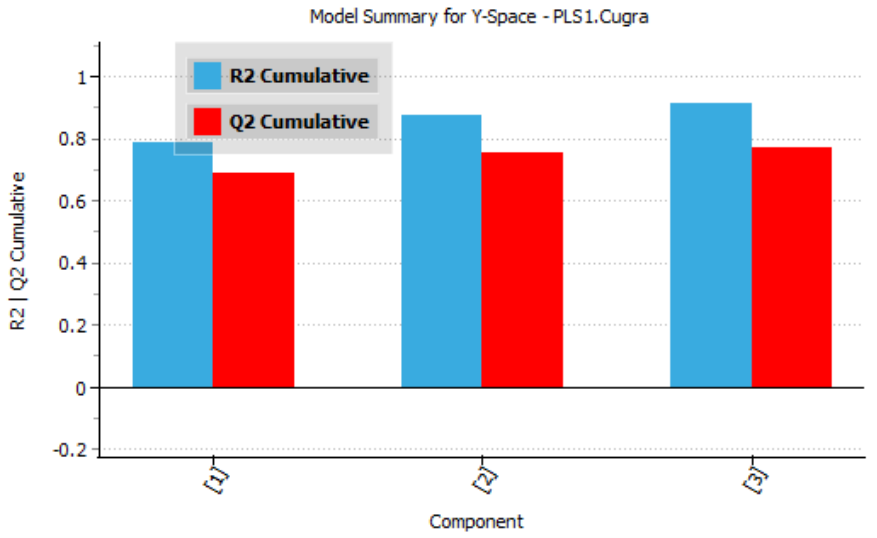

Figura 57. R^2 y Q^2 del PLS

La Figura 58 muestra los límites de control (al 95% y 99%) para el estadístico T^2 de Hotelling. Por su parte, la Figura 59 muestra los límites de control (al 95% y

99%) para el estadístico SPE. De los gráficos anteriores, se observa que las observaciones *83*, *64* y *27* muestran comportamientos anómalos.

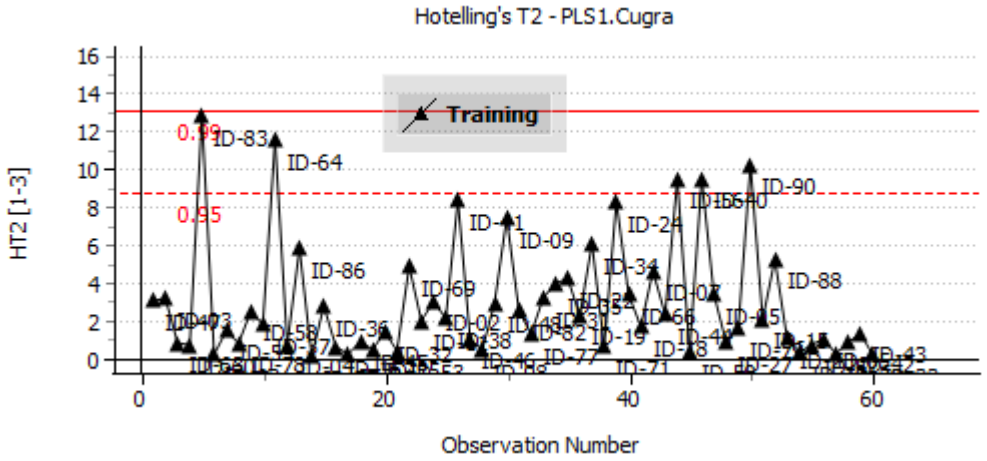

Figura 58. T^2 de Hotelling del PLS

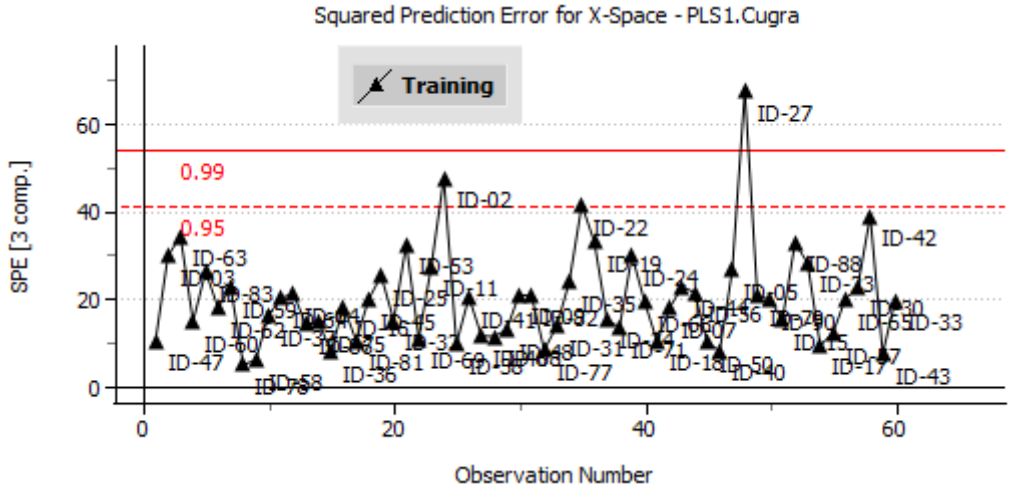

Figura 59. SPEX del PLS

La Figura 60 muestra el gráfico de contribución de la observación *27*, donde se observa altos valores del elemento carbono.

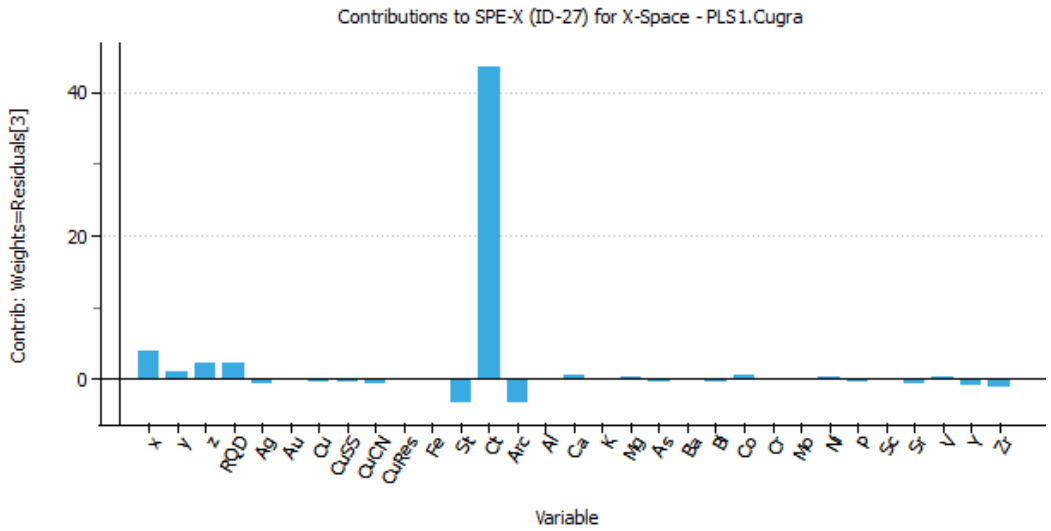

Figura 60. Contribución de observación 27 en el PLS

La Figura 61 muestra la relación interna del modelo. Se aprecia una tendencia lineal. No se piensa en adicionar términos cuadráticos al modelo o interacciones.

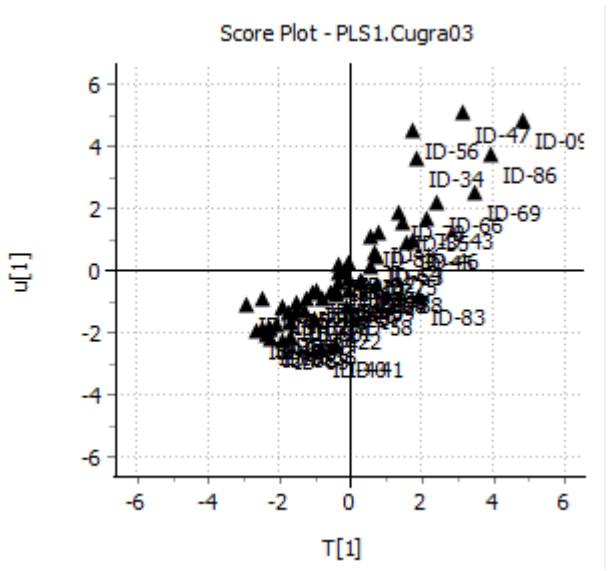

Figura 61. Estructura interna del PLS

La Figura 62 muestra los intervalos de confianza *Jackknife* (95%) para los coeficientes de las variables predictoras calculados con el número de componentes seleccionado.

