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

ANÁLISIS MULTIVARIABLE DE ALTERACIONES

TESIS PARA OPTAR AL GRADO DE MAGISTER EN MINERÍA
MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL DE MINAS

ROBERTO JESÚS MIRANDA CONTRERAS

PROFESOR GUÍA:
JULIÁN ORTIZ CABRERA
PROFESOR CO-GUÍA:
WILLY KRACHT JAGARDO

MIEMBROS DE LA COMISIÓN:
XAVIER EMERY
PAULA LARRONDO DURÁN

SANTIAGO DE CHILE
2015

RESUMEN DE TESIS
PARA OPTAR AL GRADO DE
MAGISTER EN MINERÍA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL DE MINAS
POR: ROBERTO JESÚS MIRANDA CONTRERAS
FECHA: 2015
PROF. GUÍA: JULIÁN ORTIZ CABRERA

ANÁLISIS MULTIVARIABLE DE ALTERACIONES

El objetivo general del proyecto es implementar herramientas de análisis multivariable para definir tipos de alteración a partir de concentraciones de distintos elementos químicos, a modo de proponer una metodología cuantitativa de clasificación, siendo esta utilizada para estudiar información de alteraciones presentes en Mina Escondida.

Para la creación de modelos de clasificación se tiene una base de datos que consta de 54 variables geoquímicas más la alteración de la roca mapeada en diecinueve categorias por el equipo de geología, de la cual se estableció que cinco de ellas eran diferenciables. La dificultad del análisis se origina en la gran cantidad de atributos que deben ser estudiados. Para esto se desarrollaron técnicas que ayudan al usuario a entender cómo se asemejan entre ellas las distintas categorías de alteración. Para decidir qué mediciones son las de mayor utilidad, se utilizan técnicas de selección de variables automatizadas con el fin de establecer los criterios de clasificación.

Para realizar el análisis se trabaja con cuatro técnicas de clasificación: mejor variable de clasificación, k-mean clustering, regresión logística y redes neuronales. De los métodos seleccionados el que tiene un mejor desempeño es el de redes neuronales alcanzando alrededor de 77 % de éxito promedio en la clasificación de las cinco categorías.

En la clasificación se logró identificar dos grandes grupos altamente diferenciables entre sí, constituidos por las alteraciones más tardías (argílicas y fílicas) y las más tempranas (clorita-sericita y potásicas), encontrando que los mayores problemas se originan con la ocurrencia de las arcillas, las que dificultan la clasificación de las familias. Los elementos más importantes dentro de la clasificación fueron Mg, Al, Rb, Sc, los cuales se cree que tienen una proveniencia de minerales como la biotita y la clorita principalmente. En resumen se logró crear una metodología que facilita al usuario la clasificación de categorías, siendo esta guiada y semiautomatizada.

Abstract

MULTIVARIATE ANALYSIS OF ALTERATIONS

The overall objective of the project is the implementation of multivariate analysis tools, to define the alteration type using the concentrations of the different elements. The main idea is to propose a quantitative classification methodology , in order to be used to the study Escondida Mine databases.

For the creation of these classification models, we have a database consisting of 54 geochemical variables plus the alteration type of the rock in nineteen categories logged by the geology team of the mine. We established that five of these categories were distinguishable among them. The difficulty in analysis arises from the large number of attributes to be studied. To solve this, techniques were developed to help the users understanding the similarity between the different alterations. To decide which measurements were the most useful, automated variable selection techniques were used in order to establish the criteria for the classification.

To perform the analysis, four classification techniques were used: best classification variable, k-mean clustering, logistic regression and neural networks. The selected method which has the better performance is neural networks, reaching on average about a 77 % success in the classification of the five categories.

During the classification, it was possible to identify two highly distinguishable groups, namely early alteration (chlorite- sericite and potassic) and late alteration (argillic and philic), finding that the greatest problem arises with the occurrence of clays which complicates the classification of families. The most important elements used for the classification were Mg, Al, Rb, Sc, that are believed to have a mineral provenance from biotite and chlorite mainly.

In Summary, it was possible to create a methodology that facilitates to the user the classification of alteration categories, being guided and semiautomated.

A mis padres.

Agradecimientos

Quiero agradecer a mi familia por siempre darme aliento, y estar respaldándome en todas las decisiones que he tomado, a la universidad por la formación y todos los conocimientos que me entregó, a mis amigos con los cuales viví exhaustivas jornadas de estudio y muchas aventuras, Luengo , Rolo, Guti, Chasca, Kenji, Chamo, Marco, Pancho, Bob y muchos mas. , a mi polola Fernanda por estar conmigo durante esta etapa de mi vida, a los profesores de la comisión, especialmente al profesor Julian por haberme dado la posibilidad de realizar el magister, a mis compañeros de ALGES- VIP, Pia , Efra, Mauro, Garrido, Oscar, por la buena onda y las carreras de mario Kart después del almuerzo a los chicos de ALGES Central por siempre estar dispuesto a ayudar y a Minera Escondida por haber financiado el estudio y facilitado las bases de datos que fueron objeto del análisis.

Tabla de Contenido

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El presente proyecto se desarrolla en el marco de colaboración entre BHP Billiton – Base Metals y el Laboratorio ALGES del Departamento de Ingeniería de Minas de la Universidad de Chile como base para el proyecto MQALT. El estudio “Análisis Multivariable de Alteraciones” busca generar una metodología estándar de discriminación de variables categóricas en función de medidas continuas utilizando técnicas de minería de datos como lo son el clustering, regresiones logísticas y redes neuronales entre otras, con el fin de encontrar posibles patrones ocultos en las grandes bases de información como las que se manejan en una mina. En este caso se busca refinar la diferenciación del tipo de alteración hidrotermal mapaeada asociada a la roca para la Mina "La Escondida", utilizando la información obtenida a través de ensayos de agua regia aplicados a muestras obtenidas de los sondajes (información que contiene la concentración de elementos por cada medición de 15 metros aproximadamente) .

Los pórfidos cupríferos destacan en Chile por su ubicación sobre la zona de subducción de la placa oceánica bajo la continental, asociados a diferentes alteraciones hidrotermales. La alteración hidrotermal ocurre a través de la transformación de fases minerales, donde acontece la formación de nuevos minerales, por disolución de minereales preexitentes, la precipitación de nuevos minerales. Lo anterior ocurre a través de reacciones de intercambio iónico entre los minerales constituyentes de una roca y el fluido caliente que circula por la misma. Las asociaciones mineralógicas que se generan se han estudiado y se agrupan bahi distitos tipos de alteración según los tipos de minerales y condiciones fisico-químicas de formación. Cada uno de estos tipos de alteración, posee propiedades típicas de su formación, como por ejemplo tipos de minerales principales, acidez, profundidad, temperatura, entre otros.

En minería durante los últimos años, ha existido un gran avance en las capacidades de generar y recolectar datos de diversas fuentes debido a nuevas metodologías de análisis, al gran poder de procesamiento de ordenadores como también su bajo costo

de almacenamiento, llevando esto a tener grandes cantidades de información disponible. Minera Escondida cuenta con grandes y variadas fuentes de información recolectada a partir de muestras pertenecientes a sondajes, que serán analizadas desde un punto de vista multivariable utilizando técnicas de minería de datos, con el fin de encontrar asociaciones y patrones repetitivos que se produzcan en ellas, y así poder discriminar y ayudar en el modelamiento geo metalúrgico con respecto a la definición de las diferentes unidades de alteraciones que se presentan.

1.1. Problemática

La definición de unidades de estimación geometalúrgicas es una tarea compleja, en la que se deben estudiar las propiedades de los distintos tipos de rocas frente a los diversos procesos a los cuales sean sometidos luego de su extracción, con el fin de tener un control sobre estos como por ejemplo procesos de recuperación, de chancado, etc. Conociendo cómo reaccionará el mineral y sus cualidades, es posible interpretar e inferir estos dominios, por esta razón la caracterizació geológica de los minerales a procesar es una etapa de suma importancia ya que entrega atributos al categoricos semicuantitativos mineral que son base del trabajo posterior. En particular, las alteraciones son relevantes en los procesos y difíciles de caracterizar sistematicamente mediante mapeo, que puede variar de geólogo a geólogo existiendo incluso superposiciones de dos o más alteraciones al mismo tiempo.

1.2. Motivación

Actualmente en minería es mucha la información que se encuentra almacenada, esperando ser análizada de manera integrada. El trabajo de analizar estas grandes bases de datos se encuentra mas allá de las capacidades de una persona, requiriendo de esta manera análisis automatizados, al menos parcialmente, que logren incrementar la eficiencia de los procesos. Por lo anterior, existe la necesidad de establecer metodologías que logren extraer el conocimiento de ellas, de una manera rápida para el investigador. El lograr diferenciar distintos tipos de roca (según alteración, litología, etc.), a partir de información adicional que acompaña a la medición como lo son los análisis de disolución en ácidos (que entregan concentraciones de elementos en la muestra), es una tarea de mucha ayuda, la cual logra restar un poco de subjetividad a la descripción del geólogo, añadiéndo un paso analítico, en donde se estudien las principales variables que definan a una categoría de roca, tratando de normar su descripción, siendo esta información de ayuda en el entendimiento de los procesos

a los cuales se verán sometidas.

En toda gran Mina se manejan datos de geología, la cual es información cualitativa en donde el geólogo entrega información que él puede observar en el sondaje. En zonas de superposición (más de un tipo de alteración) es latente la posibilidad de cometer errores de interpretación, resultando de gran ayuda generar a través de métodos cuantitativos una respuesta que apoye esta información.

Las bases de datos suministrada por Minera Escondida, que contienen gran cantidad de información de sondajes con análisis químicos (agua regia) y descripciones geológicas, las cuales se pretende utilizar y desarrollar en base a ellas, desarrollar herramientas de análisis cuantitativas, que permitan asignar el tipo de alteración a cada muestra. Con el objetivo final de contrastar y mejorar la codificación de los tipos de alteración con información del mapeo.

1.3. Objetivos

El objetivo general y los específicos del proyecto son:

1.3.1. Objetivo general

Implementar herramientas de análisis multivariable para encontrar relaciones entre un tipo de alteración y las concentraciones de distintos elementos químicos, a modo de proponer una metodología cuantitativa de clasificación, que logre asignar su tipo de alteración, para la definición de unidades geometalúrgicas en Mina Escondida.

1.3.2. Objetivos específicos

- Diferenciar distintas alteraciones en zonas de la mina, en sectores donde estas se superpongan.
- Crear herramientas que ayuden al geólogo a asgnar un tipo de alteración a los sondajes, en base a los análisis químicos, para su posterior interpretación en un modelo tridimensional.
  mapear los sondajes y modelar la morfología del yacimiento.
- Encontrar relaciones y patrones repetitivos en la data a través de análisis multivariable.

1.3.3. Alcances

El alcance del proyecto busca entregar un procedimiento general, basado en herramientas de análisis multivariable, que logre inferir una categoría de roca en base a mediciones cuantitativas, pudiendo ser utilizada para distintas minas que deseen estudiar su metalogénesis. El caso de estudio tiene como alcance a mina Escondida, la cual proporciona los datos con los que se trabajó, aplicando la metodología creada.

1.4. Estructura de la tesis

La estructura utilizada en este documento para exponer el trabajo realizado es la siguiente:

- **Capítulo 1. Introducción:** Corresponde a la descripción del tema, la motivación de este, los alcances y objetivos del trabajo realizado.
- **Capítulo 2. Antecedentes:** Corresponde a la revisión bibliográfica o antecedentes. En este capítulo se explican los conceptos necesarios para la comprensión y contextualización del trabajo.
- **Capítulo 3. Metodología:** Capítulo en el cual se realiza una exposición de la metodología completa, indicando los pasos que se desarrollan y cómo se resuelven los distintos problemas.
- **Capítulo 4. Resultados de los modelos:** Se exponen los resultados obtenidos a través de la metodología, detallando cada uno de ellos.
- **Capítulo 5. Análisis de resultados de los modelos:** En este capítulo se analizan los resultados de los modelos realizados, comparando cada uno de ellos, seleccionando el que presente mejor desempeño.
- **Capítulo 6. Análisis de modelo final utilizando RNA:** Corresponde al capítulo donde se realiza un estudio más profundo del modelo seleccionado, identificando las principales falencias y sus soluciones.
- **Capítulo 7. Análisis de resultados modelo final:** En este capítulo se analizan los resultados finales con el modelo mejorado. Estos resultados se estudian poniendo énfasis en el origen de los errores y la distribución de las respuestas. Además se compara el modelo creado con los sondajes mapeados de manera visual.

- **Capítulo 8. Conclusiones:** Se establecen las conclusiones del trabajo realizado y se proponen trabajos a realizar en el futuro.

Capítulo 2

Antecedentes

2.1. Encontrando el conocimiento

Durante los últimos años, en minería, ha existido un gran avance en las capacidades de generar y recolectar datos de diversas fuentes debido al gran poder de procesamiento de máquinas como también su bajo costo de almacenamiento, lo que ha llevado a tener grandes cantidades de información acumulada. A medida que el tiempo transcurre, las bases de datos se van incrementando en dos aspectos:

1. La cantidad N de mediciones que se encuentran en ella.
2. La cantidad d de atributos o variables de la medición.

Con grandes bases de datos, analizar la información es ciertamente irrealizable por humanos. Es por esta razón que los análisis deben ser automatizados o al menos parcialmente, con el objetivo de incrementar la eficiencia del proceso, reduciendo costos y aumentando el conocimiento de comportamiento de la roca. Es de urgente necesidad crear nuevas teorías y herramientas que faciliten a las personas a extraer la información útil (conocimiento) de los grandes volúmenes de información digital que se poseen. Estas teorías y herramientas son objeto de un campo emergente llamado descubrimiento de conocimiento en bases de datos o KDD por sus siglas en inglés (knowledge discovery in databases). Su nombre hace referencia al proceso general de hallar la información útil, diferenciándose del término “minería de datos” al ser éste un paso particular en el proceso, en el cual se aplican algoritmos específicos para extraer los distintos patrones ocultos que están contenidos en las bases de datos [1]. Es posible que a través de la minería de datos se encuentren patrones que parecen ser significantes estadísticamente (incluso en información aleatoria), pero de hecho

no lo son, por lo cual el KDD, toma gran importancia al tratar de entender el conocimiento y legitimarlo, siendo la minería de datos un paso legítimo, entendiendo cómo llevarlo a cabo correctamente.

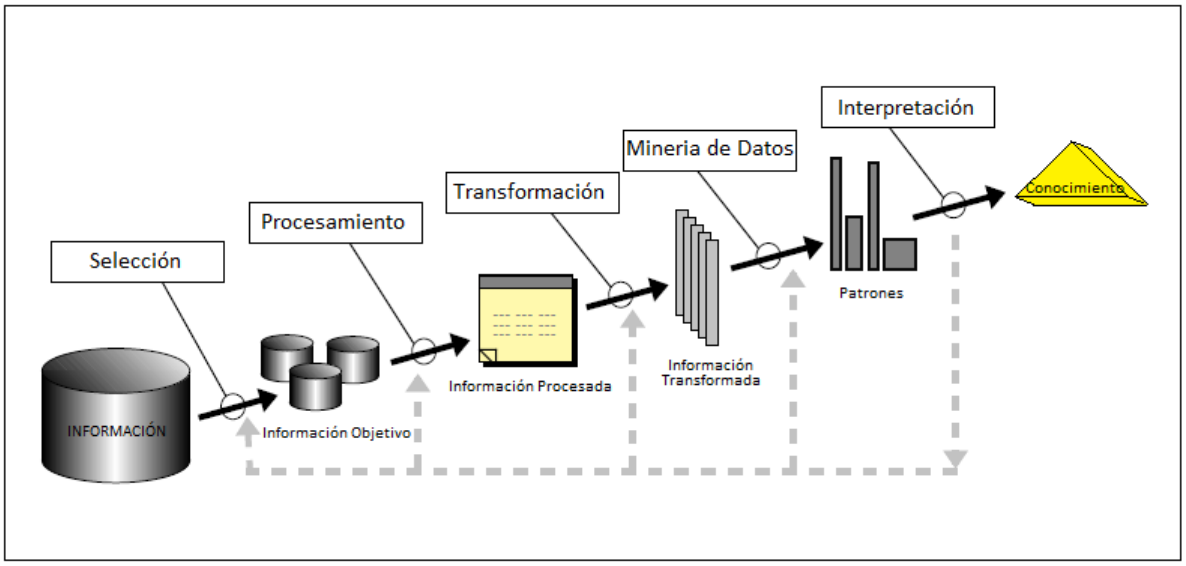


Figura 2.1: Proceso KDD

El KDD es un proceso no trivial de identificación de patrones válidos, nuevos, útiles y entendibles, en donde los datos son realidades y los patrones son expresiones de estos, describiendo subgrupos de información o un modelo que puede ser aplicable. La figura 2.1 muestra una vista general de los distintos pasos que lo conforman, siendo un proceso interactivo e iterativo, incluyendo la evaluación y las posibles interpretaciones de los patrones encontrados para determinar cuáles de estos son considerados como conocimiento innovador. Estos patrones encontrados deben ser validados con nueva información, con el fin de asegurar cierto grado de certeza de los distintos fenómenos que pueden contener estos mismos.

En minería, herramientas de análisis estadísticos que estudian relaciones geometalúrgicas han sido ocupadas con anterioridad [2], vinculando la sericita-gris verde con zonas de alto enriquecimiento, relacionando minerales de ganga con la dureza, entre otras conclusiones para un yacimiento de tipo pórfido cuprífero. Con el fin de darle un contexto geológico y exponer distintas herramientas que se utilizarán para llevar a cabo el estudio, se expondrá a continuación, información relevante que se debe tener en cuenta a la hora de realizar el estudio para analizar los resultados y concluir sobre estos mismos.

2.2. Pórfido cuprífero

Los pórfidos cupríferos son yacimientos de gran volumen de mineralización primaria de sulfuros de cobre-fierro y fierro. Estos yacimientos están asociados a arcos magmáticos de márgenes continentales y a magmatismo calco-alcalino de composición intermedia. La mineralización ocurre de forma diseminada en vetillas, stock works, brechas y como relleno [3].

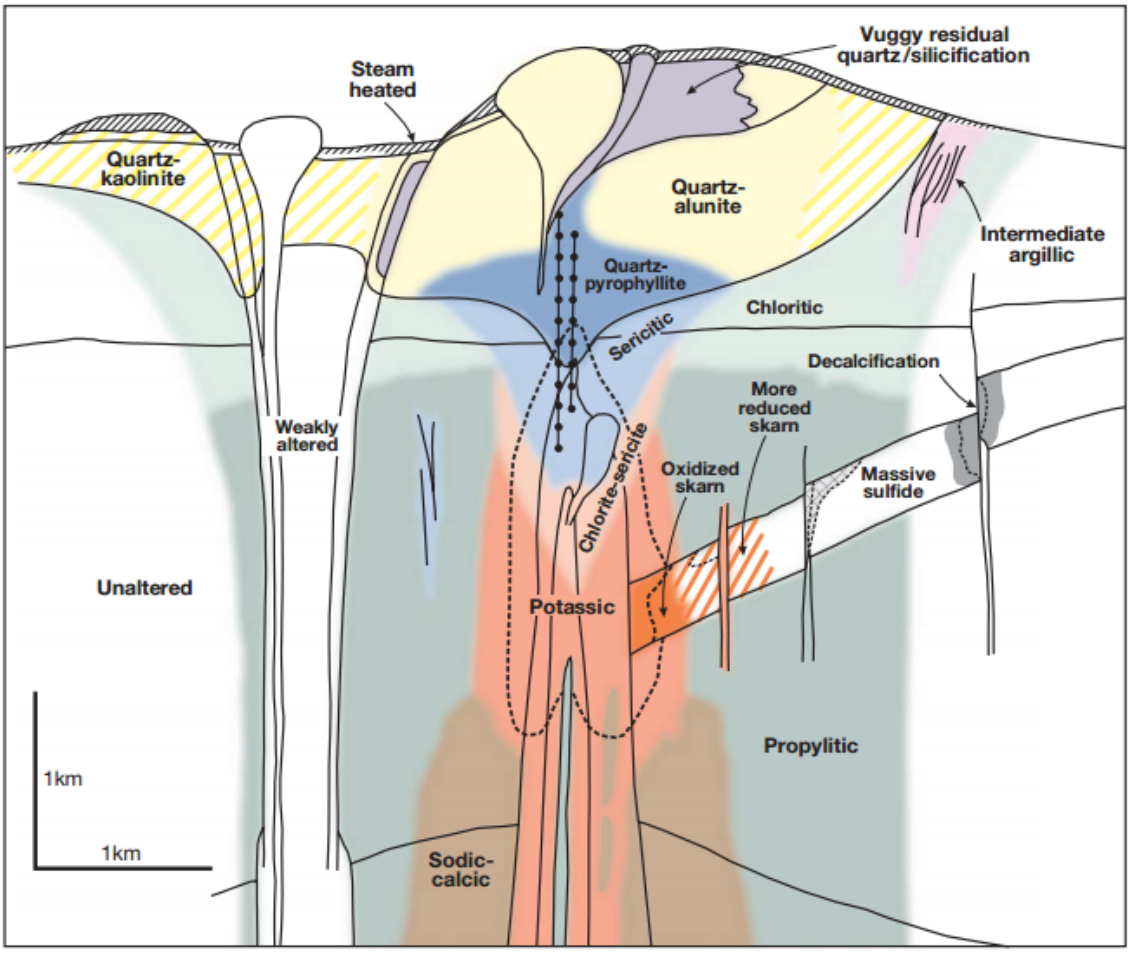


Figura 2.2: Zonación de la mineralización por alteración generalizada, Sillitoe 2010

En la figura 2.2, se puede apreciar las interacciones espaciales de un pórfido centralizado de Cu $\pm$ Mo $\pm$ Au, entre ellas con la roca caja, ubicando las distintas alteraciones hidrotermales que lo componen.

2.2.1. Alteraciones hidrotermales

Los pórfidos cupríferos están asociados a diferentes alteraciones hidrotermales, siendo este, un término general que incluye la respuesta mineralógica, textural y química de las rocas

a un cambio ambiental, en términos químicos y termales, en presencia de agua caliente, vapor y/o gas. La alteración hidrotermal ocurre a través de la transformación de fases minerales en donde acontece el crecimiento de nuevos minerales, disolución/ precipitación de minerales y a través de reacciones de intercambio iónico entre los minerales constituyentes de una roca y el fluido caliente que circula por la misma. Las principales alteraciones hidrotermales presentes en los pórfidos cupríferos son:

Alteración potásica

Caracterizada principalmente por feldespato potásico y/o biotita, con minerales accesorios como cuarzo, magnetita, sericita y clorita. Esta alteración se caracteriza por ocurrir en ambientes de pH neutro a alcalino y a altas temperaturas.

Alteración propilítica

Caracterizada principalmente por la asociación clorita-epidota con o sin albita, calcita, pirita con minerales accesorios como cuarzo, magnetita e illita. Esta alteración se origina en condiciones de pH neutro a alcalino a rangos de temperatura bajo los 200-300°C.

Alteración fílica

Caracterizada principalmente por cuarzo y sericita con minerales accesorios como clorita, illita y pirita, ocurriendo a un rango de pH entre 5 y 6 y a temperaturas sobre los 250°C.

Alteración argílica moderada

Caracterizada principalmente por arcillas (caolín) y mayor o menor cuarzo. Esta alteración ocurre en rangos de pH entre 4 y 5, y puede coexistir la alunita en un rango transicional de pH entre 3 y 4, y a temperaturas bajo los 300°C.

Alteración argílica avanzada

Caracterizada por cuarzo residual (oqueroso) con o sin presencia de alunita, jarosita, caolín, pirofilita y pirita. Ocurre dentro de un amplio rango de temperaturas con condiciones de pH entre 1 y 3.5.