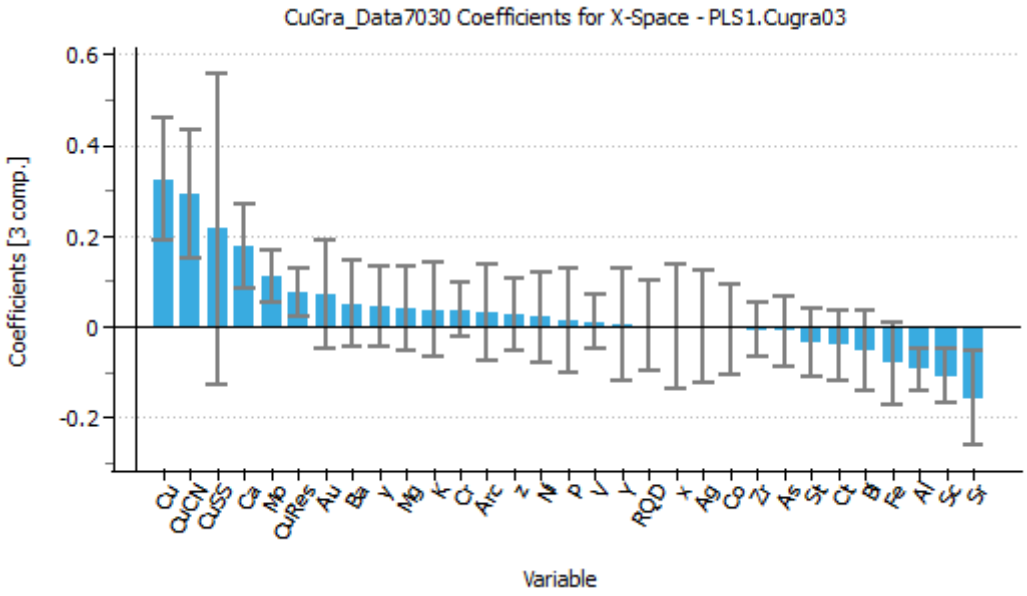

Figura 62. Intervalo de confianza Jackknife de los coeficientes del PLS

El modelo PLS se depura en este caso eliminando aquellas variables cuyos intervalos de confianza *Jackknife* contengan el valor cero (pues indicarán variables con coeficientes no estadísticamente significativos al 5%).

La Figura 63 muestra la bondad de predicción (Q^2) y de ajuste de datos (R^2), para el modelo PLS depurado.

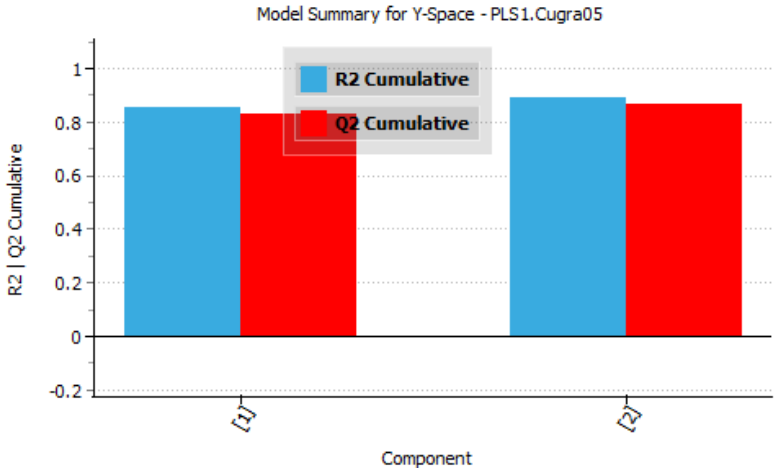

Figura 63. R^2 y Q^2 del PLS depurado

La Figura 64 y Figura 65 no detectan ninguna observación claramente anómala en el modelo.

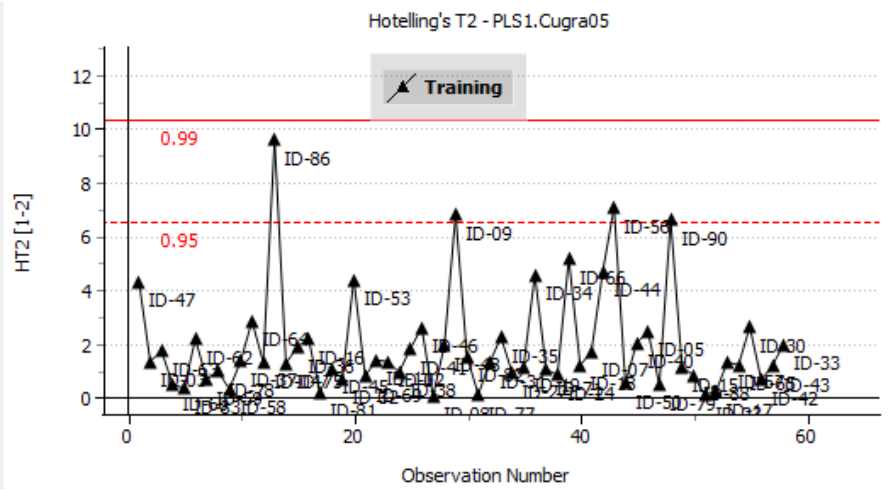

Figura 64. T² de Hotelling del PLS depurado

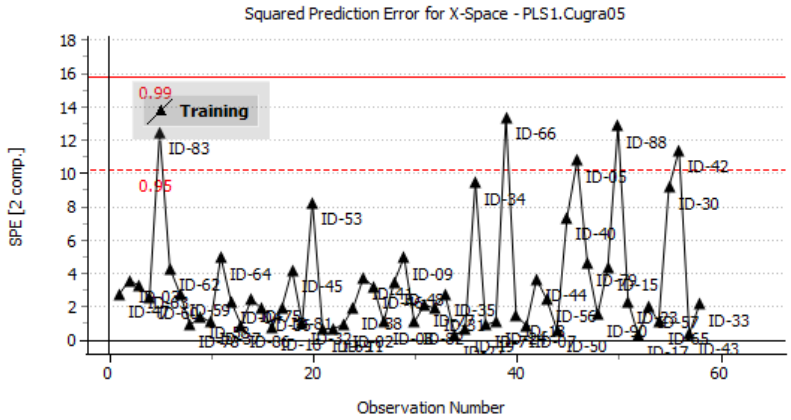

Figura 65. SPEX del PLS depurado

La Figura 66 muestra que, tras seleccionar aquellas variables cuyo intervalo de confianza *Jackknife* (95%) que no contienen al cero, quedan en el modelo las variables: Cu, CuCN, Ca, CuRes, Mo, Sr y Fe.

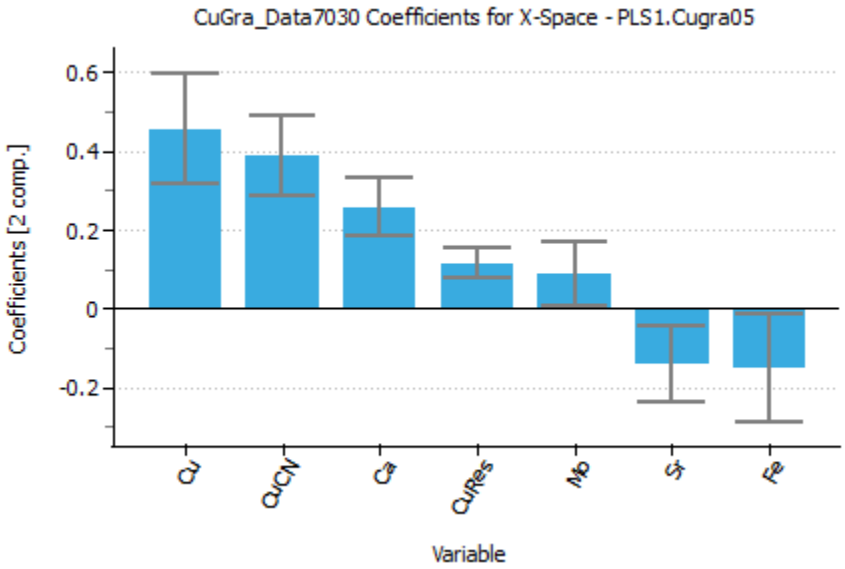

Figura 66. Intervalo de confianza Jackknife de los coeficientes del PLS depurado

La Tabla 9 muestra los coeficientes del modelo de regresión PLS depurado que predice el valor medio de la variable respuesta “CuGra” a través del contenido de Cu, CuCN, Ca, CuRes, Mo, Sr y Fe.

Tabla 9. Coeficientes del PLS depurado

Variable	CuGra
Cu	0.46
CuCN	0.39
Ca	0.26
CuRes	0.12
Mo	0.09
Sr	-0.14
Fe	-0.15

La Figura 67 muestra el gráfico *w*,c* de las dos primeras componentes. Las variables predictoras en posiciones cercanas (pero lejos del origen) a la variable respuesta “CuGra” están correlacionadas positivamente y variables predictoras en posiciones opuestas están correlacionadas negativamente.

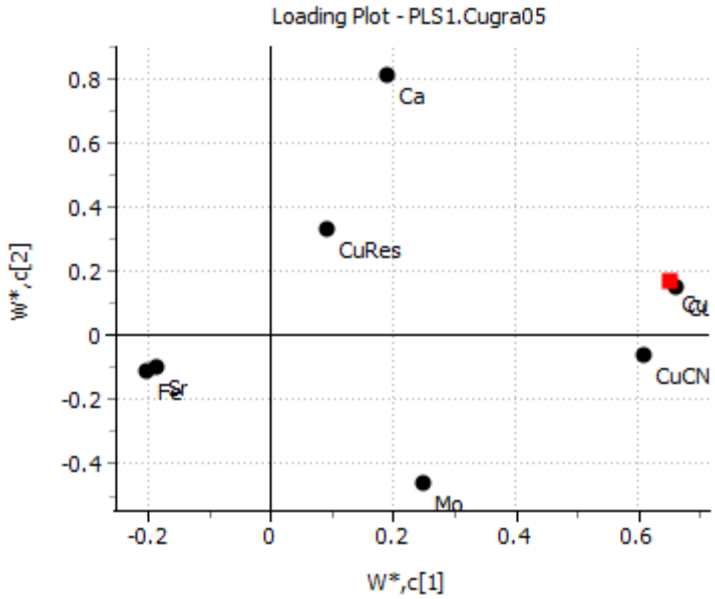

Figura 67. Gráfico w, c del PLS depurado*

La Figura 68 corrobora la identificación de estructuras de correlación entre variables, intuidas de la Figura 67. Por ejemplo, se puede intuir que existe una correlación positiva entre la variable X “Cu” y la Y “CuGra”.

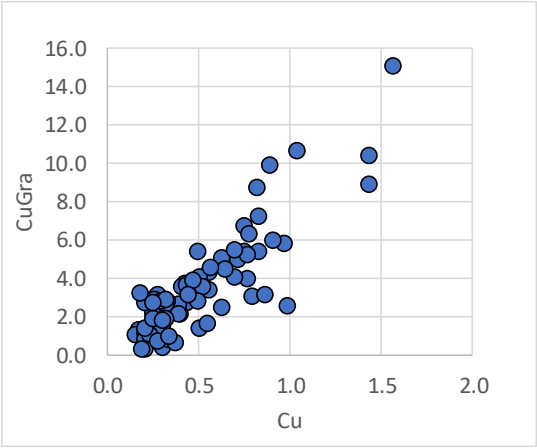

Figura 68. Estructuras de correlación positiva del PLS

La Figura 69 muestra, el gradiente de la variable Y "CuGra" con dirección noreste. Las observaciones que estén en la dirección de este gradiente presentarán mayor concentración de Cu; así la observación *90* presenta mayor concentración de Cu que la observación *06*. Este gráfico es intuido del *loading plot* (Figura 67), ya que la variable Y "CuGra" está alejada del origen y en la dirección noreste.

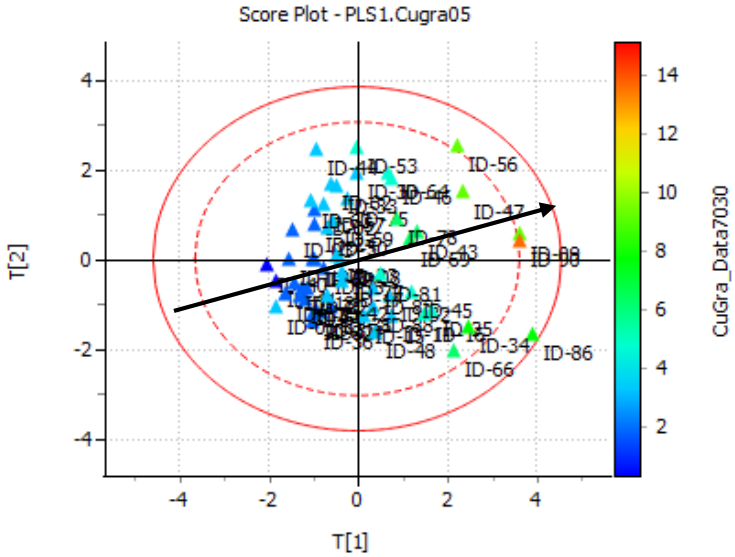

Figura 69. Gradiente del PLS

Una vez obtenido el modelo de regresión PLS, con los datos de entrenamiento, se obtiene una bondad de ajuste (R^2) de 0.89.

Las predicciones fueron realizadas con los datos de validación en el modelo PLS. La Figura 70 muestra la comparación de las predicciones con los datos

observados utilizando el modelo PLS. El error cuadrático medio 0.17, con una bondad de ajuste (R^2) de 0.60.

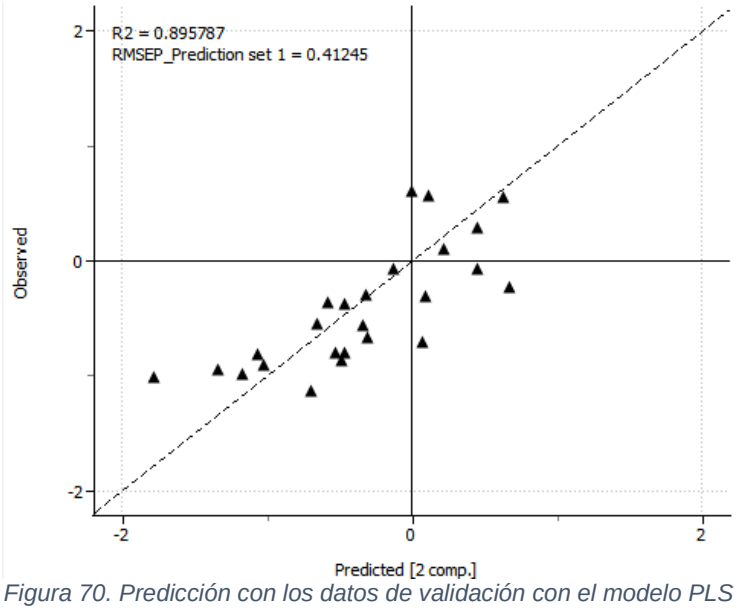

Figura 70. Predicción con los datos de validación con el modelo PLS

4.4.1.2 Regresión con Componentes Principales

Según el principio básico de un PCR, se utilizarán las variables latentes como variables predictoras. Debido a que el objetivo es predecir, se han seleccionado todas las componentes extraídas.

La regresión PCR fue realizada en dos pasos: el primero, donde se extraen las componentes del bloque de matriz **X** y el segundo, donde se realiza la regresión lineal múltiple de la variable respuesta “CuGra” con todas las componentes extraídas. La Figura 71 muestra los coeficientes de cada componente y sus p-valores asociados, los cuales nos indican su significación estadística. Se aprecia que no todos los componentes son estadísticamente significativos al 5% de riesgo de primera especie.

```
Call:
lm(formula = Y ~ ., data = pcr1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5418 -0.1053  0.0158  0.1387  0.5702

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.14e-16   4.00e-02    0.00  1.00000
PC1          1.21e-01   1.73e-02    6.97  1.4e-07 ***
PC2          2.67e-01   2.05e-02   12.99  2.2e-13 ***
PC3          2.88e-01   2.19e-02   13.14  1.7e-13 ***
PC4          1.02e-02   2.43e-02    0.42  0.67787
PC5         -2.54e-01   2.88e-02   -8.79  1.5e-09 ***
PC6          1.28e-02   3.05e-02    0.42  0.67811
PC7         -1.04e-01   3.34e-02   -3.10  0.00435 **
PC8          2.12e-01   3.43e-02    6.18  1.1e-06 ***
PC9          1.95e-02   3.90e-02    0.50  0.62130
PC10         6.21e-02   4.12e-02    1.51  0.14254
PC11        -5.26e-02   4.19e-02   -1.25  0.21993
PC12         1.15e-01   4.33e-02    2.65  0.01309 *
PC13         1.82e-01   4.45e-02    4.09  0.00033 ***
PC14         1.37e-01   4.83e-02    2.85  0.00814 **
PC15        -3.73e-02   5.29e-02   -0.71  0.48642
PC16        -6.76e-04   5.68e-02   -0.01  0.99059
PC17        -7.85e-02   6.29e-02   -1.25  0.22189
PC18        -2.04e-01   6.47e-02   -3.16  0.00377 **
PC19        -1.18e-01   7.15e-02   -1.65  0.10959
PC20         1.02e-02   7.41e-02    0.14  0.89143
PC21        -1.70e-01   7.95e-02   -2.13  0.04168 *
PC22         1.46e-01   8.26e-02    1.77  0.08698 .
PC23        -1.64e-01   9.62e-02   -1.71  0.09851 .
PC24         2.16e-01   1.07e-01    2.03  0.05204 .
PC25         1.58e-01   1.19e-01    1.32  0.19687
PC26         1.03e-02   1.27e-01    0.08  0.93614
PC27        -1.43e-01   1.46e-01   -0.98  0.33503
PC28        -4.89e-03   1.64e-01   -0.03  0.97648
PC29        -1.34e-01   2.27e-01   -0.59  0.55927
PC30        -5.19e-01   3.44e-01   -1.51  0.14216
PC31        -1.72e+00   2.27e+00   -0.76  0.45460
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.31 on 28 degrees of freedom
Multiple R-squared:  0.954,     Adjusted R-squared:  0.904
F-statistic: 18.9 on 31 and 28 DF,  p-value: 4.93e-12
```