Alteración calcosilicatada (skarn)

Caracterizada por silicatos de calcio y magnesio, dependiendo de la roca huésped.

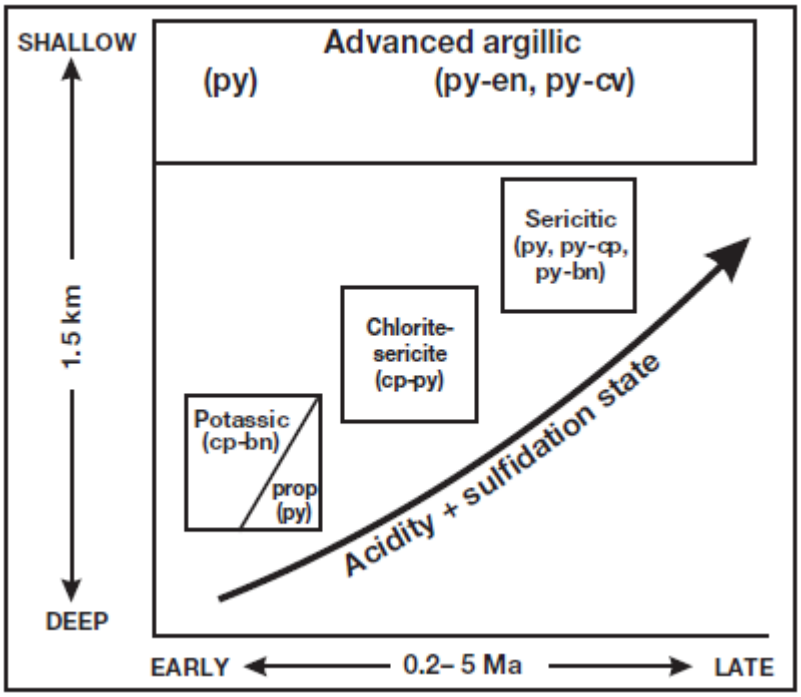


Figura 2.3: Acidez respecto a la profundidad, Sillitoe 2010

Es importante ver cómo estas alteraciones tienen una relación con la profundidad, la temperatura, la acidez y la temporalidad de los eventos. Como se puede ver en la figura 2.3, la alteración potásica se asocia a edades más tempranas de formación, teniendo una mayor profundidad. Alteraciones como las argílicas son más superficiales, siendo estas, alteradas por el fluido hidrotermal, que en estas etapas de formación es rico en ácido sulfhídrico (HS), lo que aumenta la acidez en ellas.

Corbett y Leach (1998) publicaron un diagrama de clasificación de tipo de alteración hidrotermal, en donde se incluyen los principales tipos clásicos de alteración ordenados en función del pH del fluido y de la temperatura (figura 2.4). Estos autores separaron, además, grupos caracterizados por ciertos minerales.

En las etapas formadoras de minerales, pueden ocurrir intercambios iónicos en sus estructuras, creándolos, conteniendo distintos átomos. Estos intercambios se originan debido a potenciales iónicos parecidos entre átomos. En la figura 2.5, se pueden ver grupos de átomos semejantes entre ellos, producto de su carga y tamaño parecidos. Es importante notar que el grupo LFS, LIL (low field strength, large ion lithophile) son más móviles en faces fluidas como lo que ocurre en las alteraciones hidrotermales [4], estando entre ellos el potasio (K), elemento presente en diversos minerales como la biotita y la clorita entre otros.

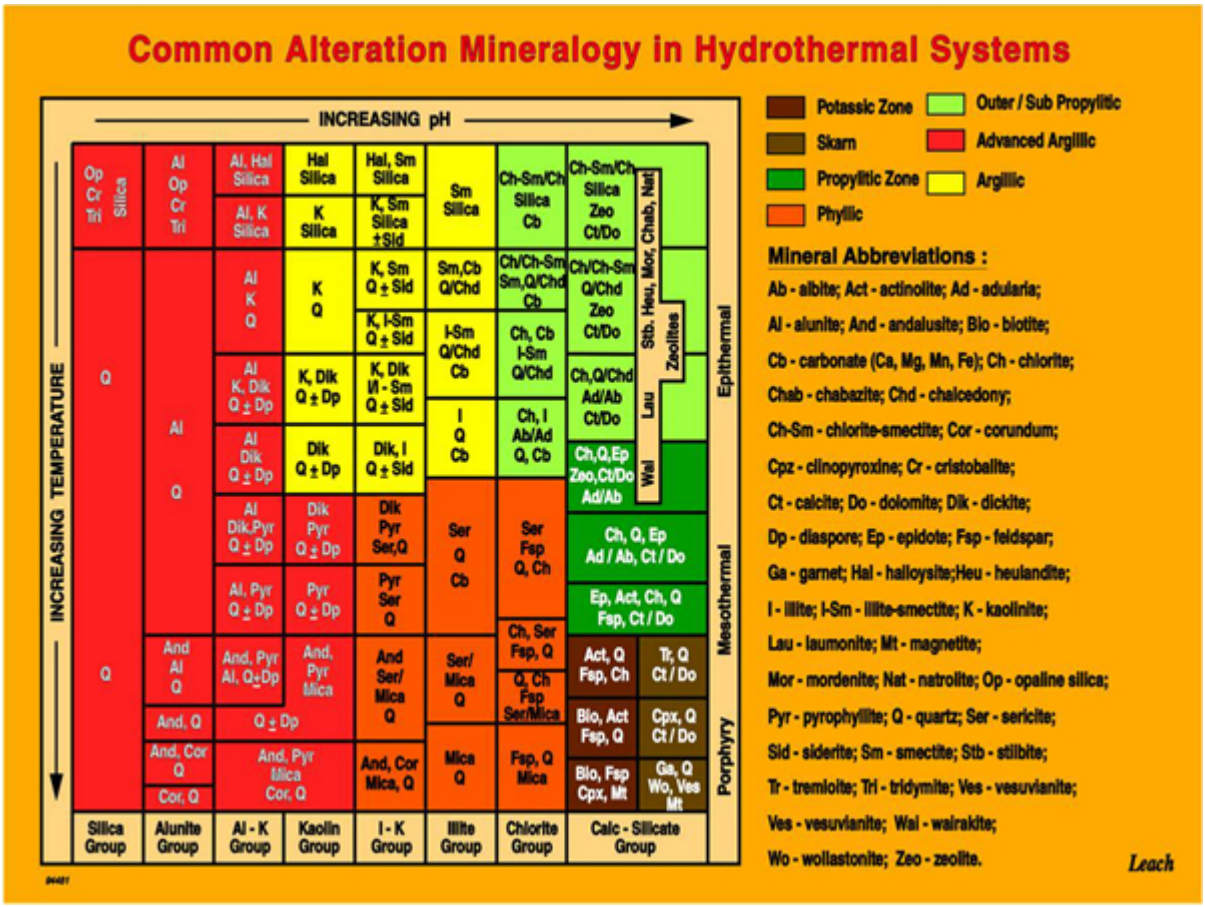


Figura 2.4: Asociaciones de minerales en sistemas hidrotermales, Corbett y leach

2.2.2. Alteración supérgena

La alteración supérgena es un proceso de reequilibrio de la mineralogía hipógena (Hidrotermal) a las condiciones oxidantes cerca de la superficie terrestre (sobre el nivel de las aguas subterráneas). La mayoría de las asociaciones de minerales sulfurados son inestables en estas condiciones y se descomponen (meteorizan) para originar una nueva mineralogía estable en condiciones de meteorización.

El proceso de alteración supérgena de depósitos minerales hidrotermales involucra la liberación de cationes metálicos y aniones sulfato mediante la oxidación de sulfuros hipógenos (lixiviación). Los sulfatos de cobre y plata generados son solubles y son transportados hacia abajo por aguas meteóricas percolantes. Los cationes descienden en solución y pueden ser redepositados por reacción con iones carbonato, silicato, sulfato o sulfuro. Estos elementos pueden formar minerales oxidados que permanecen en la zona oxidada, pero también pueden ser precipitados debajo del nivel de aguas subterráneas por los sulfuros hipógenos, y formar sulfuros de mayor ley respectivamente, siendo este proceso más eficiente para el cobre que para la plata (enriquecimiento secundario).

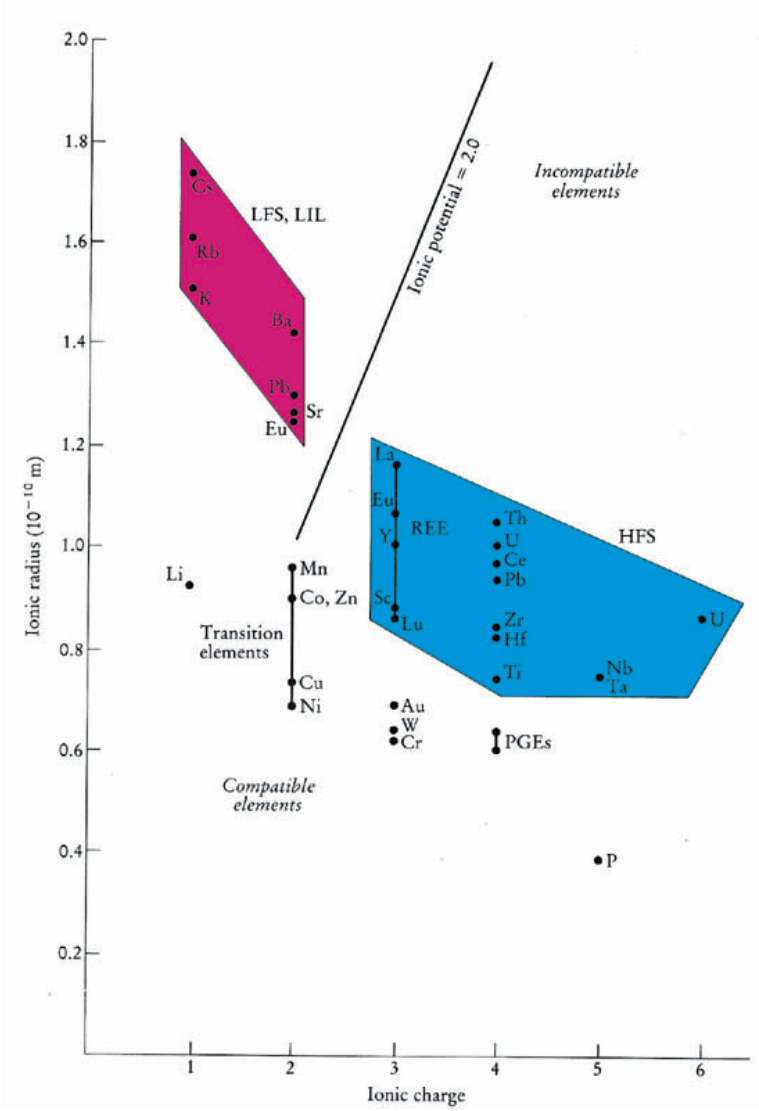


Figura 2.5: Diagrama de alteraciones supergenas

Los procesos supérgenos modifican significativamente la mineralogía de los cuerpos mineralizados de origen hidrotermal y afectan su metalurgia extractiva. Además, pueden producir importantes enriquecimientos secundarios ya sea de cobre, plata u oro. La mayor parte de los pórfidos cupríferos no son económicamente rentables, a menos que hayan desarrollado enriquecimiento secundario o supérgeno. Sin embargo, los procesos supérgenos también pueden resultar en la dispersión de los elementos metálicos o su redeposición como depósitos exóticos a cierta distancia del depósito hipógeno original.

En la zona oxidada los minerales sulfurados hipógenos son destruidos, la estructura y composición química de las menas son modificadas significativamente, teniendo repercusión en la metalurgia extractiva (ya que tienen el interés económico).

En la porción inferior de la zona oxidada que subyace a rocas lixiviadas, se forman nuevos

minerales oxidados por reacción de cationes metálicos en solución con aniones, tales como carbonatos (ej. malaquita) y silicatos (crisocola). En condiciones áridas y salinas como las del desierto de Atacama, los cloruros juegan también un rol importante como la formación de la atacamita [5].

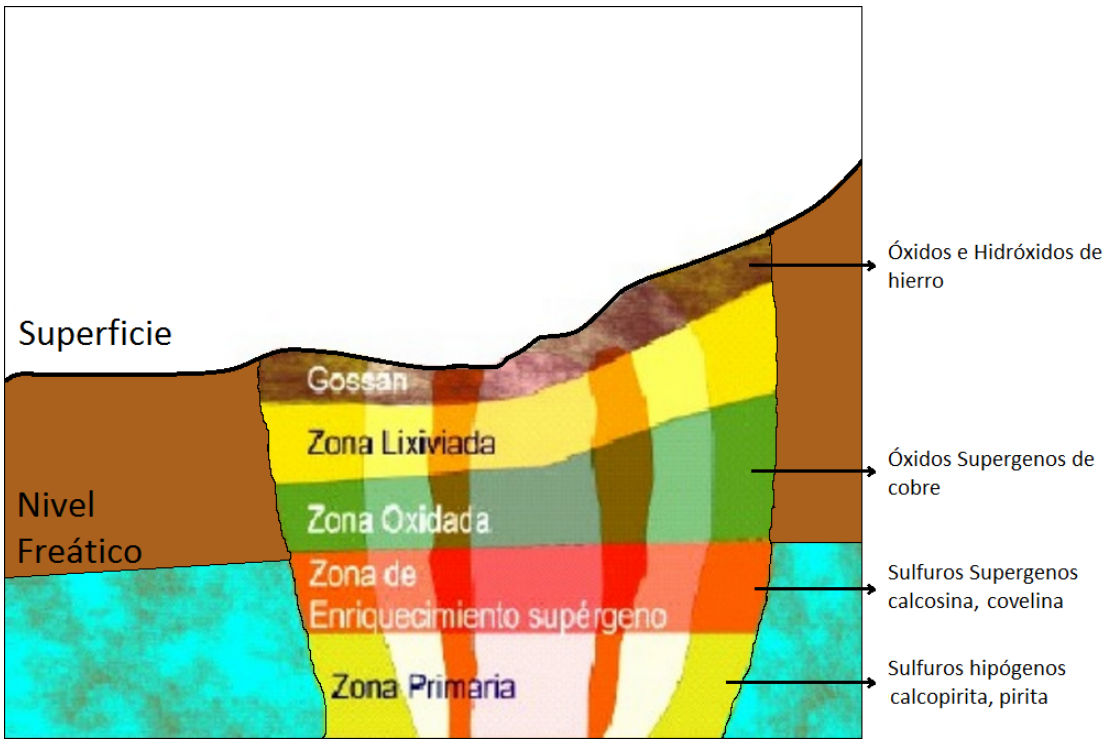


Figura 2.6: Diagrama de alteraciones supergenas

La figura 2.6 muestra la disposición espacial esperada de los distintas alteraciones supérgenas con minerales característicos que se encuentran presentes en un pórfido cuprífero.

2.2.3. Geoquímica de los minerales arcillosos

Las arcillas y micas son de gran importancia dentro del pórfido cuprífero, siendo minerales determinantes a la hora de definir una alteración, interfiriendo en diferentes procesos metalúrgicos como la recuperación en flotación [6]. Estas son, composicionalmente, complejos de silicatos de aluminio hidratado, los cuales estructuralmente son llamados filosilicatos, siendo estructuras en forma de hojas continuas apiladas que comprimen capas de cationes coordinados con grupos O^{2-} y/o OH^{-} . Las capas pueden estar separadas por cationes adsorbidos o relativamente fijos o por moléculas de agua. Las capas que forman las arcillas son descritas como octaédricas y tetraédricas. Las capas octaédricas son llamadas di-octaédricas si ellas contienen dos Al^{3+} con séxtuple coordinación con O^{2-} o OH^{-}, caso en

el que uno de los tres sitios de cationes permanece vacante. Esta capa también es llamada capa gibbsita.

Si la capa octaédrica se origina con tres Mg^{2+} coordinados con seis O^{2-} o OH^{-}, entonces todos los lugares de cationes son ocupados por Mg^{2+}. Las capas son entonces llamadas tri-octaédricas, o capas brucitas.

Las capas tetraédricas son compuestas por cationes con cuádruple coordinación con O^{2-} o OH^{-}. El catión dominante en estos casos es usualmente Si^{4+}, pero también podría ser Al^{3+} [7].

El grupo caolinita y otros filosilicatos de 2 capas

Estas son constituidas por una capa gibbsita octaédrica y una capa de Si^{4+} tetraédrico, con O^{2-} en los vértices del silicio tetraédrico, remplazando el grupo OH^{-} de la capa octaédrica y es compartido por ambas capas. En estas, no hay exceso o deficiencia de cargas en superficie porque está equilibrada en el interior de la estructura cristalina, por lo que son más difíciles de diluir al ser más estables en disoluciones como agua-regia.

Por su alternancia en capas tetraédricas y octaédricas, se dice que los filosilicatos de dos capas tienen una estructura $T : O$.

Filosilicatos de tres capas

Estos incluyen talco y pirofilitas, illita y el grupo de las arcillas esmectita, vermiculita y micas (ej: moscovita, flogopita y biotita), se dice que los filosilicatos de tres capas tienen una estructura $T : O : T$.

En el grupo de la esmectita, tiene cationes inter-capa que son adsorbidos, siendo necesarios para balancear la carga de la red insatisfecha (usualmente negativa) del enrejado del cristal arcilloso causado por substituciones estructurales y vacantes en la capa octaédrica y/o tetraédrica. Por ejemplo Mg^{2+} u otro catión con valencia dos, puede sustituir por Al^{3+} en la capa octaédrica. Al^{3+} o Fe^{3+} puede reemplazar Si^{4+} en la estructura tetrahédrica.

En el grupo de las micas de minerales $T : O : T$, un ion Al^{3+} substituye uno de cada cuatro iones Si^{4+}, en la capa tetraédrica. El exceso de carga negativa es compensado por iones K+ inter-capa que son sostenidos electroestáticamente muy pobremente entre capas tetraédricas adyacentes. Las micas son relativamente duras y elásticas, y los iones de potasio no son fácilmente intercambiables. Las micas incluyen moscovita (di-octaédrica), flogopita (tri-

octaédrica) y biotita, la cual es similar a la flogopita pero con Fe^{2+} en substitución de Mg^{2+} en los mismos sitios octaédricos.

Es importante notar que al ser menos estables electroestáticamente, tienden a ser más fáciles de diluir en disoluciones como agua-regia. En la Tabla 1: estructuras y composiciones idealizadas y típicas de filosilicatos, se puede ver una clasificación de minerales Dioctaédricos y Tri-octraédricos $T:O$ y $T:O:T$.

Tabla 2.1: Estructuras y composiciones idealizadas y típicas de filosilicatos

I DIOCTAHEDRAL (gibbsite-type layers)
Two-layer structures
Kaolinite Group: $Al_2Si_20_5(OH)_4$
Kaolinite, nacrite, dickite & halloysite (hydrated)
Three-layer structures
Pyrophyllite: $AI_2Si_4O_{10}(OH)_2$
Smectite Group:$(Na,Ca)_{0.5_{0.7}}(Mg,Fe,AI)_4(AI,Si)_8O_2(OH)_4$
Montmorillonite: $(Na,Ca_{0.5})_{0.7}(Mg_{0.7}AI_{3.3})Si_80_{20}(OH)_4$
Wyoming type:$(Ca_{0.5}(Mg_{0.5}Fe^{3+}_{0.5}AI_3)(Al_{0.5}Si_{7.5})O_{20}(OH)_4$
Cheto WY type:$(Ca_{0.5}(Mg_{0.5}AI_3)(Al_{0.5}Si_8)O_{20}(OH)_4$
Beidellite: $(Na,Ca_{0.5})_{0.7}Al_4(Al_{0.7}Si_{7.3})o_{20}(OH)_4$
$Ca_{0.5}AI_4(AI,Si_7)O_{20}(OH)_4$
Nontronite: $(Na,Ca_{0.5})_{0.7}(Fe^{3+})_4(Al_{0.7}Si_{7.3})O_{20}(OH)_4$
$(Ca_{0.5})(Fe^{3+}_3AI)(AI,Si_7)O_{20}(OH)_4$
Muscovite: $KAl_2(AlSi_3O_{10})(OH)_2$
Illite Group: $(K_{1.5-1.0})Al_4(Al_{1.5}Si_{6.5-7.0})O_{20}(OH)_4$
Illite (ideal average):$(K_{1.5})Al_4(Al_{1.5}Si_{6.5})O_{20}(OH)_4$
Illite (typical composition):$(K_{1.5})AL_4(Mg_{0.5}Fe_{3+_{0.5}}Al_3)(AI,Si_7)O_{20}(OH)_4$
Ferric illite:$K(Fe^{3+}_3+Al)(Al,Si_7)O_{20}(OH)_4$
II TRIOCTAHEDRAL (brucite-type layers)
Two-layer structures
Serpentine Group:$Mg_3Si_2O_5(OH)_4$
Lizardite, antigorite, & chrysotite (fibrous)
Three-layer structures
Talc: $Mg_3Si_4O_{10}(OH)_2$
Vermiculite Group: $(Mg,Ca)_{0.6-0.9}(Mg,Fe^{3+},A)_{6.0}(AI,Si)_80_{20}(OH)_4$
$(K,Mg_{0.5})(Mg_4Fe^{2+}_{1.5}Fe^{3+}_{0.5})(Al_{1.5}Si_{6.5}O_{20}(OH)_4$
Smectite Group: $(Na,Ca_{0.5})_{0.7}(Mg,Fe,Al)_6(AI,Si)_8O_{20}(OH)_4$
Saponite: $(Na,Ca_{0.5})_{0.8}Mg_6(Al_{0.8}Si-7.2)O_{20}(OH)_4$
$(Ca_{0,4})(Mg_4Fe^{2+}_{1,6}Fe^{3+}_{0,3}Al_{0.1})(AI_{1.2}Si_{6.8})O_{20}(OH)_4$
Hectorite: $(Na,Ca_{0.5})_{0.7}(Li_{0.7}Mg_{5.3})(Si)_80_{20}(OH)_4$
Phlogopite: $KMg_3(AISi_3O_{10})(OH)_2$
Biotite: $K(Mg,Fe)_3(AISi_3O_{10})(OH)_2$
Chlorite Group: $(Mg,Fe^{2+},Fe^{3+},Mn,AI)_{12}(AI,Si)_8O_{20}(OH)_{16}$
Sedimentary:$Mg_5Fe^{2+}_{4.5}Fe^{3+}_{0.5}AI_2(AI_{2.5}Si_{5.5})O_{20}(OH)_{16}$
Note: Cations listed first in curved brackets for the smectites and vermiculites (Na, Ca, K. and Mg) are present as exchangeable interlayer ions. All the smectites and vermiculites (and thus interlayer illite-smectites) have important amounts of interlayer water. the amount of which depends upon the clay and the nature of interlayer cations (cf. Brindley and Brown 1980). As is customary. these waters are left out of the mineral formulae.
Source: Deer et al. (1992) tSlaughter (1992)

2.3. Marco Geológico Escondida

Escondida es un yacimiento del tipo pórfido cuprífero del Eoceno Superior – Oligoceno Inferior, al que también pertenecen importantes depósitos tales como Chuquicamata, Collahuasi y el Salvador, entre otros. Este depósito corresponde específicamente a uno porfídico de cobre (-Molibdeno (Mo), -Oro (Au)) diseminado, existiendo varias fases de intrusión en él, donde las más tempranas tienden a tener las más altas leyes de minerales de mena, existiendo mineralización en la roca caja que lo hospeda [8].

Figura 2.7: Foto satelital distrito Escondida

2.3.1. Litología

Los yacimientos de Escondida y Escondida Norte están asociados genéticamente a un complejo intrusivo de composición monzonítica a granodiorítica, denominado Complejo Intrusivo Feldespático Escondida, que está claramente en contacto de intrusión, con rocas volcánicas andesíticas pertenecientes a la formación Augusta Victoria. A su vez, este complejo, está intruido por cuerpos subvolcánicos tardíos de composición riolítica y por diques dacíticos.

Formación Augusta Victoria

Esta unidad está constituida principalmente por rocas andesíticas y corresponde a la roca caja del complejo subvolcánico mineralizador del yacimiento Escondida.

Complejo Intrusivo Feldespático Escondida (CIFE)

Posee rocas de composición intermedia, variando de monzonitas a granodioritas, cuyas características principales son sus texturas, constituidas principalmente por fenocristales de plagioclasas, los cuales constituyen entre el 40 % y 60 % del volumen total de la roca. Esta unidad se encuentra en contacto por falla e intrusión con el pórfido riolítico y andesitas de la formación Augusta Victoria

Unidad pórfido riolítico

Petrográficamente esta unidad se describe como rocas de color gris claro con fenocristales de plagioclasas y un notorio aumento en cantidad y tamaño de los fenocristales de cuarzo con respecto al Pórfido Escondida. Además, presenta fenocristales de biotita subhedrales de 2mm de espesor, los que generalmente se encuentran alterados a clorita o sericita.

Brechas magmáticas-hidrotermales

Las brechas constituyen la unidad de roca que concentra la mayor cantidad de mineralización, tanto de mena como de ganga. En el yacimiento se las reconoce afectando a todas las unidades con excepción de las gravas. Se desarrollan en el interior de las unidades preexistentes, o bien, en las zonas de contactos de intrusión. Equivalen aproximadamente al 5 % del total de las rocas del yacimiento y, de acuerdo al tipo de relleno y matriz que contienen, se han agrupado en dos tipos: las de origen hidrotermal y las de origen ígneo.

Gravas

Corresponde a una secuencia de sedimentos continentales (areniscas y brechas sedimentarias) mal a moderadamente consolidadas, de color pardo-rojizo y gris blanquecino y compuesta por fragmentos polimícticos, mal seleccionados, angulosos a subangulosos, con tamaños variables milimétricos a decimétricos, cuya composición corresponde a los tipos

litológicos más antiguos. Presenta una matriz de proporción variable, compuesta por detritos tamaño arena y cemento correspondiente a yeso, carbonatos y sales indeterminadas.

En la figura 2.8, se puede apreciar la disposición de la litología en Escondida.

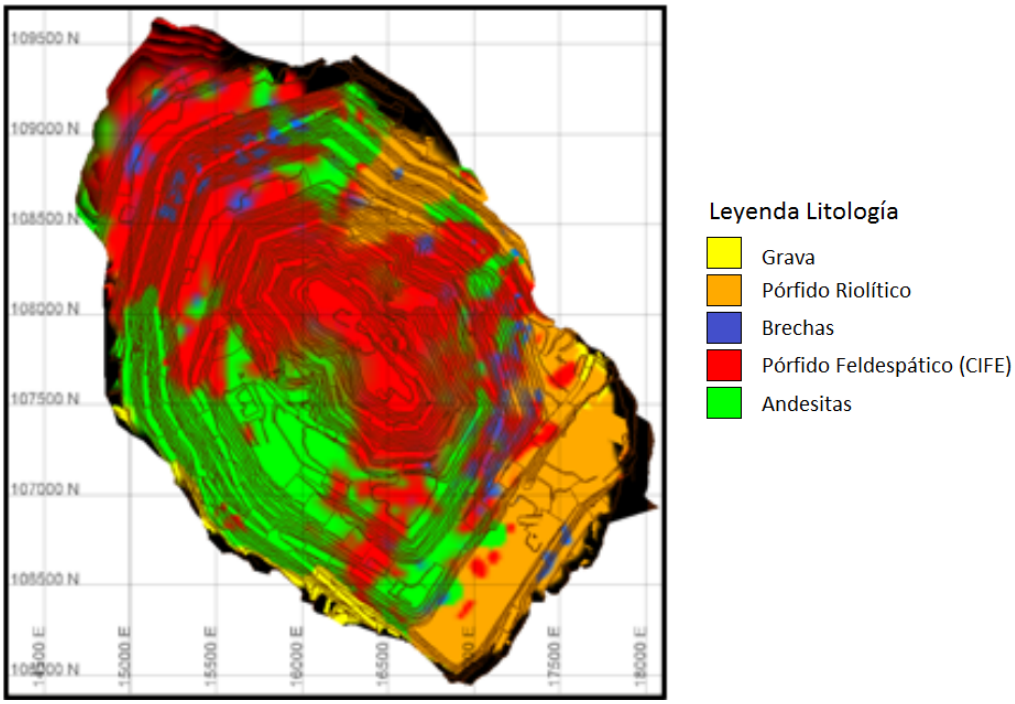


Figura 2.8: Distribución Litología Escondida Enero 2009

2.3.2. Estructuras

El rasgo estructural más relevante en las cercanías de Escondida y Escondida Norte, lo constituye el Sistema de Fallas de Domeyko (SFD). Este sistema corresponde a una amplia zona de fallas que se extiende por más de 1.000 km, desde Collahuasi hasta el río de Copiapó, la que está separada en cinco segmentos con características estructurales bien definidas.

2.3.3. Alteraciones

El depósito Escondida presenta una clara zonación lateral y vertical de las alteraciones hidrotermales que afectan a las rocas. En la figura 2.9, se muestra la distribución que presentan las alteraciones en Escondida. Se puede apreciar una cierta simetría en torno a un núcleo potásico de los demás tipos de ellas, lo que concuerda con el modelo presentado por Lowell y Guilbert (1970), el que corresponde a una zonación idealizada para las alteraciones

hidrotermales en un sistema de pórfido cuprífero. En este modelo se denota un centro potásico, en torno al cual se disponen las alteraciones Fílica, Propilítica y Argílica.

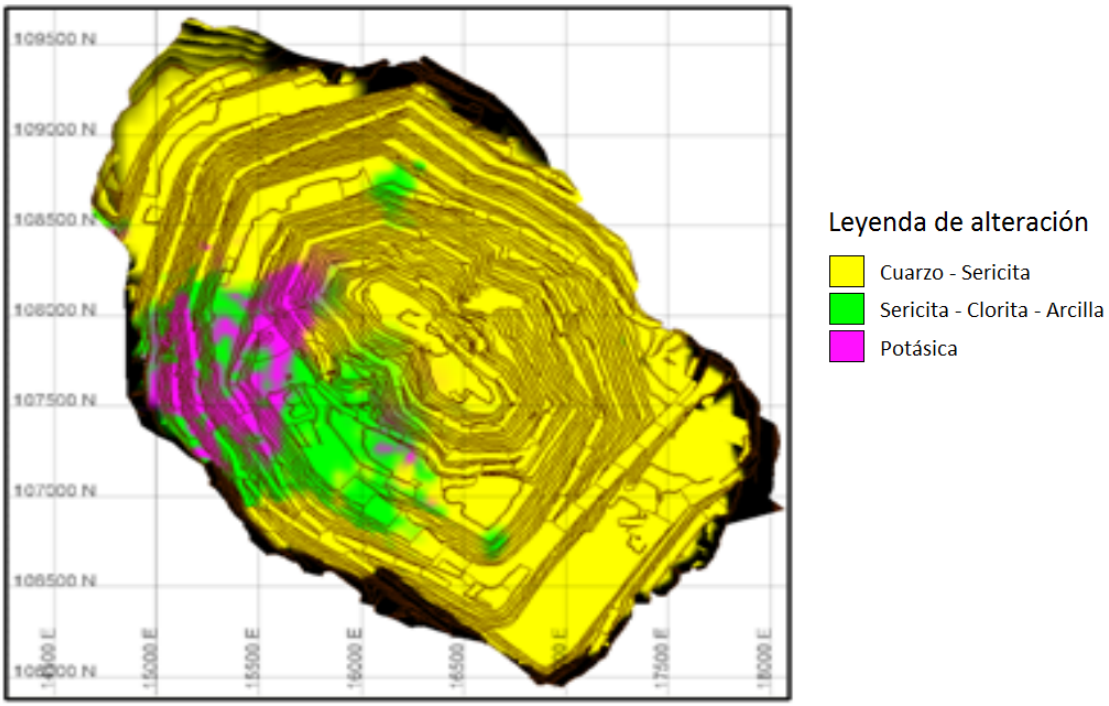


Figura 2.9: Distribución Alteración Escondida Enero 2009

Alteración potásica (KK)

La alteración potásica se presenta como un núcleo, el cual cubre una superficie relativamente menor, y se expone principalmente en la pared oeste del rajo Escondida. Se encuentra afectando a las andesitas de la Formación Augusta Victoria. Está caracterizada principalmente por la asociación feldespato potásico, cuarzo, biotita $\pm$ sericita, en donde se pueden encontrar vetillas de feldespato potásico, las que en ocasiones se encuentran asociadas con cuarzo y además, se puede observar una diseminación de estos dos minerales más biotita en la matriz de las rocas.

Por otra parte, y con mayor desarrollo en las andesitas, se encuentra la alteración biotítica, la que se caracteriza por una casi total alteración de los máficos a biotita. La asociación para esta etapa está representada por biotita-sericita y cuarzo y se interpreta como parte de las etapas más tardías de la alteración potásica.

Alteración sericita-clorita-arcilla (SCC)

La alteración Sericita-Clorita-Arcilla corresponde clásicamente a la alteración propilítica. Las biotitas y las plagioclasas son reemplazadas por clorita-sericita-pirita y sericita-arcilla, respectivamente. Las arcillas asociadas a este tipo de alteración corresponden a caolinita y esmectita. Comúnmente se conservan relictos (parches) de alteración potásica (biotitización y feldespato potásico).

Alteración cuarzo-sericita (QS, QSA)

Esta alteración presenta una relación más directa con las zonas de más alto enriquecimiento mineral. Esta etapa fue denominada como la etapa hidrotermal principal. En Escondida presenta su mejor expresión en el CIFE, principalmente desde el techo de sulfuros hasta la base del enriquecimiento y es la de mayor extensión en el rajo y coincidiendo con las mejores leyes de cobre del depósito.

Alteración argílica

La alteración argílica se encuentra caracterizada por una parte, por la asociación cuarzo-caolinita, a la que se le denomina alteración argílica intermedia típica de esta etapa y para efectos de modelamiento. Cabe recalcar, que antiguamente se presentaban juntas las alteraciones cuarzo-sericita y argílica avanzada provocando que estas fueran mapeadas como una sola, bajo el código 50 (alteraciones blancas).

Alteración argílica avanzada

La alteración argílica avanzada corresponde a una etapa hidrotermal tardía y se reconoce macroscópicamente por exhibir la formación de cúmulos irregulares de color blanco intenso, que presentan una asociación sericita, alunita, pirofilita y diásporo, insertas en una masa subtranslúcida de color gris, conformada principalmente por sílice y sericita.

2.3.4. Mineralización

La etapa de mineralización está directamente relacionada con la intrusión del Complejo Intrusivo Feldespático Escondida. Este complejo fue el que aportó la mineralización primaria y los fluidos hidrotermales responsables de la etapa de alteración hidrotermal, a la que se

asocia el enriquecimiento secundario. Por otra parte el emplazamiento de la mineralización muestra un claro control estructural, presentándose principalmente alojada en fracturas, tanto en el caso de la mineralización hipógena, como la supérgena.

Mineralización hipógena

La mineralización hipógena se puede clasificar en tres principales asociaciones minerales relacionadas a tres etapas hidrotermales. Durante la etapa de alteración potásica destaca la asociación magnetita-bornita-calcopirita con un contenido de sulfuros menor al 0.5 % en volumen. En la zona de alteración sericita-clorita es característica la ocurrencia de vetillas de molibdenita y pirita intercrecidas con sericita. Además en esta etapa se puede observar un aumento en la cantidad de sulfuros a un 2 % en volumen. Por último, la etapa hidrotermal tardía se ve caracterizada por la presencia de vetas polimetálicas con sulfuros como calcopirita, bornita, pirita lamelar, covelina, calcosina, enargita, esfalerita, galena y tenantita.

Mineralización supérgena

Esta mineralización puede ser representada en zona de lixiviados, zona de óxidos y zona de sulfuros secundarios. La zona de lixiviados se encuentra representada por minerales de fierro oxidados, como goethita, magnetita y jarosita, distribuyéndose principalmente en las zonas superiores.

La zona oxidada de cobre se reconoce por la coexistencia de óxidos y sulfatos de cobre. Estos últimos corresponden a brochantita y antlerita, y en forma subordinada atacamita, crisocola, pseudomalaquita, libetenita, turquesa, cuprita, cobre nativo, copper wad y copper pitch. Bajo esta zona se encuentra una zona de mixtos, la cual corresponde a mezclas de sulfuros de cobre, primarios y secundarios, con óxidos de cobre. Los minerales de óxidos de cobre en Escondida están principalmente representados por sulfatos de cobre hidratados (brochantita y antlerita), equivalentes a un 90 % del total de óxidos con menor proporción de atacamita, crisocola, malaquita, azurita, cuprita, cobre nativo y turquesa.

La zona de sulfuros secundarios corresponde a la ocurrencia dominante y/o continua de mineralización sulfurada de cobre de carácter secundario, tales como: calcosina y covelina, localizada entre el techo dominante de sulfuros y la zona primaria. Al interior de esta zona secundaria se puede reconocer una clara zonación mineralógica vertical, con intensidad variable de enriquecimiento, representados desde los niveles superiores a inferiores por:

1. calcosina + pirita

2. calcosina + pirita (- calcopirita)
3. calcosina + pirita (- calcopirita - covelina)
4. pirita + calcopirita + covelina (- calcosina)
5. pirita + calcopirita + covelina (- calcosina - bornita)

Según esta zonación mineralógica vertical, los sulfuros secundarios se subdividen en dos: una subzona superior, denominada Alto Enriquecimiento (HE), que presenta las asociaciones “1, 2 y 3” presentando como mena principal calcosina en pátina sobre pirita, en vetillas y en reemplazo, y en forma subordinada y esporádica, covelina y calcopirita. Una subzona inferior, denominada Bajo Enriquecimiento (LE), caracterizada por la asociación “4 y 5”, distinguiéndose por la presencia continua de calcopirita, junto a covelina (sulfuro más importante de esta subzona), pirita y menor cantidad de calcosina (incluyendo djurleita y digenita).

En la figura 2.10, se puede apreciar la disposición de la mineralización en Escondida.

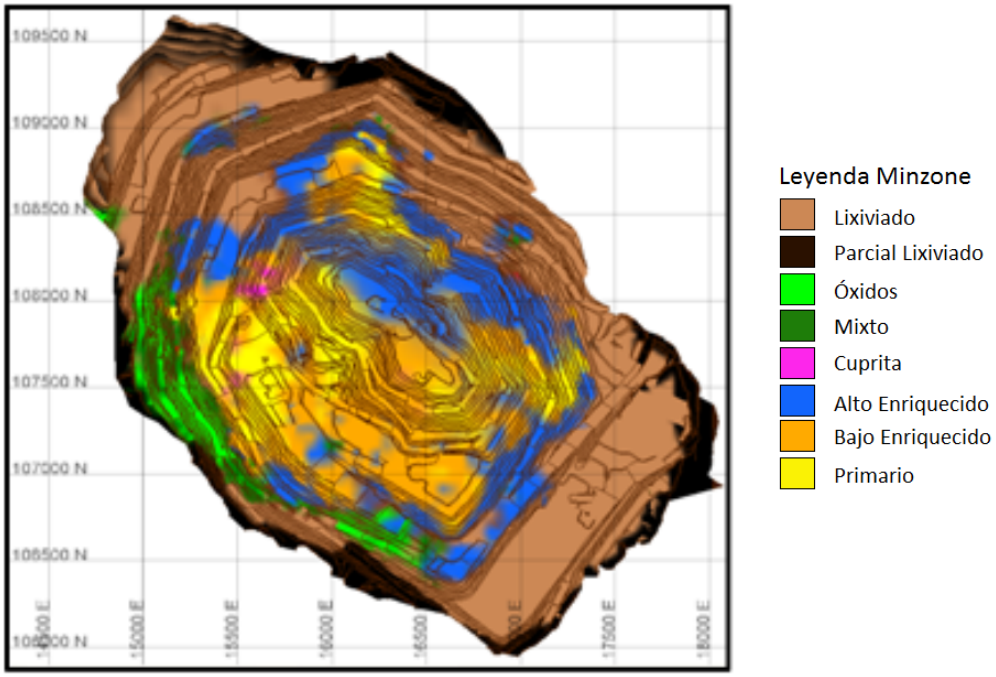


Figura 2.10: Distribución Mineralización Escondida Enero 2009

2.3.5. Mapeo geológico en Escondida

En la mina se realizan diferentes tipos de sondajes exploratorios, siendo estos:

- Aire Reverso (RC).

- Aire Reverso y Diamantina (RD).
- Diamantina Horizontal (HD).
- Diamantina (D).

La codificación puesta en paréntesis es seguida por un guion y un número, el cual da la identificación, por lo que es importante tener la siguiente información:

- Fecha de mapeo
- Escala del mapeo
- Las coordenadas
- La elevación
- El rumbo
- La inclinación
- Control fotográfico

Esta información es ingresada a la base de datos, almacenando la fecha de ingreso. El mapeo geológico se realiza en dos ubicaciones:

- La Chimba: Aquí se estudia la litología, alteración, mine zone y la ley visual.
- Minera Escondida: Donde se analiza la geología de mina, mapeo de sondajes y mapeo geológico.

Toda la información recolectada se valida para luego ser ingresada a la base de datos final de la mina.

Los sondajes, al ingresar a la muestrera, vienen acompañados de información que indica el pozo del cual provienen y la longitud de estos, siendo asignados a diferentes geólogos, que tienen la tarea de ingresar la información de geología y geotecnia, en el programa Acquire (programa de base de datos en formato SQL), con sus coordenadas y sus características correspondientes. Mapeado el sondaje, se realiza una codificación para obtener la información resumida por zonas, para así subirlo finalizado al servidor, esperando que este sea validado por Minera Escondida Ltda (la pauta de descripción se encuentra en anexos A). Posteriormente, en la etapa de generación de esquemas analíticos, llegando las leyes de cobre total y el PtXt (extracción parcial de mineralogía, el cual es un dato que se observa para ver la correlación de minerales), se realiza la codificación nuevamente pero esta vez, con ley en mano se crea una nueva planilla, teniendo presente estos datos, entregándola nuevamente para su validación, la cual se realizará viendo la geología y las leyes, ingresando a la base de datos final.

En resumen se realizan dos validaciones:

- Validación por litología, alteración y mineralización.
- Validación con leyes químicas de laboratorio, realizando la validación con las leyes en mano.

Actualmente, se está implementando realizar una validación en terreno, en donde se analice si el sondaje validado corresponde a la zona mapeada. Es importante mencionar que existen errores al mapear, ya que para algunos sondajes no existe un código adecuado al tratarse de zonas nuevas. Al ocurrir esto, el geólogo debe poner el código más semejante al que crea que tenga en realidad, ingresando esta información como una observación. Es muy importante que los techos y pisos de las mine zones sean reconocidos para definir bien los minerales. En litología se captura:

- Tipo de litología (andesitas, pórfidos, brechas, volcanita, etc.).
- Características particulares de la litología (en el caso de que sean brechas, qué matriz posee, qué porcentaje de esta tiene, si tiene clastos, qué porcentaje de clastos, etc. En el caso de pórfido es importante buscar si se encuentran ojos de cuarzo y fenocristales de feldespato, tamaño y porcentaje).

Cuando el geólogo mapea la litología, debe notar los minerales formadores de roca, como lo es el cuarzo tratando de ver la cantidad que se presenta en la roca. Por ejemplo si hay más de un 60 %, le dará al geólogo una mayor seguridad de que el sondaje está dentro de un pórfido, si es menor se podría inferir que se está dentro de una andesita, analizando de igual manera el contenido de feldespato, el de plagioclasas observando los tamaño y sus forma ayudando a relacionar el tipo de roca, con el objetivo de disminuir la cantidad de errores.

Al mapear los minerales, se tienen distintos eventos, en el área de las texturas. Donde, por ejemplo, se trata de diferenciar un pórfido cuarcífero de uno reolítico, para esto se tiene que analizar qué minerales están presentes en la muestra y ver la existencia de vetillas. Se debe apreciar un conjunto de variables al mismo tiempo, las cuales asociadas, debiesen dar una lógica a una litología y así relacionarla con el sondaje.

En general, se tiene dos formas de ver las alteraciones presentes; dominante (primaria) y la subordinada (secundaria). A cada una de estas se le asigna una intensidad de alteración en la base de datos, siendo dichas alteraciones definidas por medio de los minerales que se puedan apreciar en la muestra. En la base de datos se adjunta cuáles son los minerales presentes y en qué intensidad se encuentra cada uno de estos, asignándole valor 1 (débil),

2 (moderado) y 3 (fuerte), según la presencia de mineral en la roca; con esto se realizan asociaciones mineralógicas de cada alteración determinada por diferentes relaciones, por ejemplo: la asociación cuarzo-sericita-arcilla está asociada a una alteración fílica QS, en donde existen diferente alteraciones QS, dependiendo si se tiene arcilla o no. Si existe clorita y no hay arcilla, se tiene una alteración clorita-sericita SCC. Es recurrente que se mapee bien la mineralogía y que se asocie a una alteración diferente equivocándose en la codificación.

Se mapean los minerales accesorios y sus intensidades, ingresando también la intensidad de alteración de los máficos y la destrucción de los feldespatos. Los minerales accesorios son relevantes, ya que por ejemplo, si el geólogo pone en una alteración que existe caolín o arcilla el feldespato debería estar completamente destruido, y el máfico, que podría ser la biotita u hornblenda, debería estar completamente blanco. Debido a esto, se puede ver si existe alguna equivocación en cuanto a si es arcilloso o no.

Al ver la información, se puede observar del sondaje, el metraje, el tipo de alteración dominante, su intensidad, el tipo de alteración subordinada, su intensidad, la alteración de los máficos, la destrucción de los feldespatos, los minerales presentes y sus intensidades.

Es importante ver los techos y los pisos como el de óxidos, el de sulfuros, de calcopirita, el de bornita y el techo primario. Estos techos deben coincidir con la zona mineralógica, en donde si se marca uno de estos, no debiesen aparecer minerales asociados a esa alteración fuera de ese intervalo.

2.4. Métodos de Análisis Multivariable

A continuación se presentan de algunas herramientas para el análisis multivariable con el fin de realizar la etapa de minería de datos y obtener información de la base de datos.

2.4.1. Correlación

Es una herramienta estadística cuyo valor se mueve en un intervalo entre -1 y 1, en donde, -1 son las relaciones inversas y 1 las perfectamente lineales. Para obtenerla se calcula de la siguiente manera. Sean:

$$\mu_a, \mu_b \qquad \text{medias de variables A y B}$$

σ_a, σ_b desviaciones estándar de las variables A y B

n número de mediciones de (A,B)

Covarianza:

$$\sigma^2_{a,b} = \frac{1}{n}\sum_{i=1}^{n}(a_i - \mu_a)(b_i - \mu_b) \tag{2.1}$$

Correlación:

$$\rho_{a,b} = \frac{\sigma^2_{a,b}}{\sigma_a \sigma_b} \tag{2.2}$$

2.4.2. Gráfico de dispersión

Permiten comparar las muestras para dos variables y los valores que toman. De esta forma, es posible identificar gráficamente dependencias entre los valores de dos variables y encontrar tendencias lineales, cuadráticas o de otro tipo.

En la figura 2.11, se aprecia un gráfico de dispersión para dos variables x e y, incluyendo información categórica.