Figura 71. Resultados del modelo PCR

Un método utilizado en regresión para detectar las variables importantes es realizar la regresión *stepwise*. En R la función *stepwise* para detectar las variables adecuadas para el modelo está basada en el criterio de información de Akaike (AIC). Se utiliza el método de selección automática en ambas direcciones, *backward* y *forward*, donde el modelo a escoger es el que tiene menor valor de AIC. Al ser las componentes ortogonales, la regresión *stepwise* es equivalente a eliminar las componentes menor valor de AIC. La Figura 72 muestra el modelo PCR final obtenido.

```
Call:
lm(formula = Y ~ PC1 + PC2 + PC3 + PC5 + PC7 + PC8 + PC10 + PC11 +
    PC12 + PC13 + PC14 + PC17 + PC18 + PC19 + PC21 + PC22 + PC23 +
    PC24 + PC25 + PC30, data = pcr1)

Residuals:
   Min     1Q Median     3Q    Max
-0.562 -0.138  0.012  0.131  0.625

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.25e-16   3.57e-02    0.00   1.0000
PC1          1.21e-01   1.54e-02    7.82  1.7e-09 ***
PC2          2.67e-01   1.83e-02   14.57  < 2e-16 ***
PC3          2.88e-01   1.96e-02   14.73  < 2e-16 ***
PC5         -2.54e-01   2.57e-02   -9.86  3.8e-12 ***
PC7         -1.04e-01   2.98e-02   -3.48   0.0013 **
PC8          2.12e-01   3.06e-02    6.93  2.7e-08 ***
PC10         6.21e-02   3.67e-02    1.69   0.0986 .
PC11        -5.26e-02   3.74e-02   -1.41   0.1673
PC12         1.15e-01   3.86e-02    2.97   0.0051 **
PC13         1.82e-01   3.97e-02    4.58  4.6e-05 ***
PC14         1.37e-01   4.30e-02    3.19   0.0028 **
PC17        -7.85e-02   5.61e-02   -1.40   0.1691
PC18        -2.04e-01   5.77e-02   -3.54   0.0010 **
PC19        -1.18e-01   6.38e-02   -1.85   0.0714 .
PC21        -1.70e-01   7.09e-02   -2.39   0.0216 *
PC22         1.46e-01   7.36e-02    1.99   0.0537 .
PC23        -1.64e-01   8.58e-02   -1.92   0.0626 .
PC24         2.16e-01   9.50e-02    2.28   0.0284 *
PC25         1.58e-01   1.06e-01    1.48   0.1462
PC30        -5.19e-01   3.07e-01   -1.69   0.0983 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.276 on 39 degrees of freedom
Multiple R-squared:  0.95,      Adjusted R-squared:  0.924
F-statistic: 36.7 on 20 and 39 DF,  p-value: <2e-16
```

Figura 72. Resultados del modelo PCR con stepwise

Una vez obtenido el modelo de regresión PCR, con los datos de entrenamiento, se obtiene una bondad de ajuste (R^2) de **0.95.**

La Figura 73 muestra la comparación de las predicciones con los datos observados utilizando el modelo PCR con *stepwise*. El error cuadrático medio (MSE) es 0.51, con una bondad de ajuste (R^2) de 0.26.

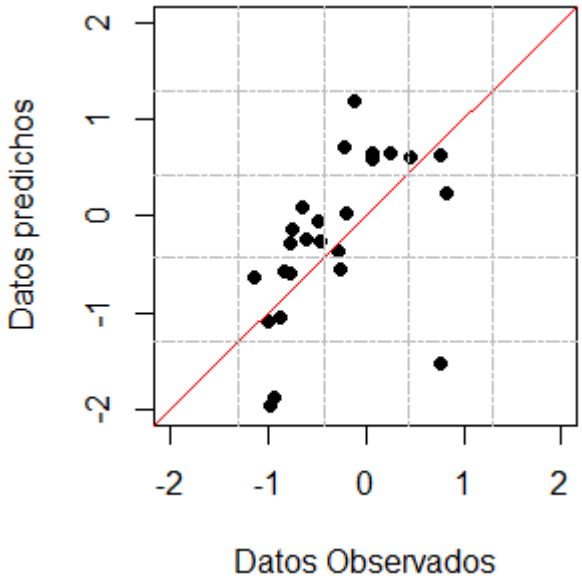

Figura 73. Predicción con los datos de validación con el modelo PCR

4.4.1.3 Árbol de Regresión

La variable respuesta “CuGra” es numérica, por tanto, un modelo de árbol de regresión es construido con el conjunto de variables predictoras.

En el modelo de árbol de regresión se obtiene la tabla de complejidad, que proporciona información de todos los árboles considerados para llegar al modelo final.

Para evitar el sobreajuste del modelo, se realiza la poda del árbol. La poda o *pruning* consiste en reducir el tamaño del árbol agrupando nodos en vez realizar excesivas divisiones. Generalmente es preferible tener un modelo de árbol de regresión sencillo.

La Figura 74 y Figura 75 muestran la tabla y el gráfico de complejidad, respectivamente, indicando el factor de complejidad cp (*cost complexity*), el error relativo por validación cruzada (X-val) y el tamaño del árbol creado (registro del criterio del número de hojas finales). El cp óptimo se obtiene cuando el error relativo es el mínimo. En este caso, se obtiene se obtiene cp mínimo con tamaño de árbol igual a 7.

```
Regression tree:
rpart(formula = y.data.tr ~ ., data = t.data, method = "anova",
    cp = 0.001, minsplit = 10)

Variables actually used in tree construction:
[1] Ca Cr Cu St z  Zr

Root node error: 72/60 = 1.2

n= 60

        CP nsplit rel error xerror  xstd
1  0.55228      0    1.0000  1.034 0.305
2  0.18168      1    0.4477  0.699 0.172
3  0.06057      2    0.2660  0.600 0.231
4  0.05156      3    0.2055  0.591 0.231
5  0.02353      4    0.1539  0.566 0.234
6  0.01297      5    0.1304  0.556 0.235
7  0.00671      6    0.1174  0.528 0.235
8  0.00512      7    0.1107  0.535 0.235
9  0.00483      8    0.1056  0.538 0.234
10 0.00477      9    0.1008  0.538 0.234
11 0.00158     10    0.0960  0.538 0.235
12 0.00100     11    0.0944  0.540 0.235
```

Figura 74. Resultados del modelo de árbol de regresión sin podar

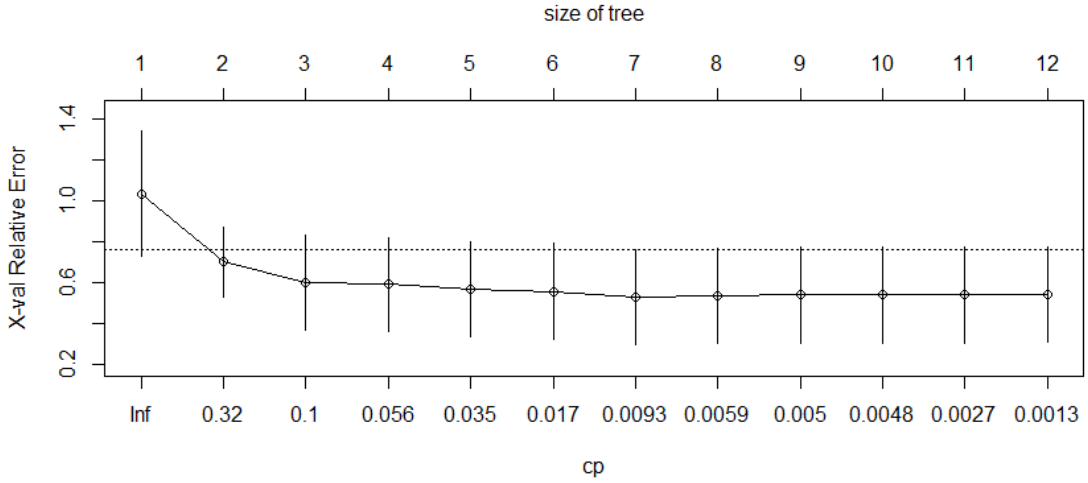

Figura 75. Gráfico de complejidad del modelo de árbol de regresión

La Figura 76 muestra el árbol podado, donde las variables Cu, Ca, Cr, St y z son utilizadas como criterio de decisión. Las variables predictoras Cu y Ca, se repiten cuando se compara con las variables más importantes en el modelo PLS.