2.4.3. Boxplots

Los boxplots son una técnica gráfica que consisten en realizar un despliegue de la distribución de variables. Es particularmente útil para la comparación de clases. Estos diagramas representan gráficamente 5 elementos para el estudio, los cuales son el cuartil superior F_U, el cuartil inferior F_L, la media y los extremos [9]. El esparcimiento-F, es definido como $d_F = F_U - F_L$. Los extremos que definen los datos atípicos quedan definido por:

$$Borde_{Superior} = F_U + 1.5 d_F \tag{2.3}$$

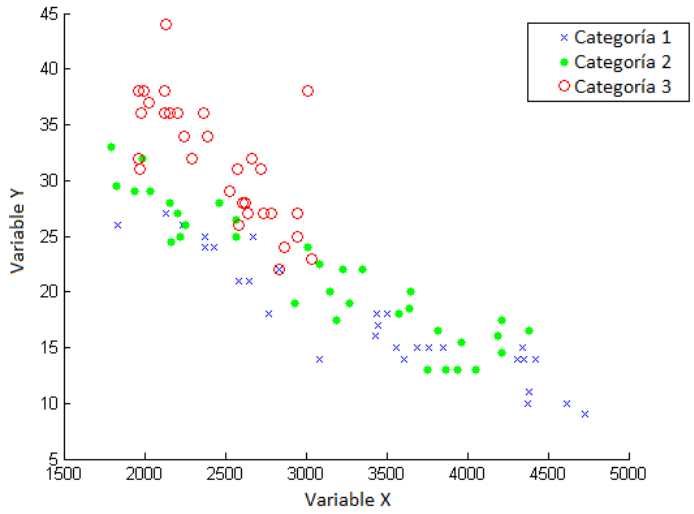


Figura 2.11: Ejemplo gráfico dispersión

(2.4) $$Borde_{Inferior} = F_L - 1.5d_F$$

La figura 2.12 muestra un boxplot graficado horizontalmente indicando como este ejemplifica la distribución comparándola con una normal.

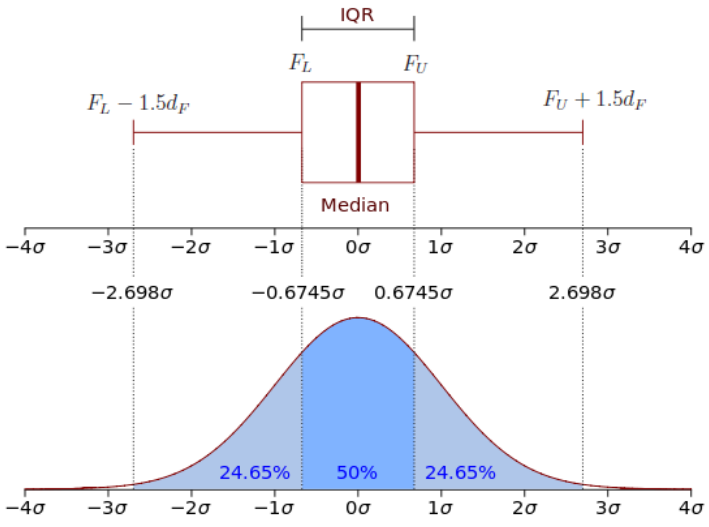


Figura 2.12: Boxplot

2.4.4. Histogramas

Los histogramas son estimadores de la densidad de los datos, entregando una buena impresión de la distribución de la información. En contraste con los boxplots un histograma puede mostrar multimodalidades de los datos. La idea de este gráfico es representar la densidad localmente contando el número de observaciones en intervalos consecutivos [9]. La figura 2.13 muestra un ejemplo de histograma con una distribución multimodal.

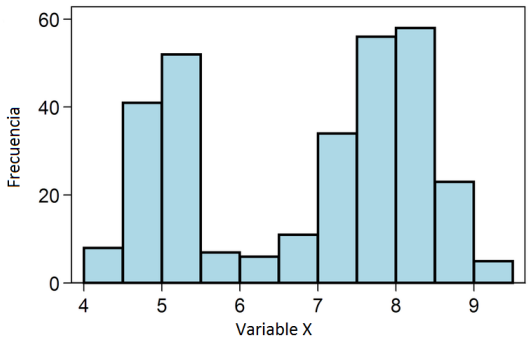


Figura 2.13: Ejemplo Histograma

2.4.5. Regresiones

Es una metodología estadística para predecir valores de una o más variables dependientes desde una colección de variables predictoras (independientes). Puede ser usada también para evaluar los efectos de las variables predictoras.

Esta trata de encontrar una función de diversas variables del tipo:

$$\hat{Y} = a + bX \tag{2.5}$$

Minimizando el Error:

$$\epsilon = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i [Y - (a + bX_i)]^2 \tag{2.6}$$

2.4.6. Discriminante y clasificación

Técnica relacionada con la separación de distintos conjuntos y/o sets de objetos (observaciones) con asignación de nuevos objetos, a los set previamente definidos, siendo este método más bien de exploración. Como un proceso separativo, es a menudo utilizado para investigar diferencias observadas cuando las relaciones causales no se entienden muy bien. Esta técnica tiene como objetivo:

- Describir de forma gráfica y algebraica, a diferentes características de una medición, a partir de muchos grupos conocidos de observaciones. Se trata de encontrar discriminantes en donde sus valores numéricos son tales que las colecciones se separen lo más posible.
- Ordenar las observaciones en dos o más clases etiquetadas. El énfasis está en derivar reglas que puedan ser usadas para asignar óptimamente nuevos objetos en las clases etiquetadas.

En ciertos casos una función que separa los objetos, puede servir como un asignador, y, por el contrario, una regla que asigna objetos puede sugerir un procedimiento discriminatorio [10].

2.4.7. Clustering (agrupamiento)

Análisis exploratorios de datos son de poca ayuda al momento de entender la compleja naturaleza de las relaciones multivariables. El clustering usa medidas de similitud para contabilizar la distancia en semejanza de las variables y/o mediciones, realizando grupos entre ellos. A diferencia de la clasificación, los grupos no se conocen de antemano y tienen que ser descubiertos durante el análisis.

Las herramientas de clustering se pueden agrupar en dos grandes familias las cuales son:

Métodos de clustering jerárquicos

Es muy difícil relacionar todos los posibles grupos de puntos y variables, siendo esta clase de clustering una manera de obtener resultados razonables.

Métodos aglomerativos

Comienzan con un solo dato y luego van siendo añadidos los más cercanos convergiendo a grupos más grandes.

Single linkage: son ordenados de acuerdo a la distancia a sus miembros más cercanos (mínima distancia).

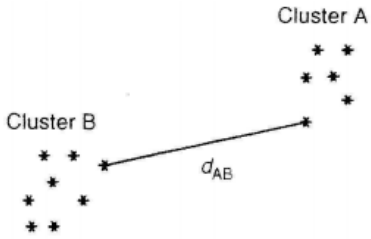


Figura 2.14: Single linkage

Complete linkage: son ordenados con respecto a la más alejada (máxima distancia).

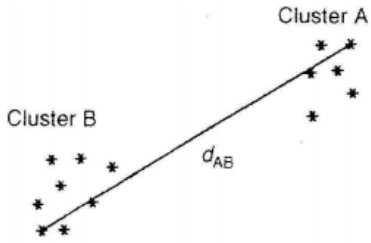


Figura 2.15: Complete Linkage

Average linkage: son ordenados de acuerdo a la distancia promedio.

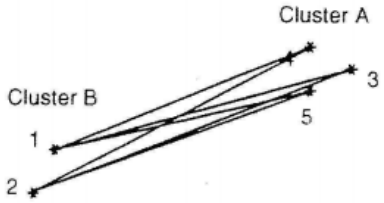


Figura 2.16: Average Linkage

En la mayoría de los métodos de clustering las fuentes de error no están formalmente consideradas en los métodos jerárquicos, significando que estos métodos son muy sensibles a outliers.

En clúster jerárquicos no existe una reacomodación de los objetos que han sido agrupados incorrectamente en etapas tempranas, por lo cual la configuración final debe ser cuidadosamente examinada para ver su sensibilidad. Es buena idea tratar con distintos métodos de clustering, y diferentes tipos de distancias. Si los métodos son consistentes entre ellos, puede ser que sean casos de clustering natural.

La estabilidad de las soluciones jerárquicas, puede ser revisada aplicando los algoritmos de clúster antes y después de que los pequeños errores han sido removidos. Si los grupos son muy bien distinguibles, ambos modelos de clustering deberían ser acordes.

Valores iguales de similitud o de matriz de distancia, pueden producir múltiples soluciones correspondientes a diferentes maneras de tratarlas, lo cual no implica que esté necesariamente erróneo. Conociendo su existencia, se podrá interpretar apropiadamente comparando con sus pares.

Dendrograma

Un dendrograma es un tipo de representación gráfica o diagrama de datos en forma de árbol usada usualmente en métodos de aglomeración jerárquicos tal como se muestra en la figura 3.15.

Cada combinación es representada por una línea horizontal. El eje de la ordenada representa el grado de similitud de los grupos que se han fusionado, llamando a esta similitud de combinación [11].

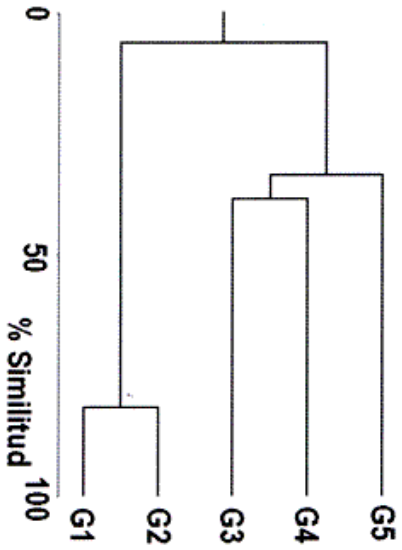


Figura 2.17: Dendrograma

Métodos no jerárquicos

Los métodos no jerárquicos son más diseñados para agrupar puntos que para agrupar variables, siendo el número k de clusters especificado o puede también ser determinado como parte del procedimiento. Como una matriz de distancia no tiene que ser calculada, pueden ser aplicados a set de datos mucho mayores que para los métodos jerárquicos. Estos métodos parten desde una división inicial de los ítems en grupos o con sets de semillas, que son el núcleo de los clúster. Buenas elecciones para configuraciones iniciales debiesen estar libres de sesgo, siendo una buena manera de iniciar escogiendo ítems aleatorios o particiones al azar.

Método K-mean

En este método, dado un set de datos vectoriales $X = [X_1, ..., X_n]$ donde n es el número de observaciones, el algoritmo agrupa los puntos minimizando la función objetivo (para distancia euclidiana):

$$J = \sum_{j=i}^{k} \sum_{i \epsilon C_j} \|X_i - \mu_j\|^2 \tag{2.7}$$

donde c_j es el j-ésimo cluster y μ_j es su centroide.

En resumen es asignar ítems a un clúster por cercanía de los centroides (promedios), siguiendo 3 simples pasos.

1. Dividir los ítems en K clusters.
2. Proceder a través de la lista, reasignando puntos al clúster con centroide más cercano, normalmente utilizando distancias euclidianas.
3. Repetir el paso 2 hasta que no haya más asignaciones.

Como comentario de estos métodos se puede decir que:

- Existen muchos argumentos para no fijar el número K de clústeres [12].
- Si dos o más semillas caen inadvertidamente en un clúster, su resultado será pobremente diferenciado.
- La existencia de outliers puede producir que al menos un grupo esté demasiado disperso.
- Si incluso, la población consiste de k grupos conocidos alguno de ellos puede ser datos dispersos que no se agruparan, forzando al clustering, a formar clústeres que pudiesen perder sentido.
- En algunos casos se sugiere realizar el algoritmo repetidas veces para un número k específico.

2.4.8. Componentes principales

Un análisis de componentes principales está relacionado con la explicación de la estructura de varianza-covariancia de un set de variables, a través de pocas combinaciones lineales de éstas. Sus objetivos generales son:

- La reducción de información.
- Una mejor interpretación de los datos.

Aunque p componentes son requeridos para reproducir la totalidad de la variabilidad del sistema total, a menudo, mucha de esta variabilidad puede ser acontecida por un pequeño número de k de componentes principales. Si entonces, hay tanta información en los k componentes como lo había en las p originales variables, y el set de datos original, consiste de n mediciones de p variables siendo reducida a un set consistente de n mediciones sobre k componentes principales. Un análisis de componentes principales a menudo revela relaciones que no fueron previamente sospechadas y así permite interpretaciones que no serían resultados ordinarios [12].

2.4.9. Redes neuronales artificiales

Un modelo de redes neuronales artificiales (RNA) es un modelo matemático desarrollado para imitar las funciones del sistema nervioso humano. Este consiste en un número de simples e interconectadas neuronas artificiales. Cada neurona recibe información de una variable de entrada o de la salida de otras neuronas, en donde realiza operaciones simples con la información enviando una señal de salida a las otras neuronas. Debido a sus ventajas en aprendizaje adaptativo, auto-organización, tolerancia a fallos y a sus operaciones en tiempo real, las RNA se han convertido en una poderosa herramienta en el reconocimiento de patrones [13].

Las RNA pueden ser clasificadas en diferentes tipos de redes de acuerdo a tres principios:

1. Arquitectura: determina cómo las neuronas de una red están conectadas.
2. Aprendizaje: determina cómo la red es entrenada y como la información es almacenada en ella.
3. Llamada: determina cómo recuperar los datos almacenados.

La figura 2.18 está constituida por dos imágenes. La superior, muestra una neurona biológica y bajo ella una artificial (perceptrón). Las variables de entradas más el sesgo (bias) son ponderados por los pesos sinápticos siendo estas sumadas, para luego, aplicarles alguna función de activación que será la salida de la neurona.

Las RNA “feed-forward” (FNN) son una de las más utilizadas, estando organizadas en estructuras de capas conectadas de manera estrictamente progresiva (la información de salida de una neurona no se devuelve a neuronas que estaban en la misma o en capas anteriores).

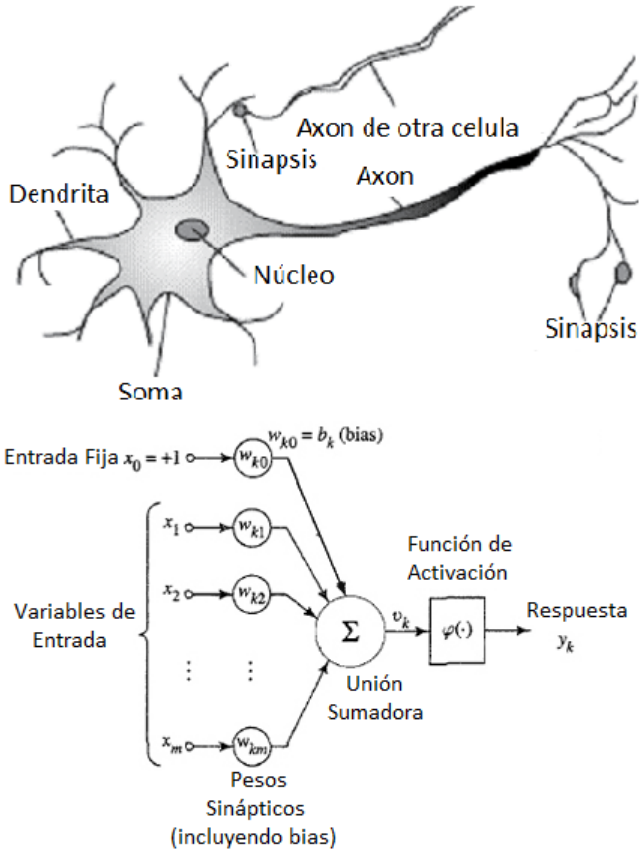


Figura 2.18: Neurona artificial

La red es entrenada con diferentes algoritmos de aprendizaje en donde cada uno de estos incluye la información de entrada (variables) y de salida (patrón) siendo la información de la red almacenada como pesos de conexión, los cuales son actualizados durante el entrenamiento en pro de minimizar el error entre las actuales salidas generadas y las salidas esperadas del sistema, siendo esto llamado entrenamiento supervisado.

La estructura básica de una FNN es presentada en la figura 2.19, en donde solo se tiene una capa oculta en la red pudiendo existir más de una, dependiendo de la configuración de esta. En la capa de entrada y en las capas ocultas siempre existe una neurona bias (variable con valor constante comúnmente 1 o -1). Dada la necesidad de entrenamiento de la FNN, la neurona bias sirve como herramienta para incrementar los grados de libertad de la red (permitiendo a las neuronas tener control efectivo sobre los umbrales de activación) y para ir actualizando los valores de los pesos sinápticos permitiendo que exista una convergencia de los pesos en pro de obtener una solución aceptable.

A continuación se explicará el desarrollo tradicional para el desarrollo de una FNN.

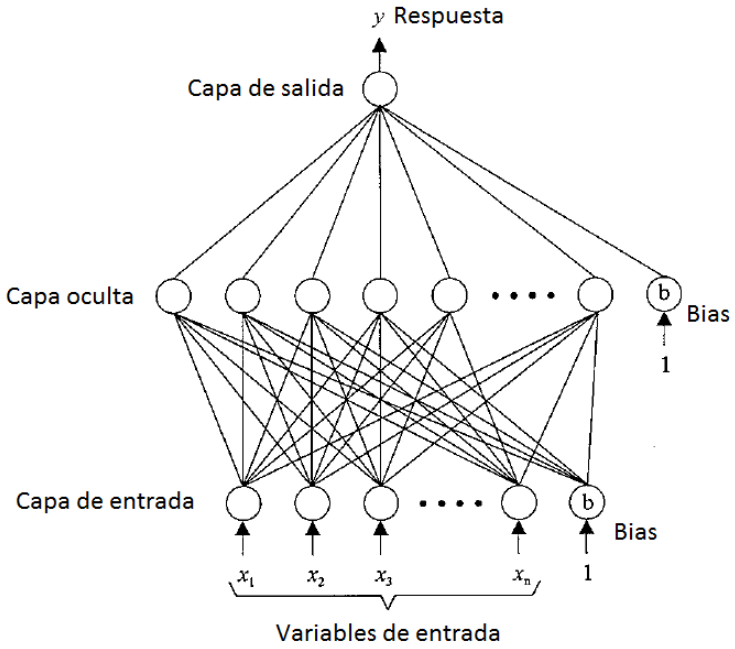


Figura 2.19: Estructura básica de una FNN

Arquitectura

El procedimiento tradicional usado en la aplicación de una FNN a un problema de clasificación, es resumido en la figura 2.20. Como muestra la figura, el procedimiento puede ser separado en dos etapas, teniendo como objetivo de la primera, definir la arquitectura, y de la segunda testear, el entrenamiento.

El número de capas y la cantidad de neuronas en cada una de ellas determina la arquitectura de una FNN. En general, solo una capa oculta en una FNN ha sido utilizada.

Entrenamiento

Entrenar una FNN es lo mismo que resolver un problema de programación no lineal (PNL). Las variables del programa son los pesos de la FNN y la función objetivo, es el promedio de los cuadrados de los errores entre el valor de salida deseado y el valor generado por la red. Si existiese más de una neurona de salida, el error será el promedio del error entre cada una de estas clases. Los algoritmos de gradiente descendente son comúnmente usados para resolver PNL, siendo conocidos como “back-propagation error”.

Al momento de entrenar se debe tener cuidado con el sobreajuste y sobreentrenamiento (over-fitting/ over-training), ya que son problemas comunes en las RNA. Dos condiciones influencian estos problemas:

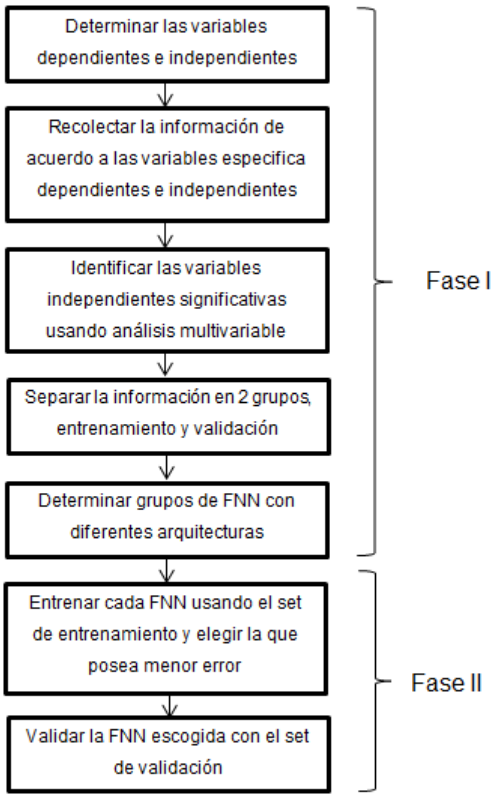


Figura 2.20: Aproximación al desarrollo de una FNN

1. Tamaño de una RNA.
2. Tiempo de entrenamiento.

El sobreajuste ocurre cuando el tamaño de una RNA excede su óptimo y el sobre-entrenamiento hace referencia al tiempo que se efectúa este, el cual podría resultar en una escasa habilidad como predictor de la red [14]. Esto se debe a que la red comienza a memorizar información ruidosa, impidiéndole identificar los patrones más importantes, en la figura 2.21,se puede apreciar el error de entrenamiento en azul y el de validación en rojo, en función de los ciclos de entrenamiento. Si el error de validación se incrementa mientras que el error de entrenamiento sigue disminuyendo, lo más seguro es que se ha producido sobreajuste en el modelo, en donde este almacena patrones específicos de la base de datos de entrenamiento perdiendo de esta manera generalidad y reproducibilidad. El mejor modelo predictivo se encuentría cuando el error de validación se encuentra en su mínimo (modelo mas general).

Ventajas y desventajas de las redes neuronales

- Los modelos de redes neuronales requieren menos entrenamiento formal estadístico para ser desarrolladas.

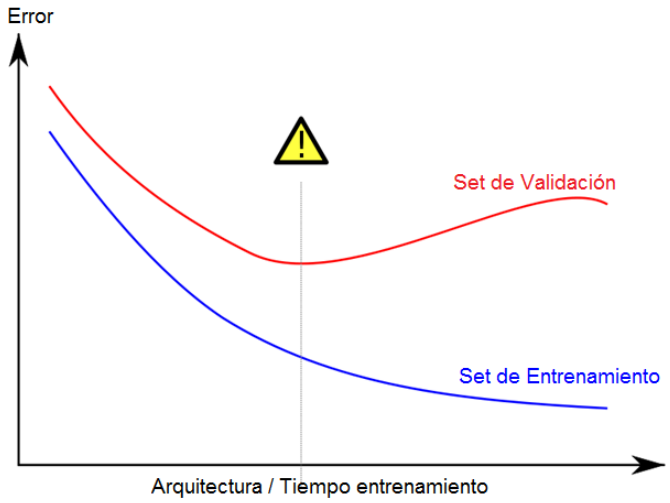


Figura 2.21: Sobreajuste en RNA

- Pueden detectar implícitamente relaciones no lineales complejas entre las variables independientes y las dependientes.
- Tienen la habilidad de detectar todas las posibles interacciones entre las variables de predicción.
- Pueden ser utilizadas usando diferentes algoritmos de entrenamiento.
- Son “cajas negras” y tienen una habilidad limitada para identificar explícitamente las posibles causales de las relaciones.
- Son más difíciles de utilizar en terreno.
- Requieren mayores recursos computacionales.
- Son proclives al sobreajuste.
- Su desarrollo es empírico y muchos problemas metodológicos esperan por resolverse [15].

2.4.10. Regresión logística

El modelo de regresión logística permite estimar la probabilidad de un suceso binario, el cual depende de los valores de ciertas covariables. Supongamos que un suceso (o evento) de interés A puede presentarse o no en cada uno de los individuos de una cierta población. Considerando una variable binaria y que toma los valores:

(2.8) $$y = 1 \quad \text{si A se presenta}, \quad y = 0 \text{ si A no se presenta}$$

Si la probabilidad p de que ocurra el evento depende de los valores de ciertas variables

$x_1 \dots x_p$, es decir, si $X = (x_1 \dots x_p)'$son las observaciones de un cierto individuo w sobre las variables, entonces la probabilidad de acontecer A dado X es $p(y = 1|X)$. Indicando esta probabilidad por $p(x)$. La probabilidad contraria de que A no suceda dado x será $p(y = 0|X) = 1 - p(x)$. Pretender que $p(x)$ sea una función lineal de x no puede funcionar correctamente, pues $p(x)$ está comprendido entre 0 y 1 [16].

Por diversas razones, es conveniente suponer un modelo lineal para la llamada transformación logística de la probabilidad:

$$ln[\frac{p(x)}{1-p(x)}] = B_o + B_1 x_1 + \cdots + B_p x_p = B_o + B' x \tag{2.9}$$

Equivale a suponer las siguientes probabilidades para A como:

$$p(x) = \frac{e^{B_o + B' x}}{1 + e^{B_o + B' x}} \tag{2.10}$$

2.4.11. Ventajas y desventajas de usar redes neuronales o regresiones logísticas

Muchos investigadores han comparado ambos métodos utilizando distintos conjuntos de información. Algunos de ellos han logrado encontrar un mejor desempeño ocupando redes neuronales. Otros en cambio, no han observado diferencia alguna. Estas comparaciones son altamente dependientes de la naturaleza del conjunto de datos, por lo cual no se puede concluir qué método es superior al otro.

Ventajas y desventajas de la regresión logística

- Simplicidad: la ecuación de la regresión logística es muy simple.
- Interpretabilidad: se puede conocer cómo afecta al modelo un cambio en el valor de una variable (exponencialmente).
- El modelo de regresión logística la expresa en términos de probabilidad.
- Es más difícil manejar relaciones no lineales complejas.
- No funciona bien con pocos datos.
- Es muy sensible a datos aberrantes [17].

2.4.12. Estrategias de Selección de variables

Al trabajar con distintos modelos matemáticos es difícil decidir qué variables pueden predecir un evento en orden a construir un modelo adecuado. Al tener demasiadas variables en el modelo, es común que muchas de ellas no sean de ayuda y solo aporten ruido en la investigación, por lo que resulta ventajoso seleccionar las mejores, previo al proceso de modelaje, conociendo la existencia de variadas formas de selección de variables, como lo son:

Selección forward

Consiste en comenzar buscando la variable que genere un menor error en el modelo. Luego esta es examinada conjuntamente con cada una de las variables restantes, para encontrar la mejor tupla, siguiendo así hasta que el error aumente en vez de disminuir, o se cumpla algún criterio de detención .

Selección backward

Este algoritmo opera en la dirección opuesta al anterior. En un set de datos compuestos por p variables se crea un modelo utilizándolas todas, a continuación, se hace un nuevo modelo utilizando $p-1$ variables, en donde cada una de ellas es omitida por turnos reteniendo el subgrupo de variables óptimo. Luego con este subgrupo de $p-1$ variable se le es omitida nuevamente cada una de las variables, teniendo como resultado un subgrupo de $p-2$ variables. Este proceso es repetido hasta alcanzar un subgrupo óptimo conformado por m variables según el criterio de detención que se tenga.

Selección stepwise

Es una modificación de la selección forward, comenzando con un modelo que solo incluye 1 variable, luego en las siguientes iteraciones, más variables son añadidas, 1 a la vez, seguida por una eliminación backward en cada iteración, siendo aplicada hasta alcanzar un criterio de detención.

La selección forward modificada por Chen [18] se inicia separando la base de datos en dos, entrenamiento y validación, tomando la de entrenamiento para crear un modelo con cada una de las variables (teniendo n modelos, donde n es la cantidad de variables inicialmente), calculándoles a cada uno, el error asociado que se genera al ingresar la información de

la base de datos de validación, seleccionando de esta manera la variable que genere el mejor resultado (en validación no en entrenamiento), buscando que no exista sobreajuste del modelo. El siguiente paso, al igual que la metodología forward normal, es crear modelos con la variable escogida conjuntamente con cada una de las $n-1$ variables restantes (se tendrán $n-1$ nuevos modelos). Se evalúan y si el error del modelo anterior es mayor o se cumple un criterio de detención se escoge la variable, de no ser así se vuelve a iterar.

Una limitante de estas selecciones, es que solo una pequeña porción del espacio de búsqueda es tomado en cuenta. Por ejemplo, dos variables a y b pueden ser insignificantes separadamente, pero en conjunto, pueden proveer información útil. A pesar de esto, dichos métodos son mucho mejores que la selección más simple que se puede realizar, la que sería generar una búsqueda exhaustiva de todos los subgrupos que se puedan formar con las variables que se tienen, lo que se traduce en buscar los $\sum_{i=1}^{N} = p!/m!(p-m)!$ subgrupos posibles de conformar, siendo esto, computacionalmente demasiado caro (en la mayoría de las situaciones imposible) [19]. Otros métodos de selección, son los estocásticos, estos se ven ejemplificados por algoritmos genéticos, programación evolucionaria entre otros, siendo más lentos computacionalmente en comparación con los otros (ocurriendo sobreajustes en el modelo, es posible que estos algoritmos tarden demasiado) [20], por lo cual no se ocuparán en el modelo, ya que se busca tener una respuesta rápida a las necesidades que se plantean.

2.4.13. Curva roc (característica operativa del receptor)

Las curvas características operativas del receptor o curvas ROC (por su nombre en inglés) son curvas paramétricas que resultan ser una técnica útil de visualización, organización y selección de clasificadores basados en su desempeño. Son usadas en distintas áreas de la ciencia, siendo ocupadas en métodos de “machine learning” demostrando ser una herramienta de gran valor al evaluar y comparar distintos algoritmos [21].

Algunos métodos de clasificación producen una salida continua en la predicción, que pueden ser aplicados a diferentes umbrales de corte para predecir a qué grupo pertenece. De esta manera se presentan cuatro escenarios distintos de salida, estos son:

- Positivo Verdadero: si es positiva y es clasificada como positiva.
- Negativo Falso: si es clasificado como negativa siendo positiva.
- Negativo Verdadero: si es negativa y es clasificada como negativa.
- Positivo Falso: si es clasificado como positiva siendo negativa.

Estos pueden ser vistos en el recuadro de la tabla 2.2.

Tabla 2.2: Matriz de confusión ROC

		Clase Verdadera: Positivo	Clase Verdadera: Negativo
Clase Hipótesis	Si	Positiva Verdadera	Positiva Falsa
	No	Negativa Falsa	Negativa Verdadera

$$Tasa\ Positivo\ Verdadero\ (\boldsymbol{TPR}) = \frac{PV}{Pos}.$$
$$Tasa\ Positivo\ Falsa\ (\boldsymbol{FPR}) = \frac{NF}{Neg}.$$

La curva ROC, es un gráfico de dos dimensiones, que se representa a través de la Tasa Positivo Verdadero (TPR) en el eje de la ordenada, y la Tasa Positivo Falsa (FPR) en eje de la abscisa.

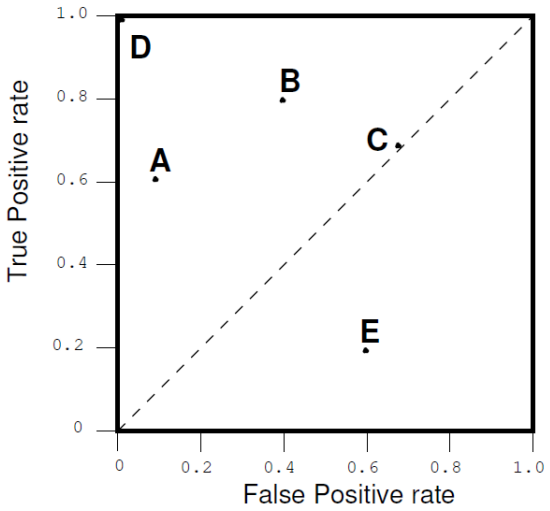


Figura 2.22: ROC básico, mostrando clasificadores discretos

Un clasificador discreto (0 ó 1), produce un punto en el espacio del gráfico ROC (FPR, TPR), estos se pueden apreciar en la figura 2.22, que muestra 5 distintos clasificadores. En el gráfico es importante notar el punto $(0, 0)$, el cual, representa que el clasificador nunca realizó una clasificación positiva, siendo su opuesto el punto $(1, 1)$. El punto $(0, 1)$ representa la clasificación perfecta, para clasificadores donde el valor de salida es una variable continua es posible generar una curva a medida que se va cambiando el umbral de corte. En la figura 2.23 se puede apreciar las distribuciones de los valores de salida para dos distintos clasificadores, donde uno clasifica de forma perfecta y el otro totalmente errado.

El área bajo la curva en un gráfico ROC es una medida de la calidad del clasificador, la cual tiene valores entre 0.5 y 1, tomando el 1 como el discriminador perfecto y 0.5 el peor.

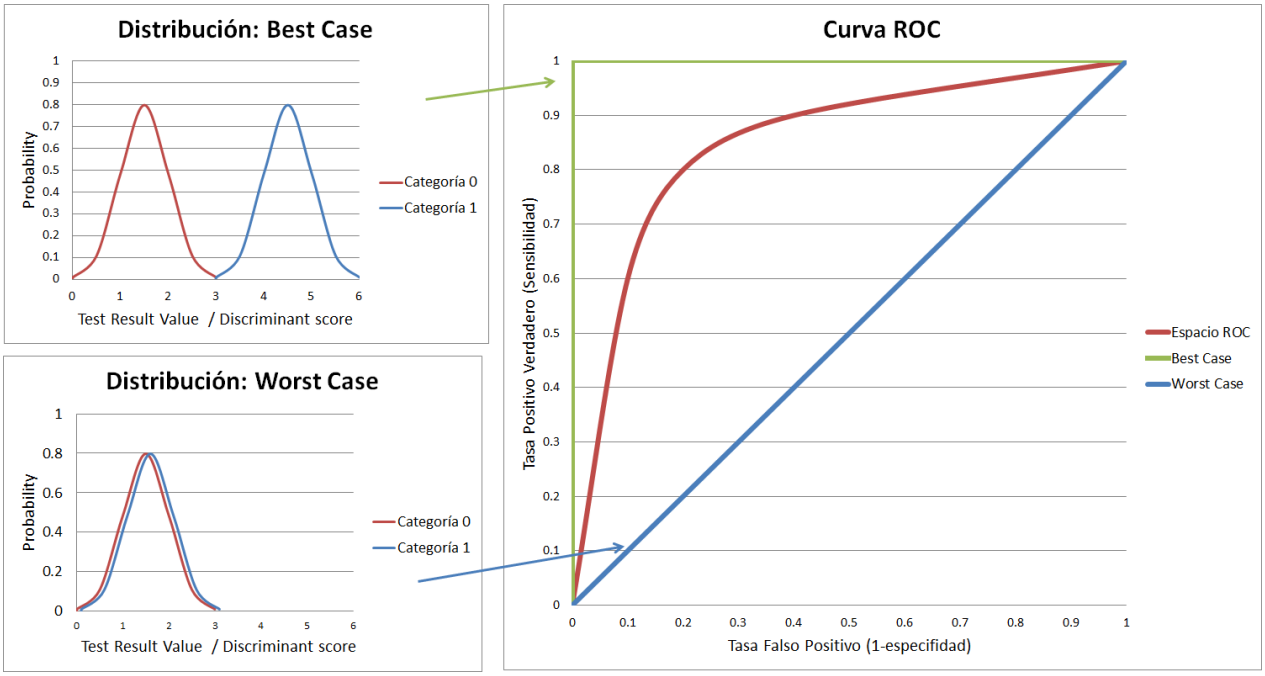


Figura 2.23: Curva ROC Distribución

Curva roc en k-mean clustering

Al contrario de metodologías como RNAs o Regresiones logísticas, k-mean clustering no entrega un valor continuo, lo cual transforma la construcción de una curva ROC en un problema no trivial. Se ha creado un método para poder realizarlas a través de un proceso iterativo no-paramétrico, sin tomar en cuenta las funciones de distribución de los clústers utilizados, en pro de obtener una cantidad de puntos (TPV, TFP) que generen la curva [22].

El proceso se describe de la siguiente manera:

1. Aplicar K-mean Clustering con $k = 2$.
2. Computar TPV y TFP para los clústeres C_1 y C_2 respectivamente.
3. Determinar el clúster verdadero positivo comparando TPV y TFP de la siguiente manera:
   - C_1= Cluster positivo verdadero si $TPV_1 > TPV_2$.
   - C_2= Cluster positivo si $TPV_2 > TPV_1$.
4. Computar (TPV, TPF).
5. Reubicar un punto X_i de C_1 a C_2 tal que el cambio en J sea mínimo ($J = \sum_{j=i}^{k} \sum_{i \epsilon C_j} \|X_i - \mu_j\|^2$), de esta manera se cambian de grupo los puntos mas cercanos al otro cluster, que aseguraría un menor cambio en J.
6. Computar (TPV, TPF)
7. Repetir el paso 5-6 hasta que C_1 se transforme en un clúster vacío.

8. Reinstalar el resultado del clustering original.
9. Reubicar los puntos x_i de C_2 a C_1 talque el cambio de J sea mínimo.
10. Computar (TPV, TPF).
11. Repetir los pasos 9-10 hasta que C_2 se convierta en un clúster vacío.
12. Una vez que todos los puntos (TPC, TPF) sean obtenidos, la curva ROC puede ser generada, graficando estos en el espacio.

En la figura 2.24, se aprecian dos clusters con sus respectivos centroides, con el objetivo de generar el menor cambio en J se tendría que cambiar el punto 1 del cluster 2 al cluster 1 que se encuentra a una distancia d_1 del centroide 1 menor que d_3 del centroide 1.

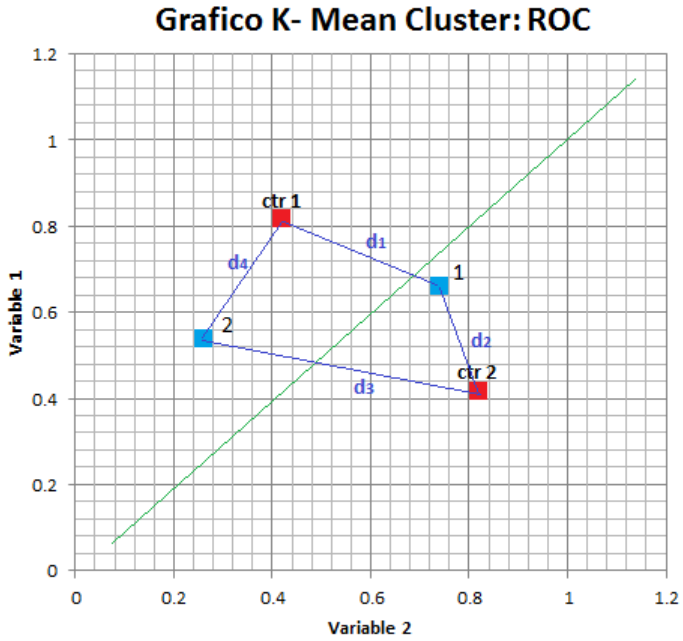


Figura 2.24: Ejemplo confección de curva roc con K-mean

Capítulo 3

Metodología

En este capítulo se expondrá la metodología completa desarrollada para la realización del tema, el que, a manera de resumen, se compone básicamente de cuatro pasos, que son los siguientes:

1. Realizar un estudio exploratorio de datos, para entender cómo se encuentran distribuidas las variables en el yacimiento.
    - Realizar estadísticas básicas acerca de las variables.
2. Preparar la información: conciliar las diferentes bases de datos de las que se dispone, en una sola que contenga todas las características de manera de facilitar su investigación y análisis.
    - Estudiar las estadísticas básicas de la nueva base de datos creada, con el fin de encontrar posibles cambios con respecto a los valores iniciales.
    - Estudio exploratorio según tipo de alteración, lo cual correspondería a una etapa primordial dentro del ejercicio, ya que unidades demasiado parecidas, podrían no ser discriminadas adecuadamente.
3. Utilizar herramientas de análisis estadístico uni- y multi- variable con la finalidad de encontrar un agrupamiento en la información y poder dividir en poblaciones.
    - Aplicar algoritmos de selección de variables tratando de encontrar una relación entre ellas y entender cómo se relacionan.
    - Estudiar variables seleccionadas con respecto a la alteración.
    - Utilizar distintos algoritmos de clasificación de variables en base a la selección y estudio de estas.

4. Análisis: analizar la información y sus resultados buscando posibles implicancias de la metodología en la mina, concluyendo sobre los patrones encontrados.

Teniendo en cuenta lo antes expuesto, se procederá a explicar cada etapa de la metodología, indicando sus características y detallando cómo se realizaron.

3.1. Base de datos Iniciales

Para la realización del proyecto se ha trabajado con tres bases de datos diferentes, estas son:

3.1.1. Geo-química (Agua Regia)

Cuenta con 15,408 datos descritos por 50 variables (elementos químicos) obtenidos a través del proceso analítico a base de agua regia. La tabla 3.1 muestra las variables acompañado de sus estadísticas básicas.

En la base de datos todos los elementos están completamente informados, a excepción del oro el cual tiene 8257 datos ausentes (-99).

La distribución de los largos de los sondajes se puede apreciar en la figure 3.1, donde se puede ver que en su mayoría tienen un largo de 14 y 16 metros.

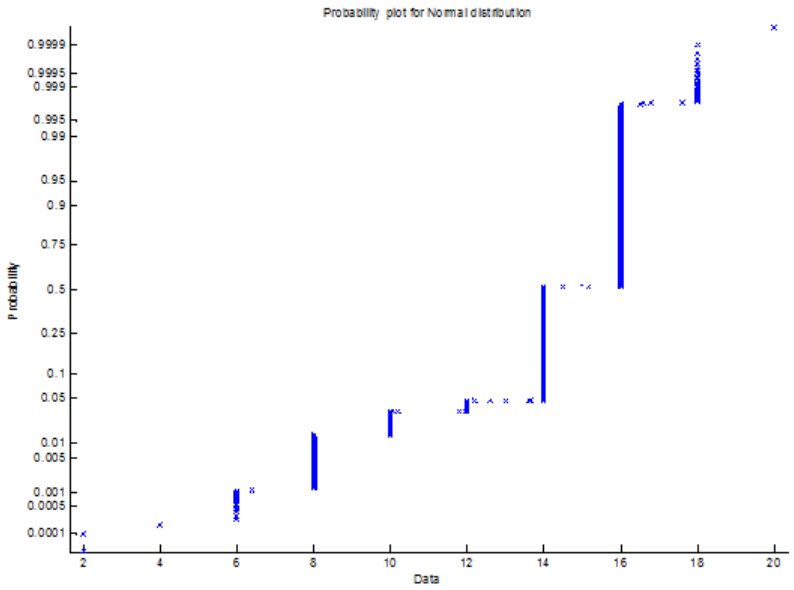


Figura 3.1: Probability plot para largo sondaje Geo-química

Tabla 3.1: Estadísticas básicas geoquímica

Elemento	Media	Varianza de la muestra	Desviación estándar	Min	Max	Datos validos
au ppm	0.21	0.01	0.07	0.2	4.5	7151
ag ppm	3.04	13.33	3.65	0.01	83.1	15408
mo ppm	89.84	13314.55	115.39	0.1	2890	15408
re ppm	0.52	0.96	0.98	0	25.5	15408
cd ppm	4.15	184.45	13.58	0.01	601	15408
pb ppm	71.38	64170.11	253.32	0.2	10000	15408
zn ppm	503.34	857856.21	926.21	2	10000	15408
te ppm	0.75	3.84	1.96	0.01	68.3	15408
bi ppm	1.21	3.42	1.85	0.01	59.1	15408
sb ppm	2.09	119.55	10.93	0.05	472	15408
hg ppm	0.03	0.01	0.09	0.01	6.87	15408
co ppm	12.39	94.86	9.74	0.1	349	15408
ni ppm	23.25	2366.67	48.65	0.2	785	15408
se ppm	3.35	2.79	1.67	0.2	26.5	15408
al porc	0.93	0.45	0.67	0.06	4.7	15408
as ppm	41.57	28467.83	168.72	0.1	7390	15408
b ppm	10.11	1.7	1.3	10	70	15408
ba ppm	40.86	547.18	23.39	0.2	270	15408
be ppm	0.25	0.04	0.2	0.05	1.76	15408
ca porc	0.26	0.28	0.53	0.01	7.41	15408
ce ppm	16.48	74.69	8.64	0.02	51.9	15408
cr ppm	106.69	9877.56	99.39	1	1560	15408
cs ppm	1.13	0.46	0.68	0.05	15.5	15408
cu ppm	5948.84	7997433.8	2827.97	23	10000	15408
ga ppm	2.48	4.15	2.04	0.05	13.35	15408
ge ppm	0.06	0	0.03	0.05	1.86	15408
hf ppm	0.02	0	0.01	0.02	0.22	15408
in ppm	0.27	0.2	0.45	0.01	16.1	15408
k porc	0.23	0.01	0.1	0.01	1.61	15408
la ppm	8.07	17.27	4.16	0.2	26.7	15408
li ppm	5.07	45.9	6.78	0	39	15408
mg porc	0.41	0.32	0.56	0.01	3.13	15408
mn ppm	264.45	191664.22	437.79	5	7830	15408
na porc	0.06	0	0.04	0.01	0.5	15408
nb ppm	0.06	0	0.03	0.05	0.53	15408
p ppm	437.97	215128.77	463.82	10	4420	15408
rb ppm	12.64	43.36	6.58	0.1	109	15408
s porc	2.26	2.73	1.65	0.01	10	15408
sc ppm	1.64	3.68	1.92	0.1	21.5	15408
sn ppm	0.81	3.55	1.88	0.2	101	15408
sr ppm	91.72	10987.56	104.82	0.2	1240	15408
ta ppm	0.01	0	0	0.01	0.26	15408
th ppm	1.74	1.22	1.1	0.2	37.3	15408
ti ppm	0.01	0	0.02	0.01	0.32	15408
tl ppm	0.18	0.01	0.11	0.02	2.06	15408
u ppm	0.42	0.1	0.32	0.05	8.35	15408
v ppm	17.11	551.05	23.47	1	196	15408
w ppm	1.05	59.48	7.71	0.05	540	15408
y ppm	4.68	31.91	5.65	0.05	83.9	15408
zr ppm	1	0.01	0.1	1	7	15408

3.1.2. Leyes

Contiene 239,296 datos descritos por CuT, CuS, Fe, As. La tabla 3.2 muestra los estadísticos básicos de estas variables.

Es posible apreciar cómo el cobre total tiene la mayor cantidad de valores válidos y el arsénico la menor cantidad.

En la figura 3.2 se puede apreciar la distribución de largos de los sondajes para esta base de datos, en donde se puede ver que la mayoría de las mediciones son de dos metros.