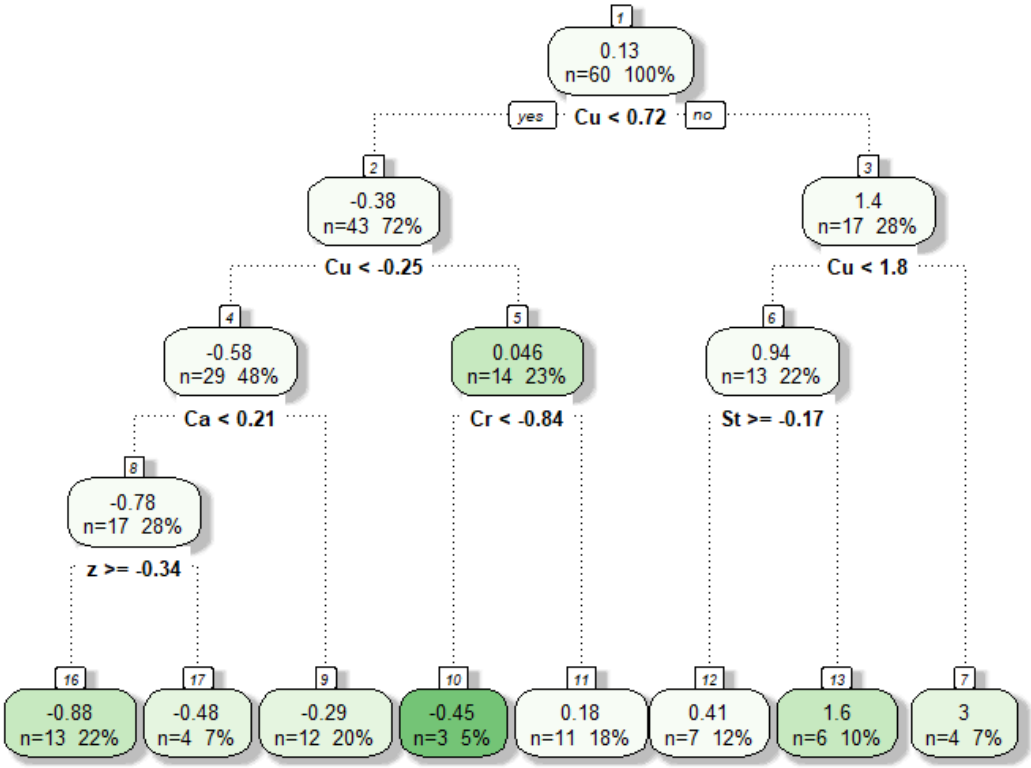

Figura 76. Modelo de árbol de regresión podado

Una vez obtenido el modelo de árbol de regresión, con los datos de entrenamiento, se obtiene una bondad de ajuste (R^2) de **0.89**.

La Figura 77 muestra la predicción sobre los datos de validación con el modelo de árbol de regresión, mostrando una correlación aceptable entre los datos predichos y observados. El error cuadrático medio (MSE) es 0.31, con una bondad de ajuste (R^2) de 0.34.

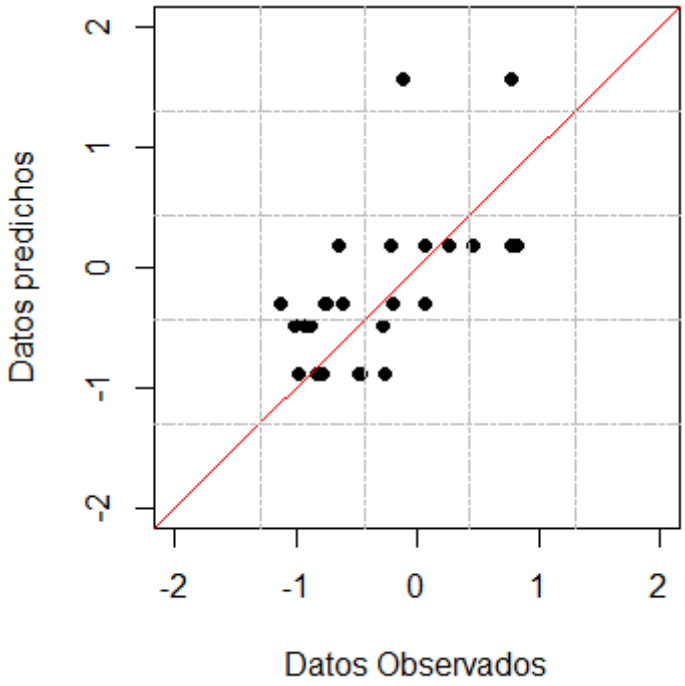

Figura 77. Predicción con los datos de validación con el modelo de árbol de regresión

4.4.1.4 Bosque Aleatorio

Utilizando la función *randomForest*() en R se puede generar un modelo de predicción de bosque aleatorio (*Random Forest*), usando como criterio de parada la generación de un máximo de 500 árboles.

El paquete utilizado en R permite identificar las variables predictoras más influyentes en el modelo. Esto es muy útil para seleccionar variables predictoras con más importancia en el modelo

La Figura 78 muestra la variación del error con el número de árboles. Se observa que a partir de 100 árboles el error del modelo se estabiliza, por tanto, que se haya seleccionado 500 árboles no generará más problema que el tiempo de procesamiento.

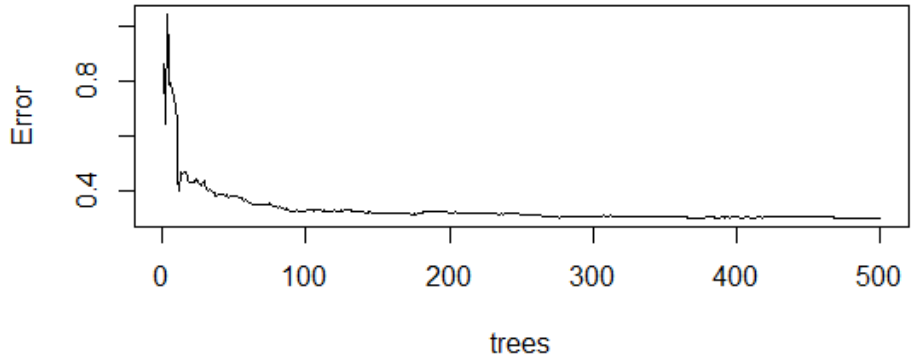

Figura 78. Variación del error con el número de árboles

La Figura 79 muestra que las variables Cu y CuCN son las más importantes en la creación del modelo de la concentración de Cu en el concentrado de flotación con bosque aleatorio . El gráfico de la izquierda (%IncMSE) indica el descenso promedio en la precisión de las predicciones cada vez que se excluye una variable. El gráfico de la derecha (IncNodePurity) indica cuanto aumenta el error cuadrático media cuando aquella variable es permutada aleatoriamente.

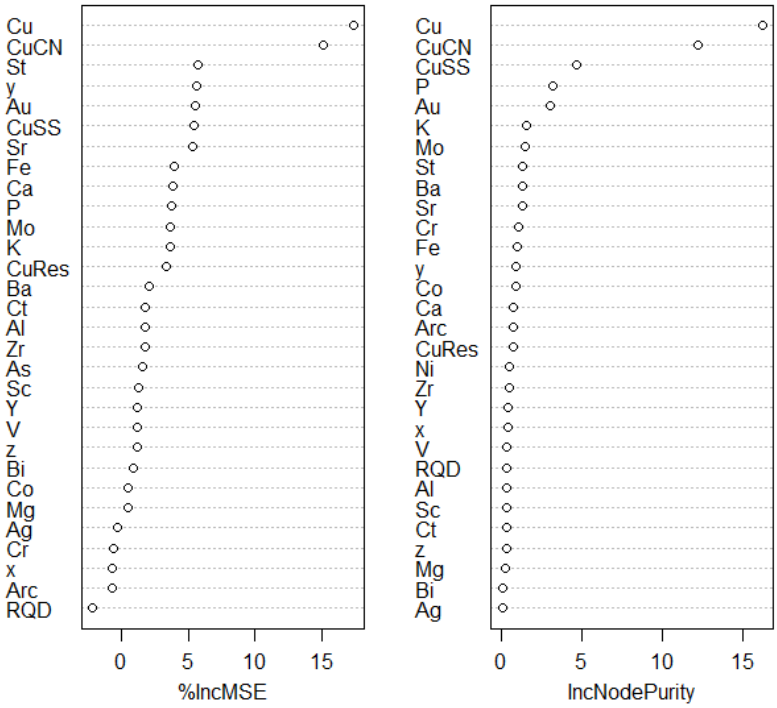

Figura 79. Medida de importancia de las variables

Una vez obtenido el modelo de regresión bosque aleatorio, con los datos de entrenamiento, se obtiene una bondad de ajuste (R^2) de 0.96.

La Figura 80 muestra la predicción sobre los datos de validación con el modelo de bosque aleatorio. Aunque la capacidad predictiva del modelo es muy buena, el modelo es difícil de interpretar. Las variables predictoras Cu y CuCN, se repiten cuando se compara con las variables más importantes en el modelo PLS. El error cuadrático medio (MSE) es 0.23, con una bondad de ajuste (R^2) de 0.39.