Tabla 3.2: Estadísticas básicos: leyes CuT, CuS, As y Fe

Elemento	Media	Varianza	Desviación estándar	Min	Max	Datos Validos
CuT porc	0.66	0.83	0.91	0	41.32	234729
CuS porc	0.08	0.1	0.31	0	16.9	232803
Fe porc	2.7	4.98	2.23	0	93.3	216991
As ppm	46.81	90318.2	300.53	1	47429	196033

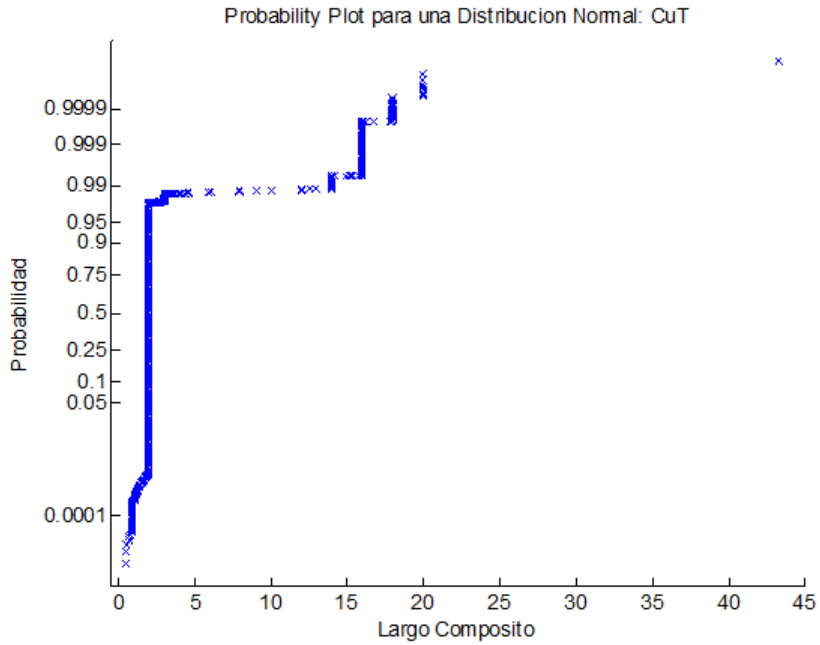


Figura 3.2: Probability plot para largo sondaje de leyes

3.1.3. Geología

Esta base de datos cuenta con 35,934 datos descritos por tres variables categóricas, las cuales son litología, alteración y min zone. Se trabajará solo utilizando la alteración de la que se busca crear un modelo clasificatorio, dejando la litología y la min zone aparte al tratarse de variables cualitativas. En la tabla 3.3 se aprecian las variables contenidas y sus estadísticas.

La figura 3.3 muestra cómo se encuentran distribuidos los largos de los sondajes para esta base de datos.

La nomenclatura de las alteraciones se pueden apreciar en la tabla 3.4. En ella se destaca en “negrita” las alteraciones que presentan una mayor cantidad de muestras en la base de datos.

Tabla 3.3: Base de datos de geología

Unidad Geológica	Moda 1	Moda 2	N° categorías presentes	Min	Max	Datos Validos
litología	50	31	25	1	73	23496
mine zone	70	10	26	10	70	23130
alteración	50	51	19	10	80	23385

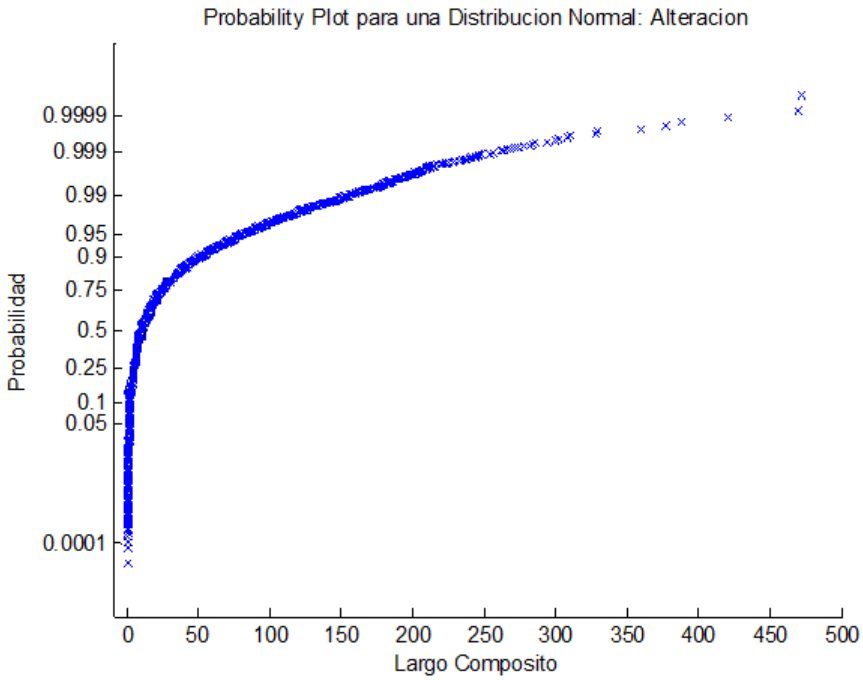


Figura 3.3: Probability plot para largo sondaje según alteración

3.2. Creación base de datos

Para la unión de las bases de datos se desarrolló un programa que busca el mismo intervalo de un sondaje en las mediciones de otra base de datos, indicando el porcentaje de ocurrencia. Para el caso de variables categóricas, se indica las dos ocurrencias mayores y para las variables continuas, un promedio ponderado por su porcentaje de ocurrencia.

En la figura 3.4 se ejemplifica cómo se realizó la creación de la nueva base de datos. Si se tiene el mismo sondaje presente en dos bases de datos el programa busca el intervalo del sondaje de la base de datos 1 en la base de datos 2. En este caso utilizaremos el intervalo 1.1 como referencia. Si las variables de la base de datos 2 son del tipo categóricas este le asignará el porcentaje que le corresponde de 2.1, 2.2 y 2.3 en el intervalo que se estudia, creando una nueva base de datos con la información de la primera, añadiéndole las dos cualidades de ocurrencia mayor indicando también este porcentaje. En el caso de ser una variable continua, se calculará un promedio ponderado por el porcentaje de pertenencia, indicando también la

Tabla 3.4: Codificación de alteraciones en Escondida

Alteración	Abreviación	Código
Fresco: sin Alteración	F	10
Propilítico	P	20
Potásico (Sólo Bt secundaria)	***K1***	***30***
Feld K > Bt	***K2***	***31***
Feld K < Bt	K3	32
(Bt sec +/- Feld K) > Cl	K4	33
Clorita-Sericita-Arcillas	***SCC1***	***40***
Clorita-Sericita-Cuarzo	***SCC2***	***41***
Clorita-Biotita +/- Feld K	SCC3	42
Clorita	SCC4	43
Sericita-Cuarzo	***S1***	***51***
Sericita-Cuarzo-Arcilla	***S2***	***52***
Alteracione Blancas (Histórica)	***S-***	***50***
Histórica	S-	53
Sericita Gris Verde	S4	54
Argilización Supérgena	***AA***	***61***
Argilización Avanzada	AAV	62
Silicificación	Q	70
albitización	alb	81
actinolita/epidota	acep	82

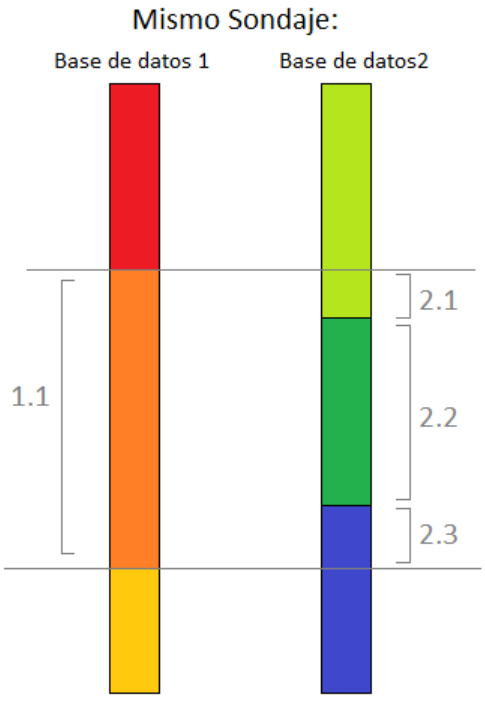


Figura 3.4: Cruce de bases de datos

proporción (largo del sondaje) que se encuentra de él.

De esta manera la base de datos con la que se iniciará la investigación está constituida por 54 variables continuas (leyes y geo-química) y una categórica (alteración).

3.3. Tratamiento base de datos

Analizando la base de datos se concluyó que el oro como variable no tiene mayor importancia en el análisis debido a la falta de valores que presentaba (-99), teniendo solo alrededor de siete mil datos válidos (menos de la mitad del total de la base de datos y casi sin variabilidad) por lo cual fue descartado. Se eliminaron siete variables por presentar en su mayoría el mismo valor (ver en anexos A los histogramas), lo cual es asociado al límite de detección del mecanismo de análisis. Estas son Zr, Nb, Ge, B, Ta, Ti, Hg, Hf.

Se encontraron pares de variables con una alta correlación (mayor a noventa y cuatro por ciento) [23], eliminando por co-linealidad una de ellas para que no tengan una doble importancia dentro del estudio. Los pares de variables con alta correlación son:

- Li – Mg (94 %de correlación) eliminando Li.
- Ga-Mg (94 % de correlación) eliminado Ga.
- Sc – V (94 % de correlación) eliminando V.
- Ce - La (98 % correlación) eliminando La.

El magnesio se mantuvo debido a su presencia en importantes minerales de alteración como la biotita y clorita, el escandio debido a su potencial de ionización parecido a este último [24], y el cerio y el lantano son elementos muy parecidos entre ellos (ambos lantánidos uno al lado del otro) comportándose de la misma manera existiendo la posibilidad de trabajar con cualquiera, seleccionando el cerio finalmente.

De esta manera, hasta este punto, se eliminó en total 12 variables ruidosas dentro del estudio.

3.4. Adición de variables sintéticas

Se utilizaron 15 variables sintéticas tomadas a partir del trabajo de GeoAV Geochemistry & Exploration S.A. [25], las que son utilizadas en la litogeoquímica, a partir de la información

contenida en la base de datos. La idea es utilizar las operaciones aritméticas para realzar las intensidades de las distintas alteraciones, donde la multiplicación trata de realzar los elementos cuando estos se encuentren en una elevada concentración y deprimirlos cuando estén en una baja. La división nos permite conocer cómo se mueve un elemento con respecto a otro, es decir, muestra el cociente de incremento de un elemento con respecto al denominador. La suma y la resta dejan comprender composiciones de elementos.

Las variables utilizadas son las siguientes:

Alteración fílica:

- KxAl : Esta variable sintética trata de realzar la aparición de sericita/moscovita en el yacimiento al contener estos elementos en su estructura $(KAl_2(Si_3Al)O_{10}(OH,F)_2)$
- (KxNa)/Al: En la estructura de la moscovita al tratarse de una mica, quedan átomos de sodio atrapados en las intercapas, además, el sodio puede entrar por intercambio catiónico en esta estructura por el potasio, la división por aluminio es para estudiar cómo varía según la concentración de este elemento, presente en importantes minerales de distintas alteraciones como biotita y la alunita.
- Na/Al: Trata de estudiar el sodio (visto en la variable anterior) con relación al aluminio
- (Al+K)/ (Na+Ca+Mg): Destacan el proceso de intercambio entre los cationes aluminio y potasio con el grupo de cationes sodio, calcio y magnesio, los cuales se comportan de manera semejante.
- (Al+K+Na)/(Ca+Mg): Destacan el proceso de intercambio entre los cationes aluminio, potasio y sodio con el grupo de cationes calcio y magnesio, los que se comportan de manera semejante.

Alteración propilítica:

- (Ca+Na)/(K+Al): Destacan el proceso de intercambio entre los cationes de calcio y sodio con el grupo de cationes de potasio y aluminio, los cuales se comportan de manera semejante.

Alteración argílica:

- Al/(Na+Ca+K): Se espera que en esta alteración exista un enriquecimiento de aluminio debido a que está presente en arcillas sobre el contenido de potasio presente en mayor cantidad en otras alteraciones, más los cationes intercambiables sodio y calcio.

- Al/Mg: Se espera que en esta alteración exista un enriquecimiento de aluminio debido a que está contenido en arcillas características de esta alteración, con respecto al magnesio, el que se encuentra en mayor cantidad en alteraciones cloríticas y biotititas.
- 3*Al/(K+Na): se espera que en esta alteración exista un enriquecimiento de aluminio debido a que está contenido en arcillas sobre la concentración de potasio, el que se encuentra en minerales como biotita, además, el componente (K+Na) tiene relación con la sericita.

Alteración argílica avanzada:

- (K+Al+S)/(Fe+S): Busca resaltar la alunita $(Na, K)Al_3(SO_4)_2(OH)_6$ presente en esta alteración en pro de la pirita FeS_2.
- (CuxAsxSbxS)/Fe: Busca realzar la mineralización con respecto al Fe como lo son las sulfosales, tales como eneragita tenantita $(Cu_3(As, Sb)S_4/(Cu, Ag, Fe, Zn)_{12}As_4S_{13})$.
- (CuxAsxSbxS/Fe)x(Al+K+S): Es igual a la anterior pero incluye un factor que hace referencia a la alunita $(Na, K)Al_3(SO_4)_2(OH)_6$.

Alteración potásica:

- K/(Ca+Na): Trata de resaltar el potasio contra el sodio y el calcio, ya que un aumento en potasio se debería a una baja de estos últimos dos, debido a que son cationes intercambiables.
- K/Mg: Trata de resaltar la sericita sobre la clorita.

Supergéna:

- Mg/Al y Mn/Al: en las alteraciones supérgenas se espera que exista una mayor cantidad de magnesio y manganeso (los cuales tienen una distribución parecida), buscando diferenciar alteraciones más superficiales (supérgenas).

3.5. Unión de base de datos

Con las bases de datos unidas, según lo descrito anteriormente, más la adición de las variables sintéticas ya explicadas, se construyó la base de datos final con la cual se trabajará. En la figura 3.5 se puede ver cómo se encuentran los largos de los compósitos distribuidos en

la base de datos unida, permaneciendo igual que la base de datos geo-química sin grandes cambios.

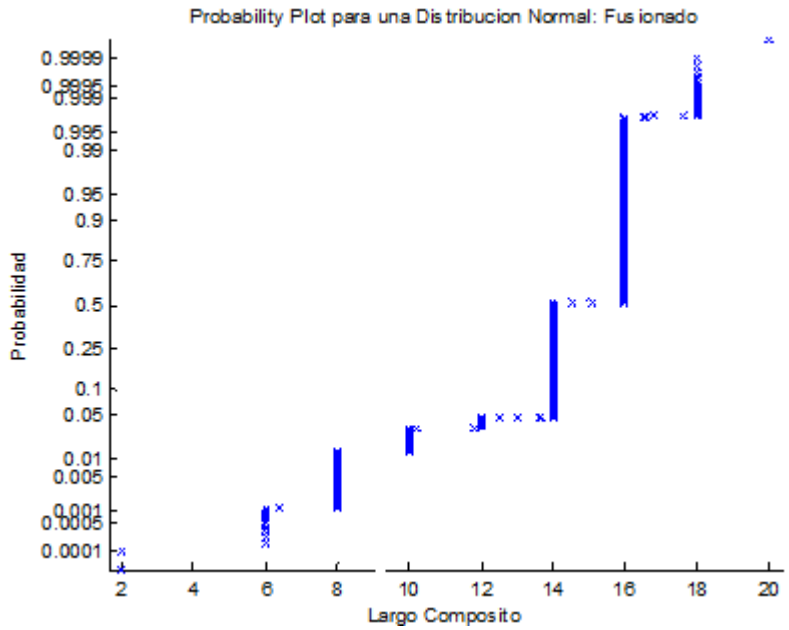


Figura 3.5: Probability plot largo de compósitos

Para estudiar la pertenencia de los sondajes de leyes (CuT, Cus, Fe) dentro de la base de datos de geo-química, se realizó la figura 3.6 sobre la base de datos fusionada en donde casi la totalidad de los datos se encuentran totalmente contenidos(La figura se origina ocupando la variable creada al unir las bases de datos señalando el porcentaje del segmento que está contenido en la otra base de datos como se ve en la figura 3.4).

Para estudiar la pertenencia de los sondajes con respecto a la geología, al igual que en la imagen anterior, se realizó la figura 3.7, donde un 70 % de los datos se encuentran totalmente contenidos tomando solo los que tienen una ocurrencia mayor o igual al 70 % (que el tramo contenga al menos un 70 % de la alteración mayor) contenido en esta base de datos.

En la figura 3.8 se puede apreciar cómo están distribuidas dentro de la base de datos. Las alteraciones se analizarán más adelante dentro del estudio.

Se realizó un filtro de datos para trabajar solo con los datos que poseen al menos un 70 % de pertenecía de una cierta alteración (se escogió este valor para no quedar con muy pocos datos en alteraciones de las cuales ya se tenían bajas mediciones), quedando alrededor de trece mil seiscientos datos. El resto de ellos serán dejados de lado debido a que no aportarían información confiable para la generación y validación de un modelo de clasificación producto de sobreimposiciones y/o confusiones que podrían producirse en estos tramos, por lo cual podrían inducir a errores.

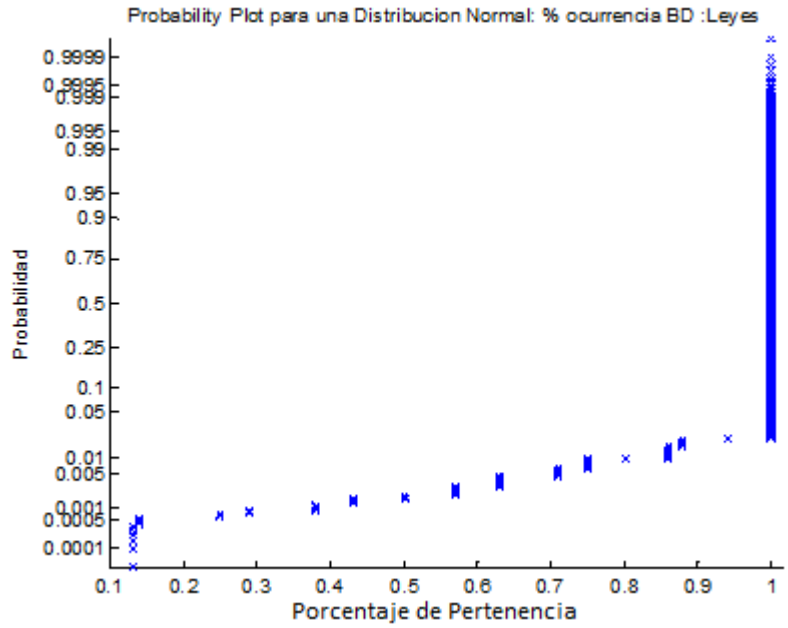


Figura 3.6: Probability plot para porcentaje de pertenencia de CuT, CuS y Fe

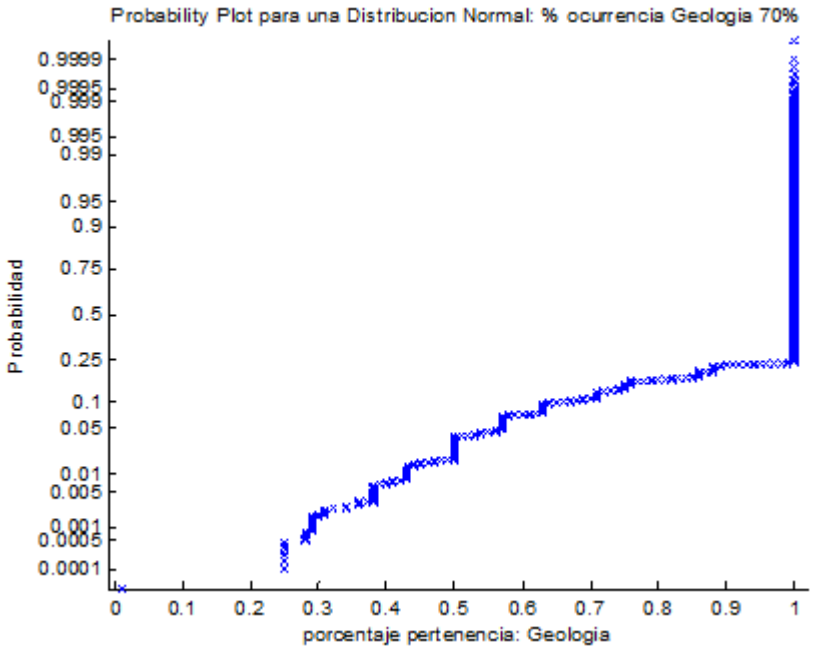


Figura 3.7: Probability plot para porcentaje de ocurrencia de geología

Estudiando nuevamente la correlación en la base de datos se encontraron elementos con alta similitud, los cuales son:

- Al+K/(Na+Ca+Mg) y (Al+K+Na)/(Ca+Mg) con correlación del 95 % eliminando (Al+K+Na)/(Ca+Mg).
- (CuxAsxSbxS/Fe)x(Al+K+S) y (CuxAsxSbxS/Fe) con correlación del 99 % eliminando [CuxAsxSbxS/Fe]x[Al+K+S].

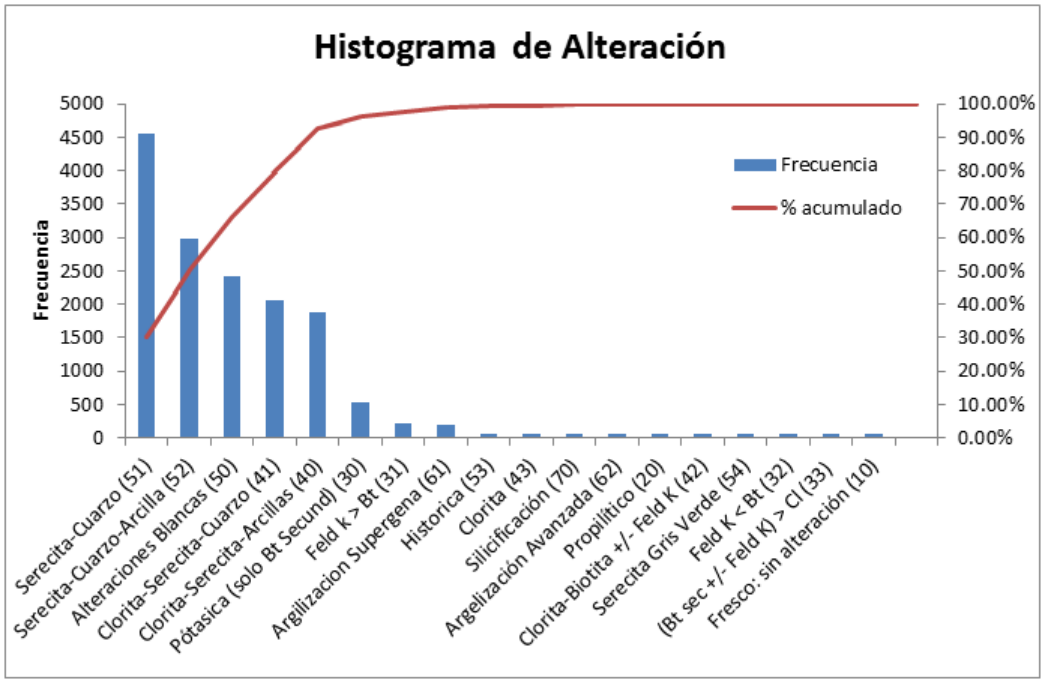


Figura 3.8: Histograma alteraciones

- As (geo-química) y As(leyes) con correlación del 94 % se decidió eliminar As(leyes) y volver a repetir la unión, ya que esta variable contenía menos datos válidos, y así tener una mayor cantidad de datos con CuT, CuS y Fe.

3.6. Estudio base de datos final

Para entender cómo se desenvuelven las variables en la zona a estudiar, se volvió a realizar un histograma de las alteraciones según los datos que se tienen después de filtrar. En la figura 3.9, se puede apreciar cómo la zona de estudio está dominada por alteraciones tipo 50 (Sericita-Cuarzo o fílica) con cerca de un 70 % de los datos, seguida por las tipo 40 (transición fílica – potásica) con alrededor de un 25 %, un 2.5 % restante para alteraciones 30 (potásica: biotita) y 61 (argílica supergena) con la 31 (potásica: feld K > biotita) con alrededor de un 1 % para cada una de estas.

Como la alteración 50 (alteraciones blancas) ya no es válida (actualmente no se mapea), dentro de la base de datos se decidió dejarla aparte dentro de lo que es la investigación, ya que en ella se combinaban las alteraciones sericita-cuarzo (51,52) con la argílica supérgena (61), lo cual aportaría ruido al análisis. Por esta razón, se trabajará con las alteraciones: 51, 52, 41, 40, 30, 31 y 61, debido a la poca ocurrencia de las alteraciones restantes.

Para estudiar la distribución de las variables se realizaron histogramas y boxplots para

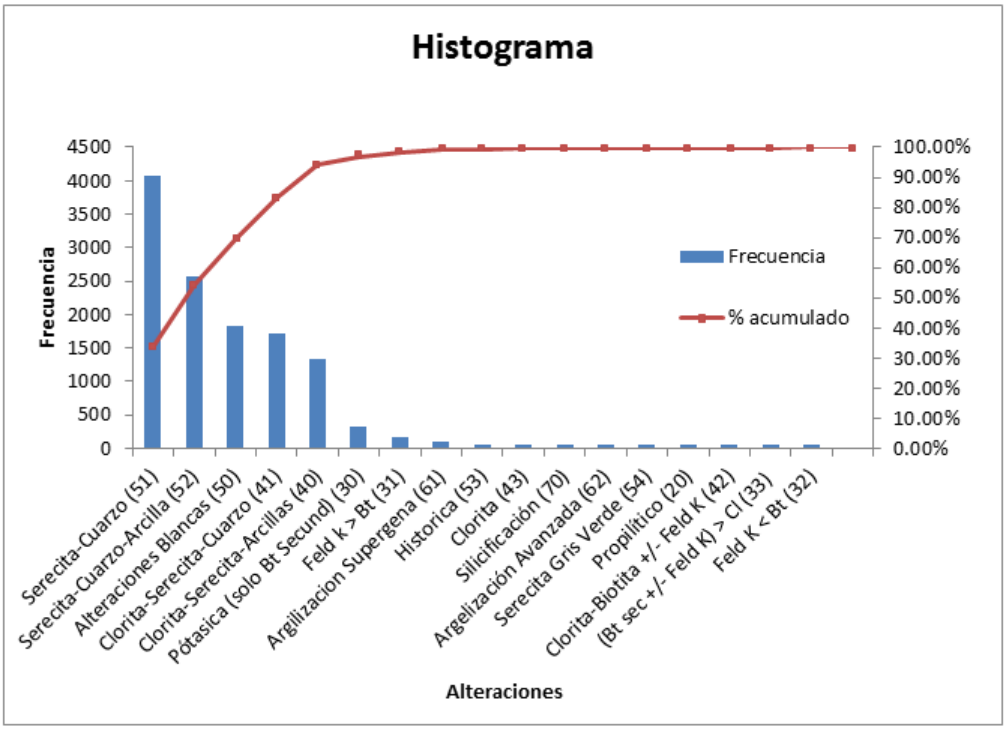


Figura 3.9: Histograma de Alteraciones

cada una de ellas según alteración, en la figura 3.10 se aprecian elementos de interés, como el cobre total, soluble y el fierro. De esta se desprende que estos elementos no son muy buenos clasificadores de alteraciones.

No es así como otros elementos que a simple vista demuestran ser buenos clasificadores como es el caso del Mg, Sc, Mn, Al, entre otros, que pueden ser vistos en las figuras 3.11 y 3.12. Los cuales al estar contenidos en los minerales de ganga (micas y arcillas) se encuentran en menor o mayor cantidad al momento de su digestión en agua regia, los que definen mejor a una alteración.

Para el desarrollo de un modelo de clasificación que logre diferenciar cada tipo de alteración es necesario estudiar el comportamiento de variables en conjuntos, lo cual se hace muy difícil debido a la gran cantidad de estos que pueden formarse, por lo que es necesaria la utilización de sistemas automáticos que ayuden al investigador a realizar el trabajo. Estos tópicos serán tratados más adelante.

3.7. Separación de la base de datos

Para llevar a cabo la investigación se dividirá la base de datos en tres partes que serán utilizadas para entrenar, afinar y validar el modelo.

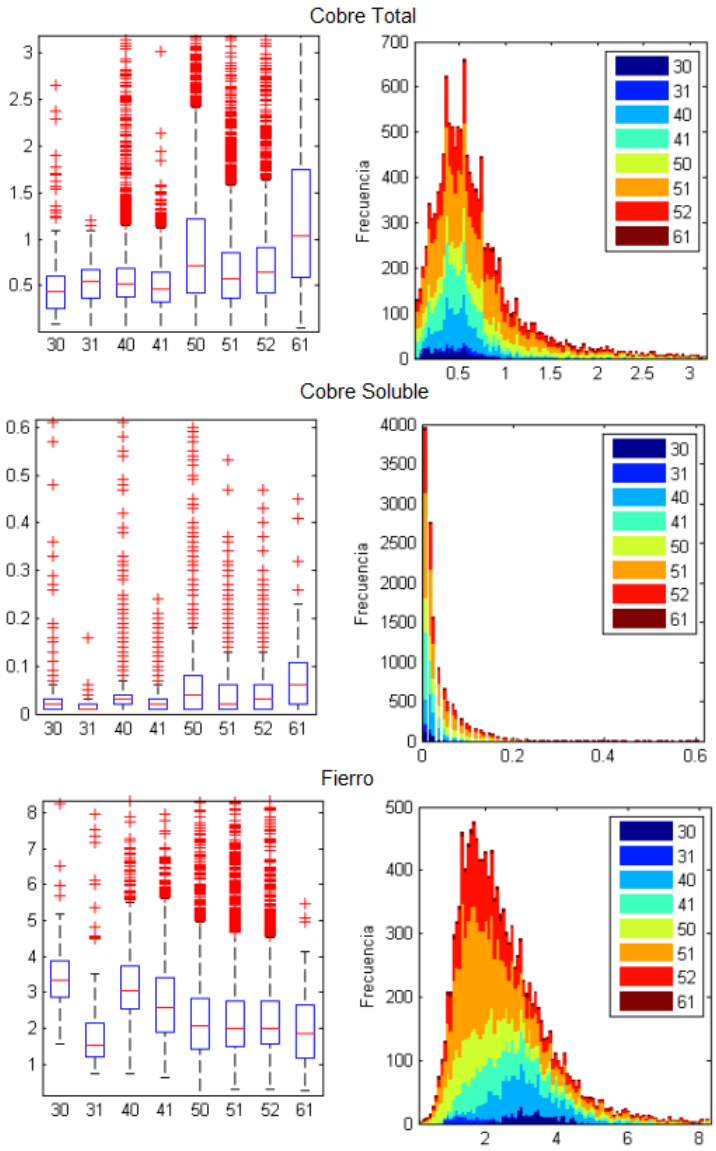


Figura 3.10: Distribución Cobre total, soluble y fierro

- **Entrenamiento:** Para la confección del modelo se dividió la base de datos según alteración y se tomaron sondajes aleatoriamente hasta completar un mínimo de 80 datos por alteración a estudiar. Con el objetivo de aumentar la dificultad de clasificación, y así extrapolar la información contenida por sondaje y no por dato unitario, este conjunto de datos representa un 4 % de la data, su finalidad es la de crear los modelos de clasificación en base a su información.
- **Ajuste:** Compuesta por un 17 % de los datos tomados al azar de la base de datos. Este conjunto de datos son utilizados para estudiar el error que produce el modelo, buscando que no exista sobreajuste en él.
- **Validación:** Para validar el modelo se tiene el 79 % de información, con el fin de corroborar el modelo y analizar si tiene un buen desempeño ante información nueva.

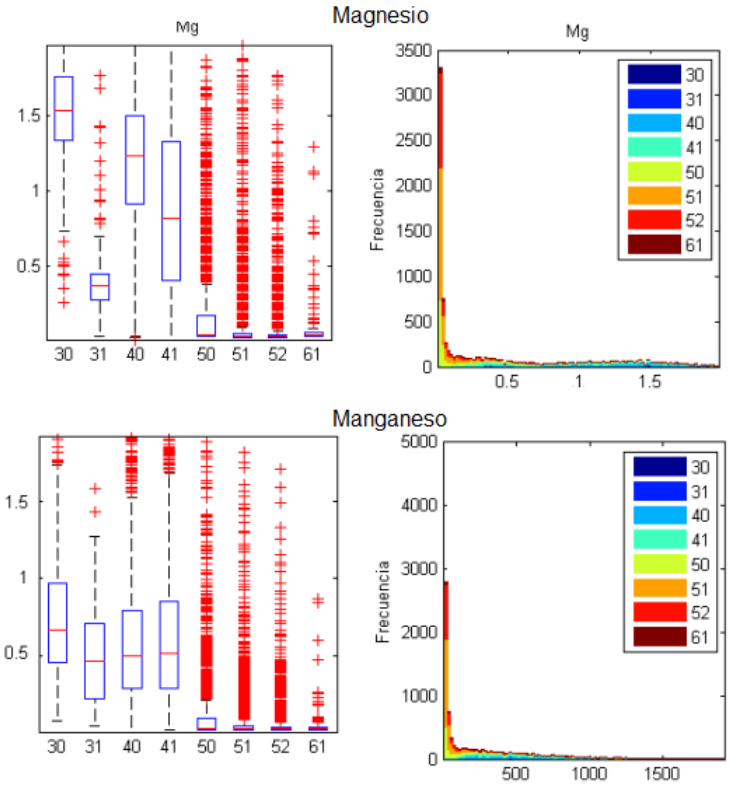


Figura 3.11: Distribución magnesio y manganeso

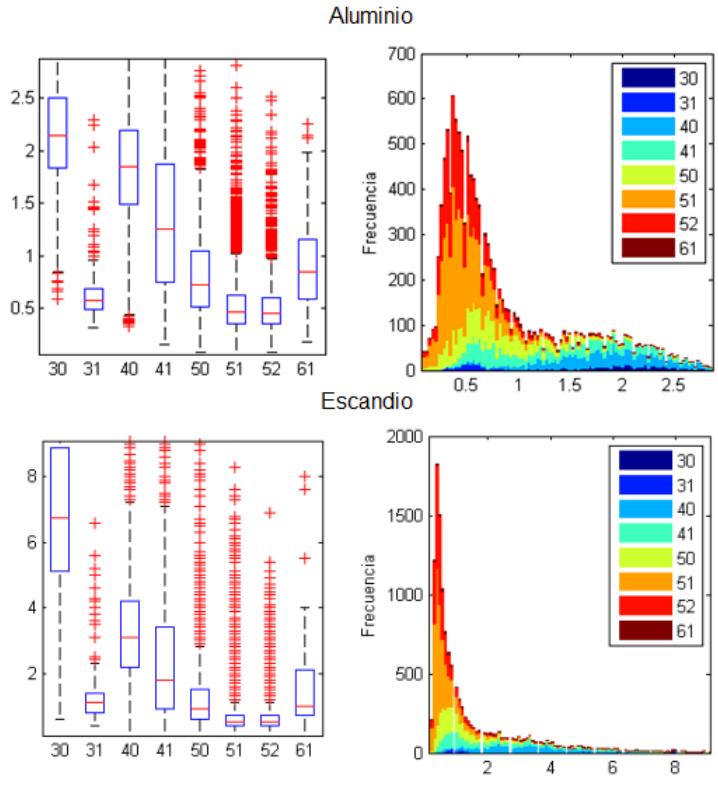


Figura 3.12: Distribución escandio y aluminio

Este conjunto de datos busca la validación del modelo, utilizándola para conocer el error que tendría el modelo al ser enfrentado contra información nueva.

3.8. Herramientas de análisis

Se han creado las siguientes herramientas de análisis para facilitar el estudio de las variables y tomar decisiones en base a ellas.

3.8.1. Traslape de poblaciones

Para la realización del tema, se creó la función "traslape de poblaciones", a partir de la idea de representar numéricamente un boxplot, señalando el porcentaje de superposición que existe entre dos poblaciones de datos sobre una misma variable, indicando la proporción de los datos a estudiar (a través de cuantiles). En figura 3.13, se pueden ver dos categorías para una misma variable, denotado por x el porcentaje de datos con el cual se desea trabajar. Los cuantiles que contendrían dicha cantidad según clase serían:

$$\text{Cuantil Superior}_{\text{Clase i}} = \frac{x+1}{2} \tag{3.1}$$

$$\text{Cuantil Inferior}_{\text{Clase i}} = \frac{1-x}{2} \tag{3.2}$$

De esta manera, es posible conocer el porcentaje de traslape de una familia con respecto a la otra. En el caso de que la otra familia esté totalmente contenida dentro de la distancia de estudio, el valor que tomará será uno. De igual manera si la familia de estudio se encuentra totalmente contenida dentro de la otra clase, el resultado sería el mismo.

Para ejemplificar el traslape de poblaciones se tomará como ejemplo un boxplot, en donde los límites están dados por el cuantil 25 % y 75 %, conteniendo el 50 % de la información, valor que sería un input del método. Al estudiar la clase uno en comparación con la clase dos, podemos ver que existe un traslape indicado por la zona verde cuyo valor porcentual será el entregado a través de esta función.

Realizando la función para cada una de las variables tomando un 75 % de los datos

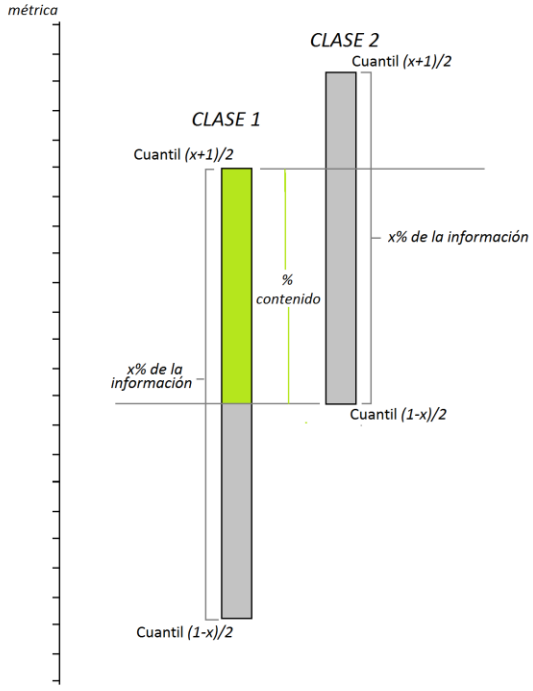


Figura 3.13: Traslape de poblaciones

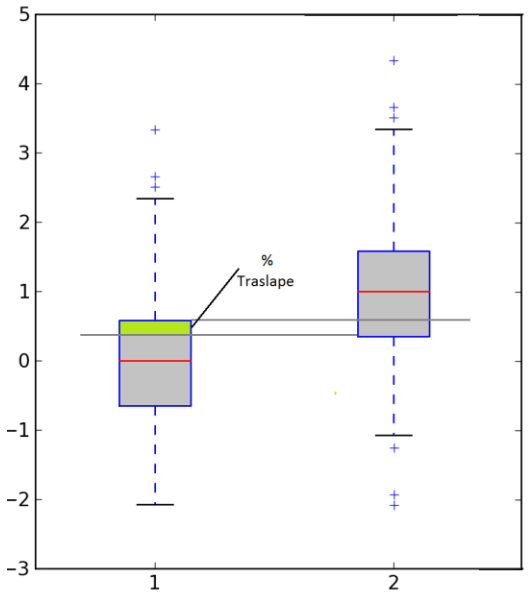


Figura 3.14: Traslape boxplot

(cuantil 12.5 % y 87.5 %) se obtienen resultados como los expuestos en la tabla 3.5. Estos resultados son un ejemplo del traslape que existe entre elementos de la alteración 30(potásica: biotita secund.) con las demás (todas las tablas de traslape se encuentran en el Anexo B: métodos de análisis). En ella es posible apreciar que elementos como el aluminio y el magnesio (resaltados) son buenos clasificadores, teniendo un traslape de 0 % con las alteraciones 31, 50, 51 y 52, significando que solo utilizando una de estas variables es posible

una separación de a lo menos el 75 % de los datos. Además son buenos clasificadores para diferenciar esta alteración con la 61, pero no son buenos separadores para diferenciar la alteración 30 de la 40 y 41.

Tabla 3.5: Valores traslape ejemplo alteración 30

Elemento (alt. 30)	31	40	41	50	51	52	61
Ag	97 %	97 %	98 %	100 %	100 %	100 %	100 %
Al	0 %	69 %	54 %	0 %	0 %	0 %	10 %
As	100 %	96 %	98 %	82 %	76 %	64 %	82 %
Ba	46 %	30 %	45 %	30 %	30 %	30 %	17 %
Be	39 %	100 %	100 %	22 %	10 %	4 %	22 %
Bi	81 %	100 %	98 %	99 %	98 %	96 %	100 %
Ca	100 %	23 %	100 %	0 %	0 %	0 %	0 %
Cd	100 %	97 %	100 %	100 %	89 %	100 %	100 %
Ce	100 %	100 %	100 %	56 %	50 %	50 %	54 %
Co	12 %	100 %	69 %	97 %	73 %	88 %	100 %
Cr	88 %	77 %	100 %	100 %	100 %	100 %	88 %
Cs	48 %	71 %	83 %	35 %	28 %	35 %	54 %
Cu	90 %	83 %	95 %	91 %	90 %	82 %	77 %
In	95 %	96 %	94 %	92 %	92 %	90 %	85 %
K	25 %	31 %	21 %	26 %	2 %	0 %	12 %
Mg	0 %	73 %	65 %	0 %	0 %	0 %	0 %
Mn	61 %	85 %	95 %	1 %	0 %	0 %	0 %
Mo	69 %	100 %	93 %	100 %	92 %	100 %	97 %
Na	31 %	37 %	20 %	7 %	0 %	0 %	0 %
Ni	48 %	75 %	29 %	79 %	22 %	25 %	23 %
P	0 %	100 %	100 %	6 %	0 %	0 %	12 %
Pb	100 %	99 %	100 %	53 %	22 %	18 %	53 %
Rb	17 %	23 %	21 %	6 %	0 %	0 %	5 %
Re	52 %	94 %	97 %	100 %	100 %	97 %	100 %
S	86 %	85 %	84 %	92 %	57 %	50 %	100 %
Sb	100 %	92 %	100 %	100 %	100 %	100 %	100 %
Sc	0 %	28 %	19 %	0 %	0 %	0 %	0 %
Se	79 %	100 %	85 %	76 %	79 %	73 %	76 %
Sn	75 %	50 %	63 %	100 %	100 %	100 %	100 %
Sr	100 %	98 %	100 %	95 %	92 %	91 %	89 %
Te	90 %	95 %	97 %	82 %	87 %	85 %	85 %
Th	86 %	100 %	100 %	43 %	43 %	36 %	43 %
Tl	22 %	54 %	27 %	100 %	76 %	100 %	100 %
U	100 %	100 %	96 %	100 %	100 %	100 %	100 %
W	100 %	100 %	93 %	87 %	91 %	92 %	97 %
Y	17 %	100 %	85 %	32 %	0 %	0 %	0 %
Zn	100 %	98 %	100 %	52 %	28 %	21 %	28 %
CuT	90 %	84 %	95 %	93 %	90 %	83 %	78 %
CuS	100 %	100 %	100 %	100 %	100 %	100 %	100 %
Fe	85 %	100 %	91 %	87 %	79 %	71 %	54 %
KxAl	58 %	57 %	100 %	71 %	41 %	44 %	89 %
KxNA/Al	84 %	28 %	30 %	58 %	36 %	35 %	23 %
Na/Al	58 %	81 %	100 %	100 %	87 %	78 %	100 %
(Al+K)/(Na+Ca+Mg)	100 %	74 %	100 %	14 %	0 %	0 %	0 %
(Ca+Na)/(K+Al)	94 %	48 %	100 %	55 %	100 %	69 %	16 %
Al/(Na+Ca+K)	53 %	64 %	100 %	100 %	96 %	97 %	77 %
3Al/(K+Na)	21 %	72 %	100 %	65 %	21 %	24 %	99 %
(K+Al+S)/(Fe+S)	69 %	75 %	58 %	100 %	48 %	44 %	100 %
CuxAsxSbxS)/Fe	99 %	99 %	99 %	95 %	91 %	85 %	96 %
K/(Ca+Na)	100 %	94 %	100 %	67 %	72 %	62 %	56 %
K/Mg	63 %	92 %	100 %	19 %	0 %	0 %	0 %
Mg/Al	83 %	80 %	83 %	16 %	0 %	0 %	0 %
Mn/Al	95 %	100 %	98 %	37 %	15 %	4 %	0 %
Promedio	**68 %**	**79 %**	**83 %**	**63 %**	**53 %**	**51 %**	**56 %**

Cada clase tiene su matriz como la antes vista, en donde se indica el porcentaje de traslape por elemento que tendrían con respecto a las otras (las tablas correspondientes a las demás alteraciones se encuentran adjuntas en anexos B). De esta manera, si se toma el promedio de los traslapes por variable con cada una de las familias, es posible tener un valor traslape promedio de una clase con respecto a las otras, además es posible modificar la cantidad de variables con la cual se calcule el promedio, indicándole que solo tome una cantidad n de variables que presenten el menor traslape, esto significaría usar variables útiles

para diferenciar alteración.

3.8.2. Clustering de categorías

Como se puede apreciar el dominio del estadístico anterior se encuentra entre [0 1]. Cero en el caso de que no exista superposición y uno, si la población está contenida en la otra. Recordando la matriz de correlación, la cual tiene un dominio [-1 1], existe una metodología para el estudio de variables que permite ver gráficamente cómo éstas se van agrupando en función de sus semejanzas, el dendrograma, definiéndose como una representación gráfica de un clustering de variables. Esta matriz, inicialmente no sería simétrica producto que el traslape que tiene una familia con respecto a la otra y viceversa no son porcentualmente iguales. En la tabla 3.6, se muestra la matriz de traslape original que se produce al someter a este técnica a las alteraciones que se estudian.

Tabla 3.6: Matriz traslape

	30	**31**	**40**	**41**	**50**	**51**	**52**	**61**
30	1.00	0.68	0.79	0.83	0.63	0.53	0.51	0.56
31	0.59	1.00	0.69	0.85	0.77	0.67	0.64	0.66
40	0.75	0.71	1.00	0.88	0.71	0.59	0.57	0.66
41	0.77	0.84	0.86	1.00	0.74	0.63	0.61	0.67
50	0.53	0.80	0.65	0.72	1.00	0.84	0.87	0.92
51	0.46	0.71	0.56	0.64	0.95	1.00	0.97	0.86
52	0.44	0.67	0.55	0.61	0.94	0.95	1.00	0.86
61	0.50	0.68	0.63	0.67	0.93	0.80	0.80	1.00

Para solucionar el problema de que la matriz no es simétrica, el valor que tomará el traslape final de dos familias será el promedio de ellas, es decir, si el traslape de la alteración 30 con la 31 es 59 %, y el traslape de la 31 con la 30 es de 68 % (resaltados en la tabla 3.6), tomará el valor final en la matriz de traslape simétrica de 63 % .

$$TraslapeFinal_{(i,j),(j,i)} = \frac{Traslape_{i,j} + Traslape_{j,i}}{2} \tag{3.3}$$

Ejemplo Traslape alteración 30 y 31

$$(Traslape_{30,31} = \frac{59\,\% + 68\,\%}{2} \simeq 63\,\%) \tag{3.4}$$

Repitiéndose esto para las demás familias, se obtiene una matriz simétrica, a la que

es posible aplicar el dendrograma y estudiar sus similitudes. Esta se ve representada en la tabla 3.7.