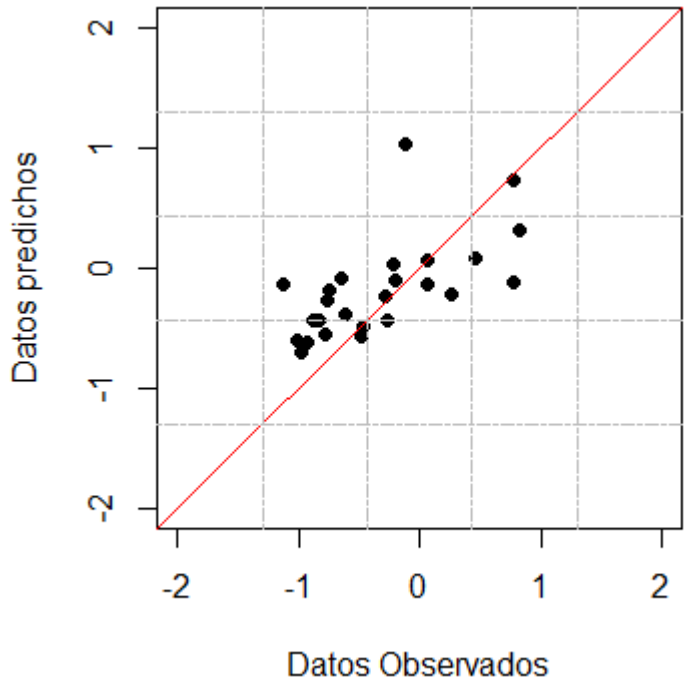

Figura 80. Predicción del modelo de bosque aleatorio

4.4.1.5 Bosque Aleatorio Depurado

El modelo de bosque aleatorio indica que las variables más importantes en el modelo son el Cu y CuCN. Un nuevo modelo de bosque aleatorio que sólo incluya estas variables es creado.

Una vez obtenido el modelo de bosque aleatorio depurado, con los datos de entrenamiento, se obtiene una bondad de ajuste (R^2) de 0.90.

La Figura 81 muestra la predicción sobre los datos de validación con el modelo de bosque aleatorio depurado. El error cuadrático medio (MSE) es 0.16, con una bondad de ajuste (R^2) de 0.51.

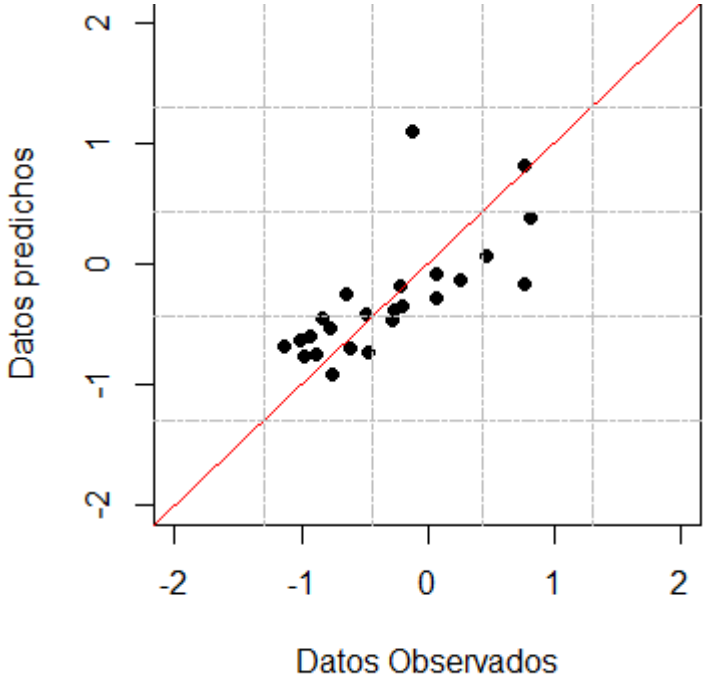

Figura 81. Predicción con los datos de validación con el modelo bosque aleatorio depurado

4.4.1.6 Máquinas de Soporte Vectorial

Se ha utilizado la función svm de R de la librería e1071, que reconoce automáticamente la variable respuesta numérica y realiza una regresión. Tres tipos de kernel fueron utilizados: el lineal, polinómico y radial, variando el parámetro de restricción de la función de *Lagrange cost* (C) en todos los casos. La Figura 82 muestra los resultados con diferentes kernel, comparando los errores de predicción con los datos de validación, siendo el kernel lineal con el que se obtienen menores errores cuadráticos medios (MSE).

```
Call:
svm(formula = CuGra ~ ., data = t.data, kernel = "linear", cost = 0.1)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  linear
       cost:  0.1
      gamma:  0.032258
    epsilon:  0.1


Number of Support Vectors:  44
Call:
svm(formula = CuGra ~ ., data = t.data, kernel = "polynomial",
    cost = 10)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  polynomial
       cost:  10
     degree:  3
      gamma:  0.032258
     coef.0:  0
    epsilon:  0.1


Number of Support Vectors:  51

Call:
svm(formula = CuGra ~ ., data = t.data, kernel = "radial", cost = 10)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  10
      gamma:  0.032258
    epsilon:  0.1


Number of Support Vectors:  47
```

Figura 82. Resultados SVR con diferentes kernel

Una vez obtenido el modelo de máquinas de soporte vectorial utilizando el kernel lineal, con los datos de entrenamiento, se obtiene una bondad de ajuste (R^2) de **0.93**.

La Figura 83 muestra la predicción sobre los datos de validación con el modelo de máquinas de soporte vectorial utilizando el kernel lineal. El error cuadrático medio (MSE) es 0.34, con una bondad de ajuste (R^2) de 0.27.

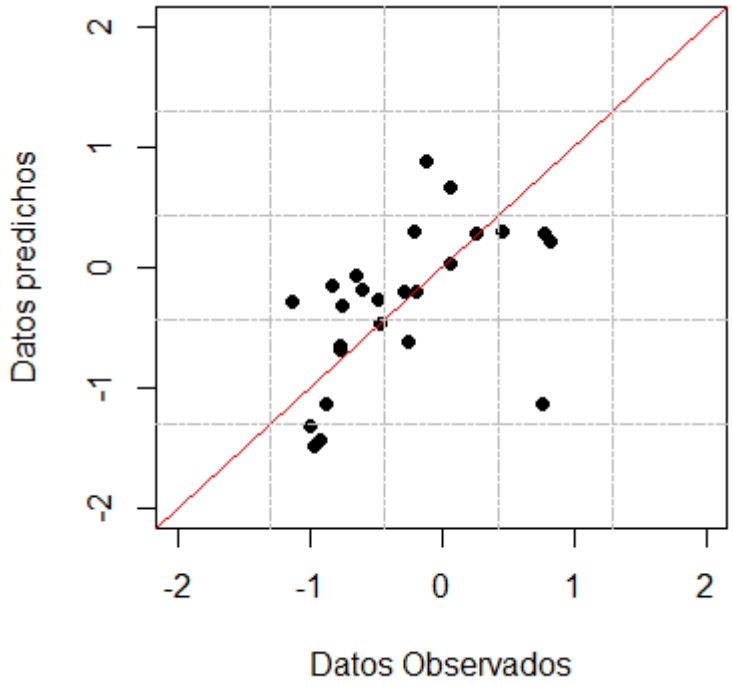

Figura 83. Predicción con los datos de validación con el modelo SVM

Los modelos de máquinas de soporte vectorial no indican las variables importantes. Utiliza un algoritmo donde todas variables intervienen haciendo difícil su interpretación. No se puede comparar con el modelo PLS.

4.4.1.7 Comparación de Modelos

La Tabla 10 compara la bondad de ajuste (R^2) de todos los modelos creados para la predicción de la variable respuesta "CuGra", con los datos de entrenamiento (*train*) y validación (*val*). El modelo de regresión en mínimos cuadrados parciales resulta ser el modelo con mayor capacidad de predicción ante nuevas observaciones.

Tabla 10. Resumen de modelos

ID	Modelo	R^2 (*train*)	R^2 (*val*)
1	Regresión en mínimos cuadrados parciales	0.89	0.60
2	Regresión con componentes principales	0.95	0.26
3	Árbol de regresión	0.89	0.34
4	Bosque aleatorio	0.96	0.39
5	Bosque aleatorio depurado	0.90	0.51
6	Máquinas de soporte vectorial	0.93	0.27

Una gran ventaja de utilizar modelo de regresión en mínimos cuadrados parciales es su sencilla interpretación con los coeficientes de regresión de las variables más importantes. El modelo de regresión en componentes principales requiere muchas componentes, además, para estimar los coeficientes de regresión de las variables originales se requieren de cálculos adicionales, lo cual hace complicado su interpretación. El modelo de árbol de regresión es útil para identificar las variables que hacen posible la división árbol, sin embargo, no se obtiene una alta capacidad de predicción. El modelo de bosque aleatorio permite identificar variables con mayor importancia en el modelo, los cuales son comparados con el modelo PLS, coincidiendo en que las variables Cu y CuCN son importantes en el modelo. El modelo de máquina de soporte vectorial termina resultando una caja negra donde utiliza todas las variables para la predicción, siendo poco útil su interpretación.

Desde el punto de vista geológico y metalúrgico, las variables Cu, CuCN y CuRes, son aquellas que están presenten en los minerales de cobre, mientras más concentración estén presente en las muestras, mayor será la variable respuesta “CuGra”. La variable Fe, está representado por las impurezas que no permiten que sea exitoso el proceso de flotación, el cual es demostrado por el signo negativo en el coeficiente de regresión. Se puede intuir que la variable Sr, está presente en algún mineral que dificulta el proceso de flotación. La variable Ca está asociada a la estructura cristalina de los minerales sulfurosos de cobre, se corrobora que los gráficos w*, c. La variable Mo aparece como un subproducto asociado al tipo de yacimiento en la cual se encuentra los minerales de cobre.

4.4.2 Predicción de Otras Variables Respuestas

La metodología empleada para la predicción de la variable “CuGra” será aplicada para la predicción de variables respuestas adicionales.

Variables repuestas:

- MassRo
- AuGra
- AuRmax
- CuRmax
- Aukinetic
- Cukinetic

Modelos evaluados:

- Regresión en mínimos cuadrados parciales
- Regresión con componentes principales
- Árbol de regresión
- Bosque aleatorio
- Bosque aleatorio depurado
- Máquinas de soporte vectorial

Los mismos datos de entrenamiento son utilizados en la construcción de los modelos. Los mismos datos de validación son utilizados para comparar las predicciones. Los datos a utilizar han sido previamente depurados en el análisis exploratorio de datos (sección 4.2).

Los modelos son entrenados con el 70% de los datos seleccionados aleatoriamente, dejando el 30% restante de los datos para la validación.

En el modelo de regresión en mínimos cuadrados parciales es validado con los límites de control (al 95% y 99%) con un determinado número de componentes. La relación interna del modelo es comprobada. Se selecciona aquellas variables cuyo intervalo de confianza *Jackknife* (95%) no contienen al cero. Se realiza este proceso iterativo hasta llegar a un modelo PLS depurado y validado.

En el modelo de regresión con componentes principales es realizada en dos pasos: el primero, donde se extraen las componentes del bloque de matriz **X** y el segundo, donde se realiza la regresión lineal múltiple de la variable respuesta con todas las componentes extraídas. Se realiza la regresión *stepwise* para detectar las componentes estadísticamente significativas en el modelo basadas en el criterio de información de Akaike (AIC).

En el modelo de árbol de regresión, todos los árboles considerados para llegar al modelo final son analizado con el gráfico de complejidad, seleccionando el tamaño de árbol. Se realiza la poda del árbol para obtener un modelo sencillo.

En el modelo de boque aleatorio se determina las variables más importantes utilizando los gráficos %IncMSE y IncNodePurity. Posteriormente se construye un modelo de bosque aleatorio depurado considerando solo las variables más importantes detectadas anteriormente.

El modelo con máquinas de soporte vectorial es construido con los tres tipos de kernel: el lineal, polinómico y radial, seleccionado el kernel con menor errores cuadráticos medios (MSE).

Las variables importantes detectadas de los modelos son interpretadas por el especialista en geología y metalurgia.

Capítulo 5

5 Conclusiones

Al realizar el estudio del proyecto metalúrgico mediante análisis multivariante y aprendizaje automático se obtuvieron las siguientes conclusiones:

1. Se ha demostrado que la metodología empleada otorga una aceptable capacidad de predicción (R^2 = 0.60) de la variable respuesta “CuGra” (concentración de cobre en el concentrado de flotación) con el modelo PLS.
2. Una gran ventaja del modelo de regresión en mínimos cuadrados en la predicción de “CuGra”, es la capacidad de interpretación del modelo PLS a través de las variables del modelo de regresión. Las variables Cu, CuCN y CuRes, se refieren a los minerales de cobre. La variable Fe y Sr indica las impurezas del mineral que dificultan la eficiencia del proceso de flotación. La variable Ca está asociada a los minerales que acompañan a los sulfuros. La variable Mo aparece asociado al tipo de yacimiento en la cual se encuentra los minerales de cobre.
3. La variable respuesta “MassRo” (recuperación de espumas en el concentrado flotación) tiene una aceptable capacidad de predicción (R^2 = 0.60) con el modelo de máquina de soporte vectorial utilizando un kernel lineal.
4. El modelo de regresión con máquinas de soporte vectorial, a pesar de utilizar un kernel lineal, es muy complejo de interpretar

5. La variable respuesta “AuGra” (concentración de oro en el concentrado de flotación) tiene una baja capacidad de predicción (R^2 = 0.44) con el modelo de regresión en mínimos cuadrados.
6. Como se comenta en el punto 2, el modelo PLS permite una interpretación sencilla del modelo a través de la selección de variables importantes. Las variables Au y Ag se refieren a los minerales auríferos que incrementan la concentración de oro en el concentrado de flotación. La variable St indica las impurezas del mineral que dificultan la colección de minerales de oro.
7. La adquisición de los datos metalúrgicos en el yacimiento geológico es un proceso muy costoso, por tanto, es importante e indispensable tener un protocolo de selección de muestras, con el objetivo que las muestras sean lo más representativas posible del yacimiento.
8. La importancia de esta metodología aplicada radica en la contribución hacía el desarrollo de un modelo que permita pronosticar variables como, por ejemplo, la concentración de cobre y oro en el concentrado de flotación, los cuales se obtendrán en cada bloque del modelo de bloques.
9. La adición de estas respuestas metalúrgicas en el bloque del modelo de bloques permite la optimización en la ingeniería del diseño del circuito de flotación.
10. La implementación de respuestas metalúrgicas en el modelo de bloques permite comprender la variabilidad de una variable respuesta dentro del yacimiento geológico.
11. Con la base de datos del modelo de bloques del proyecto se ha implementado el modelo PLS de la concentración de cobre en el concentrado de flotación, el cual es mostrado en diferentes vistas en la Figura 84, donde se aprecia que los valores más altos son obtenidos en el centro del yacimiento.

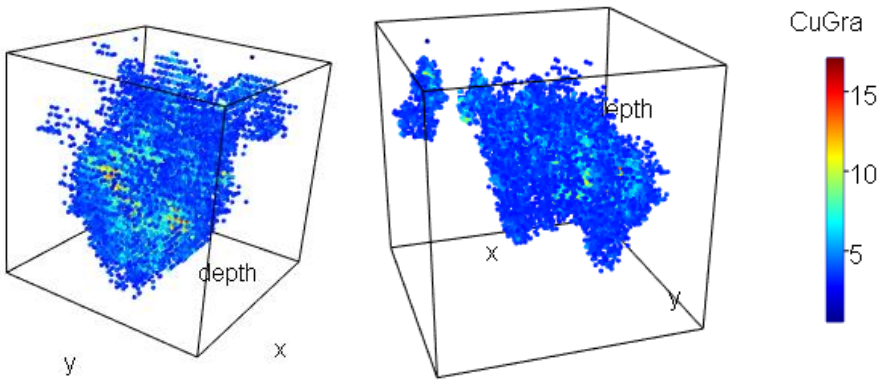

Figura 84. Modelo implementado en el yacimiento

12. La Tabla 11 muestra todos los modelos aplicados sobre las variables respuestas en la metodología propuesta. La Tabla 12 muestra la capacidad de predicción de cada uno de los modelos utilizando los datos de validación. Las respuestas metalúrgicas como AuRmax (recuperación máxima de oro), CuRmax (recuperación máxima de cobre), AuKinetic (velocidad cinética de flotación oro) y CuKinetic (velocidad cinética de flotación cobre) muestran una baja o nula capacidad de predicción con los seis modelos empleados.

Tabla 11. Resumen de MSE para todos los modelos

ID	Modelo
1	Regresión en mínimos cuadrados parciales
2	Regresión con componentes principales
3	Árbol de regresión
4	Bosque aleatorio
5	Bosque aleatorio depurado
6	Máquinas de soporte vectorial

Tabla 12. Resumen de R^2 (predicción) para todos los modelos

Respuesta	Modelos					
	1	2	3	4	5	6
MassRo	0.32	0.00	0.06	0.49	0.39	**0.60**
AuGra	**0.44**	0.26	0.00	0.53	**0.63**	0.00
CuGra	**0.60**	0.26	0.34	0.39	0.51	0.27
AuRmax	0.05	0.02	0.02	0.31	0.29	0.00
CuRmax	0.26	0.24	0.04	0.31	0.36	0.02
Aukinetic	0.04	0.00	0.03	0.17	0.00	0.16
Cukinetic	0.00	0.03	0.01	0.00	0.02	0.02

13. El modelo PLS2 que intenta predecir todas las variables respuestas con un modelo de regresión en mínimos cuadrados parciales, presenta desventajas en la selección de variables. Algunas variables predictoras pueden influir mucho en la predicción de al menos una variable respuesta, sin embargo, no es influyente en el modelo global, y por tal motivo, podría ser descartado del modelo.
14. Las técnicas de aprendizaje automático muestran ser una herramienta poco poderosa para la predicción. Además, exigen que todas las variables de la matriz **X**, utilizadas para crear los modelos, estén presenten al momento de realizar la predicción. En el ámbito geológico, obtener una variable en el modelo de bloques, donde se realizará la predicción, supone un elevado coste y mucho tiempo por cada variable interpolada. Debido a esto, a pesar de que los modelos obtenidos a partir de técnicas de aprendizaje automático logran altas capacidades de predicción, su utilización en el modelo de bloques supone un alto coste, que pocas empresas están dispuestas a cubrir.
15. La predicción realizada con el 30% de los datos seleccionados aleatoriamente, como conjunto de validación, es utilizada para confirmar la capacidad de predicción del modelo. Por ejemplo, los modelos de máquina bosque aleatorio tienen una alta capacidad para ajustar datos de entrenamiento en el modelo de “CuGra”, sin embargo, al predecir nuevos conjuntos de datos de validación se confirma que su capacidad predictiva se reduce.

6 Bibliografía

Abdi, H. (2003). Partial Least Squares (PLS) Regression. En *Encyclopedia for research methods for the social sciences* (págs. 792-795). SAGE.

Abdi, H., & Williams, L. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews*, 433-459.

Adler , D., Murdoch , D., & Nenadic, O. (28 de Marzo de 2018). *rgl: 3D Visualization Using OpenGL. R package version 0.99.16.* Obtenido de https://cran.r-project.org/web/packages/rgl/rgl.pdf

Betancourt, G. (2005). Las Máquinas de Soporte Vectorial (SVMs) . *Scientia et Technica*, 67-72.

Breiman, L. (2001). Random Forests. *Machine Learning*, 1-33.

Bulatovic, S. M. (2007). *Handbook of Flotation Reagents: Chemistry, Theory and Practice Volumen 1.* Elsevier.

Chollet, F., & Allaire, J. (2017). *Deep Learning with R.* Greenwich: Manning.

Dominy, S., O'Connor, L., & Xie, Y. (2016). Sampling and Testwork Protocol Development for Geometallurgical Characterisation of a Sheeted Vein Gold Deposit. *GeoMet*, 31.

Dunn, K. (2018). *Process Improvement Using Data.*

Gujarati, D., & Porter, D. (2009). *Basic Econometrics 5th Edition.* McGraw-Hill Irwin.

Gupta, A., & Yan, D. (2016). *Mineral Processing Design and Operations.* Elsevier.

Höskuldsson, A. (1988). PLS regression methods. *Journal of chemometrics, 2(3)*, 211-228.

Ismartini, P., Sunaryo, S., & Setiawan, S. (2010). The Jackknife Interval Estimation of Parametersin Partial Least Squares Regression Modelfor Poverty Data Analysis. *The Journal for Technology and Science*, 6.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, 1-20.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 1-18.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 18-22. Obtenido de https://CRAN.R-project.org/doc/Rnews/

MacGregor, J., & Kourtl, T. (1995). Statistical Process Control of Multivariate Processes. *Elsevier* , 12.

Matloff, N. (2011). *The art of R programming: A tour of statistical software design.* NO STARCH PRESS.

Mevik, B.-H., Wehrens, R., Liland, K., & Hiemstra, P. (2016). *pls: Partial Least Squares and Principal Component Regression. R package version 2.6-0.* Obtenido de https://CRAN.R-project.org/package=pls

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., & Lin, C.-C. (2017). *Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8.* Obtenido de https://CRAN.R-project.org/package=e1071

Oyarzun, R. (2011). *Introducción a la Geología de Minas.* GEMM.

Rossi, M., & Deutsch, C. V. (2014). Geological Controls and Block Modeling. En *Mineral Resource Estimation* (págs. 29-50). Netherlands: Springer.

Sanchez, G. (12 de Noviembre de 2012). *plsdepot: Partial Least Squares (PLS) Data Analysis Methods. R package version 0.1.17.* Obtenido de https://cran.r-project.org/web/packages/plsdepot/plsdepot.pdf

Servicio Geológico Mexicano. (22 de 03 de 2017). *Portal Gobierno Mexicano.* Recuperado el 02 de 08 de 2018, de https://www.sgm.gob.mx

Slišković, D. (2012). Multivariate Statistical Process Monitoring. En *Tehnicki Vjesnik-Technical Gazette* (págs. 33-41).

Therneau, T., Atkinson, B., & Ripley, B. (23 de Febrero de 2018). *rpart: Recursive Partitioning and Regression Trees. R package version 4.1-13.* Obtenido de https://cran.r-project.org/web/packages/rpart/rpart.pdf

Tukey, J. (1977). *Exploratory Data Analysis.* Pearson.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* New York: Springer.

Williams, G. (2011). Decision Trees. En G. Williams, *Data Mining with Rattle and R* (págs. 205-244). New York: Springer.

Wills, B., & Napier-Munn, T. (2006). *Mineral Processing Technology: An introduction to the practical aspects of ore treatment and mineral recovery.* Amsterdam ; Boston : Oxford, U.K.: Elsevier Science.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal Component Analysis. *Elsevier Science Publishers*, 37-52.

Wold, S., Sjöström, M., & Eriksson, L. (2011). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 109-130.

Wonik, T., & Olea, R. (2007). Environmental Geology. In K. Knödel, G. Lange, & H.-J. Voigt. Berlin: Springer.

7 Anexos

7.1 Análisis Univariante Clásico

Tabla 13. Análisis descriptivo univariante

Variable	Mínimo	Máximo	Promedio	SD	Coef. Asimétria	Curtosis	P25	P50	P75	IQR
x	2804.62	5066.83	3797.52	636.64	0.17	-0.99	3366.68	3761.72	4339.23	972.55
y	3690.56	4200.30	3933.64	134.75	0.01	-0.95	3832.75	3942.55	4045.60	212.85
z	248	735	534	89	0	0	467	529	596	129
RQD	0.00	0.82	0.14	0.19	1.61	2.29	0.00	0.06	0.23	0.23
Ag	0.60	3.00	0.74	0.37	3.76	17.42	0.60	0.60	0.60	0.00
Au	0.01	0.31	0.07	0.07	2.01	3.91	0.03	0.05	0.08	0.05
Cu	0.15	1.56	0.49	0.30	1.47	2.22	0.26	0.39	0.63	0.37
CuSS	0.01	0.60	0.10	0.09	2.68	11.88	0.04	0.09	0.13	0.09
CuCN	0.01	1.16	0.26	0.25	1.42	1.88	0.07	0.17	0.40	0.33
CuRes	0.00	0.46	0.11	0.10	1.68	2.62	0.03	0.08	0.14	0.11
Fe	1.35	7.90	4.15	1.31	0.79	0.82	3.32	3.91	4.72	1.41
St	0.08	7.41	2.52	1.33	0.91	1.51	1.71	2.37	3.16	1.45
Ct	0.02	1.09	0.09	0.14	5.49	34.49	0.05	0.06	0.08	0.03
Arc	0.15	1.45	0.76	0.31	0.16	-0.48	0.55	0.75	0.95	0.40
Al	0.40	4.02	1.55	0.67	0.79	1.54	1.10	1.60	1.89	0.79
Ca	0.01	1.45	0.16	0.20	4.31	23.72	0.05	0.12	0.20	0.15
K	0.04	1.08	0.41	0.25	0.47	-0.62	0.19	0.39	0.60	0.41
Mg	0.00	2.65	0.92	0.56	0.19	-0.10	0.44	0.97	1.29	0.84
As	5.00	996.00	22.81	113.97	7.75	63.14	5.00	5.00	5.00	0.00
Ba	1.00	273.00	61.76	50.67	1.94	5.13	26.34	50.50	80.97	54.62
Bi	5.00	44.00	6.81	5.66	5.43	31.96	5.00	5.00	6.12	1.12
Co	1.00	38.00	14.18	8.00	0.69	0.25	9.37	12.00	19.39	10.02
Cr	9.00	195.00	41.43	22.82	3.67	22.57	28.78	40.00	49.00	20.22
Mo	5.00	126.00	23.80	25.20	2.46	6.65	5.00	18.78	27.75	22.75
Ni	0.10	71.00	5.39	8.53	5.67	40.10	2.00	2.80	5.00	3.00
P	36.00	1650.00	468.16	286.71	1.72	4.24	293.00	444.33	562.25	269.25
Sc	0.87	16.00	4.45	3.01	1.89	4.85	2.14	4.00	5.43	3.29
Sr	3.13	133.00	31.37	30.01	1.71	2.60	9.16	22.00	38.00	28.84
V	1.15	128.00	40.32	28.54	0.78	0.54	17.25	36.01	58.51	41.26
Y	1.37	80.00	13.65	12.69	2.62	9.16	6.00	8.86	17.00	11.00
Zr	0.86	5.00	2.12	1.04	0.62	-0.48	1.00	2.00	3.00	2.00
MassRo	3.91	49.37	14.52	7.15	2.15	6.47	10.35	12.89	16.14	5.78
AuGra	0.03	2.16	0.46	0.43	2.04	4.32	0.18	0.33	0.53	0.35
CuGra	0.28	15.10	3.38	2.60	1.85	4.66	1.54	2.82	4.28	2.74
AuRmax	39.32	94.57	69.30	11.26	-0.69	0.82	64.45	70.02	76.16	11.71
CuRmax	12.00	99.13	84.75	14.62	-3.23	11.86	84.07	89.11	92.17	8.09
Aukinetic	0.16	3.15	1.43	0.65	0.25	-0.09	1.00	1.45	1.85	0.85
Cukinetic	0.18	5.77	1.75	1.26	1.15	0.73	0.87	1.18	2.46	1.59