Tabla 3.7: Matriz traslape simétrica

	30	**31**	**40**	**41**	**50**	**51**	**52**	**61**
30	1.00	0.63	0.77	0.80	0.58	0.50	0.48	0.53
31	0.63	1.00	0.70	0.84	0.79	0.69	0.66	0.67
40	0.77	0.70	1.00	0.87	0.68	0.57	0.56	0.64
41	0.80	0.84	0.87	1.00	0.73	0.63	0.61	0.67
50	0.58	0.79	0.68	0.73	1.00	0.90	0.91	0.92
51	0.50	0.69	0.57	0.63	0.90	1.00	0.96	0.83
52	0.48	0.66	0.56	0.61	0.91	0.96	1.00	0.83
61	0.53	0.67	0.64	0.67	0.92	0.83	0.83	1.00

Finalmente para aplicar el dedrograma se debe tomar la medida de disimilitud (proximidad o distancia de un ítem con respecto al otro) asociada a la matriz de traslape, en el clustering de variables se utiliza:

$$Disimilitud(x_i, x_j) = 1 - correlación \tag{3.5}$$

De manera análoga utilizamos el traslape para tener la medida de disimilitud de la muestra como:

$$Disimilitud(x_i, x_j) = 1 - TraslapeFinal_{i,j} = 1 - \frac{Traslape_{i,j} + Traslape_{j,i}}{2} \tag{3.6}$$

Así un valor relativamente grande de traslape, sugiere la idea de que las categorías se encuentran más cerca entre ellas, siendo más parecidas (una disimilitud menor), y a valores más pequeños de traslape ($Traslape \simeq 0$) estas son más diferentes, es decir, que la distancia entre ellas es mayor, por lo cual concuerda con la idea de crear una distancia de semejanza. La matriz de disimilitud realizada para la matriz de traslape se expone en la tabla 3.8.

Al realizar este procedimiento sobre las matrices de traslape se genera la figura 3.15.

Viendo los traslapes de las familias se decidió unir las alteraciones tipo 40 y 41, y las tipo 51 y 52 debido a su semejanza, lo que hace más difícil la división de estas. En la Tabla 7: Traslape 40-41 y 51-52, se puede apreciar la magnitud del traslape entre ellas para el 75 % de la información.

De esta manera desde ahora la fusión entre las alteraciones 40 y 41, serán tratadas como alteraciones tipo 40, y las 51 y 52, como alteraciones tipo 50.

Tabla 3.8: Disimilitud de la matriz traslape

	30	**31**	**40**	**41**	**50**	**51**	**52**	**61**
30	0.00	0.37	0.23	0.20	0.42	0.50	0.52	0.47
31	0.37	0.00	0.30	0.16	0.21	0.31	0.34	0.33
40	0.23	0.30	0.00	0.13	0.32	0.43	0.44	0.36
41	0.20	0.16	0.13	0.00	0.27	0.37	0.39	0.33
50	0.42	0.21	0.32	0.27	0.00	0.10	0.09	0.08
51	0.50	0.31	0.43	0.37	0.10	0.00	0.04	0.17
52	0.52	0.34	0.44	0.39	0.09	0.04	0.00	0.17
61	0.47	0.33	0.36	0.33	0.08	0.17	0.17	0.00

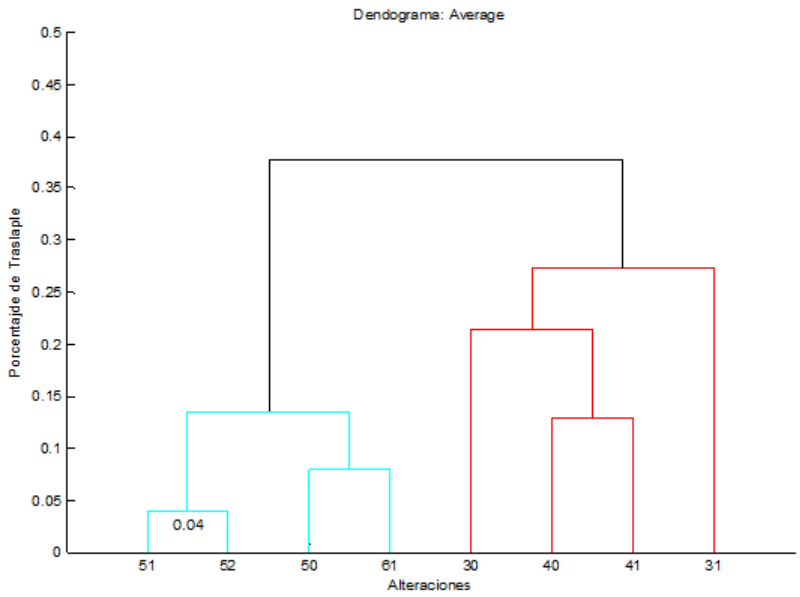


Figura 3.15: Dendrograma

Para ver si los datos de validación y entrenamiento son representativos y semejantes entre ellos, se estudió su traslape de poblaciones, pudiendo apreciar el valor promedio de traslape para el 75 % en la tabla 3.10 (para más información ver anexos B), siendo estas bastante semejantes.

En la figura 3.14 se aplica el procedimiento a la base de datos sin unir junto con la alteración 50, viendo cómo las alteraciones 51,52 y 40,41 son similares entre ellas.

Para el estudio de la base de datos de entrenamiento, validación y ajuste se decidió hacer un dedrograma de estas con sus variables, y así confirmar que no se cometieron errores en su separación, ya que se espera que sean concordantes entre ellas y representativas de la base de datos original. En la figura 3.16 se puede apreciar que la base de datos es consistente, las categorías se asemejan más con su misma familia que con otra clase.

Tabla 3.9: Traslape 40-41 y 51,52

Alteraciones	Traslape todos los elementos	Traslape 4 elementos más diferenciados
40-41	87 %	48 %
51-52	96 %	72 %

Tabla 3.10: Traslape promedio entrenamiento - validación

Alteración	30	31	40	50	61
Traslape Promedio	94 %	85 %	92 %	94 %	85 %

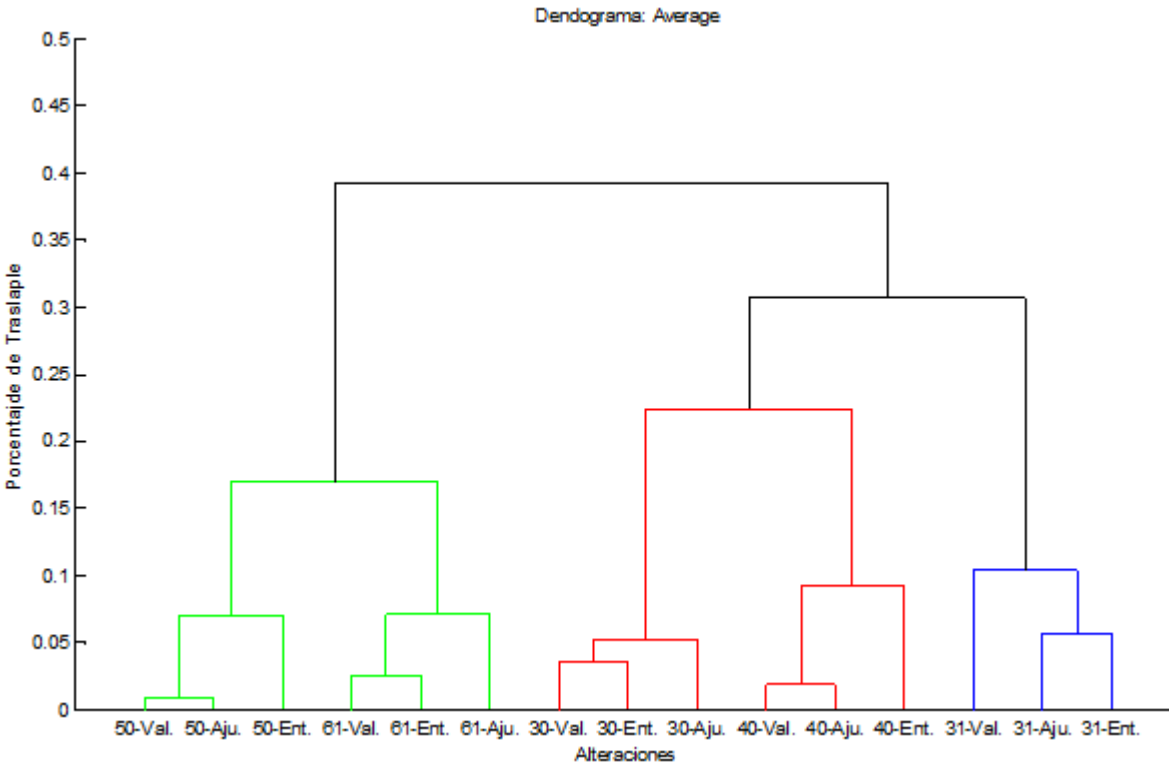


Figura 3.16: comparación entrenamiento, validación y afinamiento

En la figura 3.17 se ve el dendrograma aplicado a la información unida que se tiene. Es importante ver como el dendrograma es una herramienta que permite separar por grupos de categorías según sus semejanzas, siendo esta, una manera válida para tomar decisiones de clasificación, asemejándose a la de un árbol de decisión, dejando juntas las alteraciones más parecidas entre sí.

La clasificación en base a la figura 3.17 será el método que se tomará para lograr la clasificación de las alteraciones.

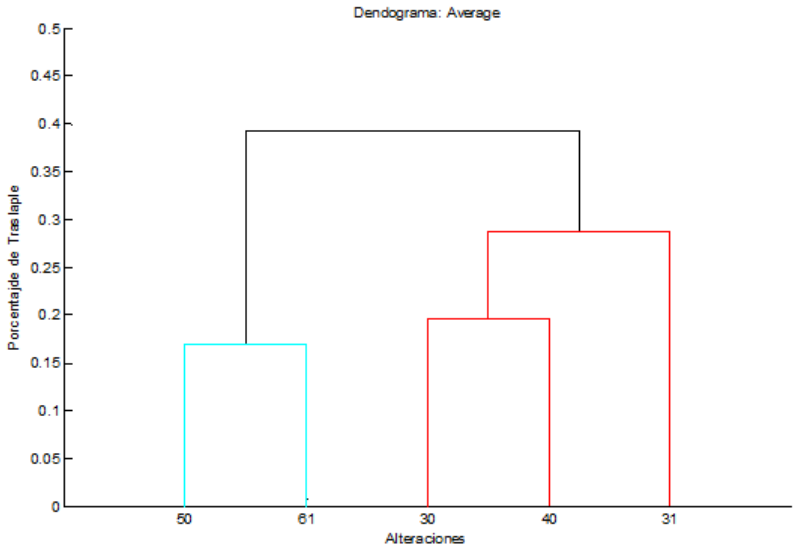


Figura 3.17: Dendrograma: Criterio de Clasificación

3.9. Error de clasificación

Para la realización de una herramienta de clasificación es necesario tener una medición del error, para tener una comparación que permita ir mejorando el modelo. Muchas técnicas estadísticas utilizan el error cuadrático medio como su medidor de desempeño básico, siendo de fácil cómputo, midiendo las diferencias entre lo que una variable predicha debería ser en contra de lo que es, teniendo como principal problema, el que si se desea clasificar un patrón dentro de varias categorías, este no dice nada acerca de la frecuencia de clasificación errónea que se tenga. Además impide distinguir entre errores menores y errores serios.

El definir una métrica de medición que no sea el error cuadrático general, se basa en que si se tuviese un 99 % de los datos pertenecientes a cierta categoría A y un 1 % de ellos a otra B, si un clasificador clasifica todo como categoría A según el error cuadrático general presentaría solo un 1 % de error y sería casi perfecto, sin tomar en cuenta que se desea tener una clasificación correcta en ambos, por lo cual la métrica de medición que se tomará para realizar el estudio está descrita por los siguientes pasos.

- Se definen las alteraciones que serán codificadas como 0 y 1.
- Se obtiene el error cuadrático por alteración codificada con 0 y 1.

$$Error\ \text{Cuadrático}_{\text{Alteración}} = \frac{|Valor_{real} - Valor_{Pred.}|}{Ndedatos} \tag{3.7}$$

- Obteniendo todos los errores cuadráticos se estudia el error promedio definido como:

$$\text{Error Promedio} = \frac{\sum \overline{\text{error alteración}_{COD:1}} + \sum \overline{\text{error alteración}_{COD:0}}}{2} \tag{3.8}$$

Con esto se mide el error de ambas clases y se genera un nuevo error con el promedio.

3.10. Metodología de selección de variables

La selección de variables, se basará en la metodología forward modificado realizada por Chen [18] (ver Estrategias de Selección de variables en el capítulo anterior) añadiendo un paso más, sumar una fase de validación global del modelo, con el fin de confirmar el desempeño real de este, utilizando una base de datos totalmente diferente de las anteriores. Al utilizar en esta metodología la base de datos de validación para seleccionar las variables se crea un sesgo (el modelo podría tomar patrones específicos que tendría esta base de datos). De esta manera, la metodología está construida para utilizar 3 bases de datos:

- Entrenamiento: Con la que se crea el modelo en cada iteración.
- Ajuste: Con la que se estudia el error que produce el modelo que se utiliza como herramienta para añadir o no otra variable.
- Validación: Con la que se estudia el error final de validación del modelo.

El criterio de detención que se aplicará para la selección de variables será que el cambio en el error producido al añadir un nuevo input sea mayor a 0.5 % para no incluir variables que no ayuden en gran medida a la clasificación, con el fin de lograr generar modelos más simples.

3.11. Modelos de clasificación

En base a la clasificación realizada por el equipo de geologos de la mina, se busca la creación de un modelo clasificador de alteraciones. Estos modelos puede ser realizado con distintos métodos de análisis, en donde a priori no se puede predecir cuál de ellos constituirá un desempeño superior, siendo una etapa crucial en la minería de datos. Los métodos al estar formulados con un trasfondo matemático distinto, generarán diversas salidas produciendo varios modelos. A causa de esto, se estudiaron cuatro técnicas para la realización de un modelo que logre la clasificación de las alteraciones según sus elementos de mayor interés, los cuales son:

- Separación simple.
- K-Mean clustering.
- Regresiones logísticas.
- Redes neuronales.

Estos modelos serán explicados a continuación, siendo elaborados a partir del clúster de categorías.

Separación simple

La separación simple hace referencia al método más sencillo de realizar, para ser contrastado contra otros más complejos, el cual es encontrar la variable (única) que mejor divide los datos en dos categorías, encontrando un umbral de corte óptimo (que minimice el error objetivo), creando un modelo para cada una de las variables y escogiendo el mejor de ellos.

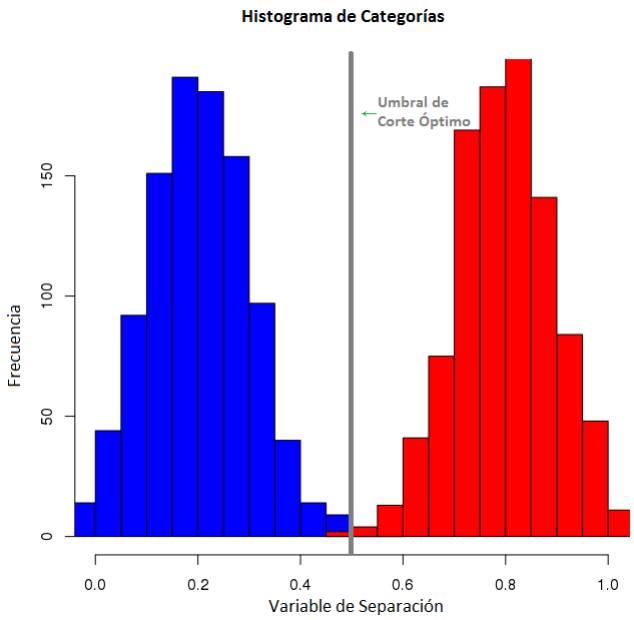


Figura 3.18: Umbral de corte en separación simple

En la figura 3.18 se puede ver la distribución de dos categorías, azul y roja en donde se genera un modelo con un umbral de corte que permite discriminar las dos como «mayor o igual que» y «menor que».

K-Mean clustering

Para generar un modelo que tome en cuenta la distribución de las variables del set de muestras y que permita una fácil visualización en el espacio a través de un gráfico de dispersión como el que se aprecia en la figura 3.19, se utilizara la técnica de clustering.

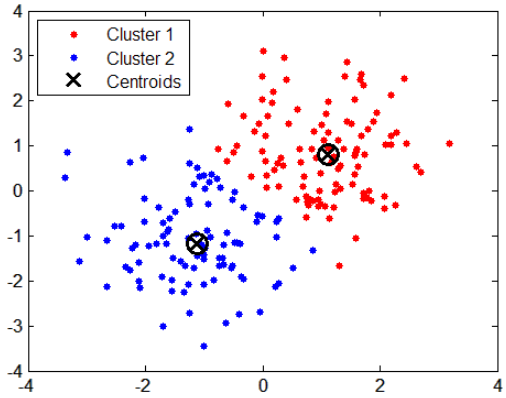


Figura 3.19: ejemplo de K-mean Clustering

De esta manera se busca tener un resultado visual que permite ser graficado en el espacio.

Regresión logística

Con la finalidad de generar un modelo semejante a uno lineal con la información y poder contrastarlo contra los demás modelos, se confeccionará un modelo a través de la regresión logística, al estudiar la probabilidad de pertenecer a una clase de la forma:

(3.9) $$ln[\frac{p(x)}{1-p(x)}] = B_o + B_1 x_1 + \cdots + B_p x_p = B_o + B^{'} x$$

El criterio de discriminación entre positivo y negativo será el valor 0.5.

Redes neuronales artificiales

Con la finalidad de crear un modelo que permita encontrar relaciones no lineales entre los datos, se utilizaron las redes neuronales, donde para definir su arquitectura se utilizaron las siguientes especificaciones:

- Se estableció solo 1 capa oculta en el sistema, ya que en la mayoría de los casos no hay razón para la utilización de capas adicionales [26], debido a que en capas extras los errores al ser “retro-propagados” generan un gradiente más inestable en el algoritmo y además, el número de falsos mínimos usualmente se incrementa drásticamente.
- Para la selección del número de neuronas en la capa oculta que componen la red, Masters [26] formuló una ecuación para determinar la cantidad mínima de estas, la cual es:

$$h = \sqrt{(m * n)} \text{ con } h >= 2 \tag{3.10}$$

En donde n es el número de variables de entrada y m la cantidad de salidas.

Teniendo como base esta configuración, para cada una de las variables se crean cinco modelos con la cantidad mínima recomendada de neuronas h (se crean cinco modelos debido a la aleatoriedad de los pesos iniciales de las neuronas, lo que genera modelos distintos debido al cambio de las valores iniciales), grabando el desempeño de cada uno, evaluados con la base de datos de ajuste, escogiendo el mejor de ellos. Luego 5 nuevas FNN son construidas, con la misma cantidad de neuronas de entrada y salida pero con $h + 1$ neuronas en su capa oculta las cuales serán evaluadas nuevamente con la base de datos de ajuste, almacenando su desempeño y seleccionando el mejor de ellos nuevamente, el que será comparado con el del modelo anterior, si el error disminuye se continuará creando modelos con $h + 2$ neuronas. Este proceso de entrenamiento y validación se continúa realizando hasta que el desempeño de la red no mejore. Métodos similares a este son utilizados en problemas de clasificación de información ergonómica [18].

La red neuronal ha sido creada usando funciones de activación sigmoideas ($tanh$), comúnmente la más utilizada [27]. Además, está confeccionada con un algoritmo de entrenamiento Levenberg-Marquardt, el cual es usualmente el más eficiente utilizando una mayor cantidad de memoria del computador. En el nodo de salida se genera una optimización del umbral de corte buscando el valor que genere el mínimo error.

En la figura 3.20 se tiene una representación de la arquitectura de la FNN con la cual se trabaja.

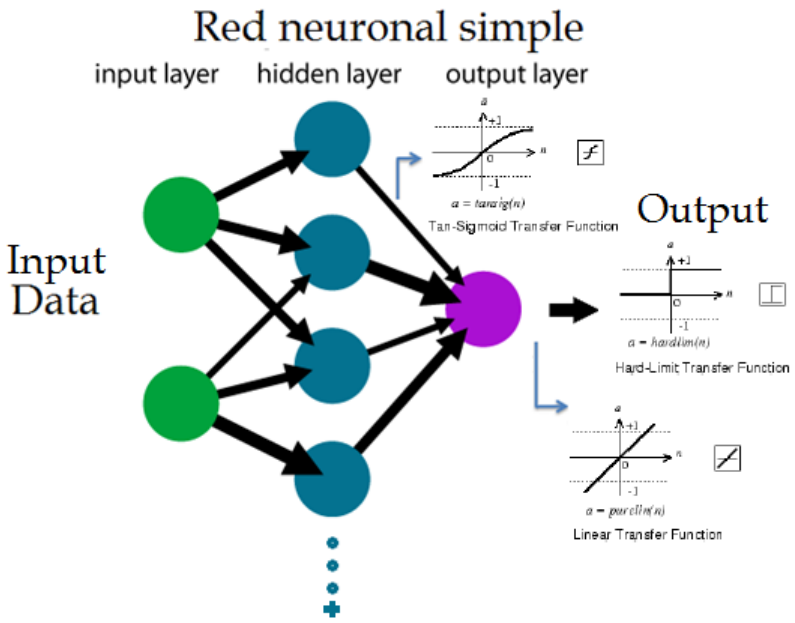


Figura 3.20: Red neuronal

Este modelo fue creado utilizando el toolbox de redes neuronales de Matlab, bajo la metodología de una selección forward modificado.

Capítulo 4

Resultados de los modelos

En el presente capítulo se expondrán los resultados obtenidos, al realizar los distintos modelos matemáticos desarrollados a partir del clustering de categorías, el cual originó un árbol de clasificación siendo ilustrado en la figura 4.1. Este constituye la base para cada uno de los modelos, tratando de generar una división como esta.

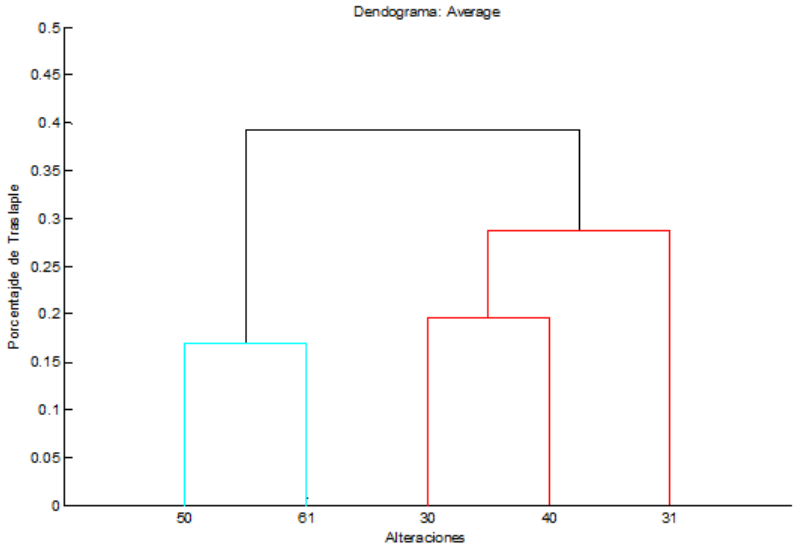


Figura 4.1: Criterio de Clasificación

4.1. Separación Simple

Los modelos generados a partir del método de separación simple entregaron los siguientes resultados:

4.1.1. Alteración 50, 61 vs 30, 31, 40

El elemento de interés es la variable sintética: $\frac{Mg}{Al}$, el cual tiene un umbral de corte menor que: 0.22, para pertenecer a 50 o a 61. Los errores que presentó el modelo fueron para la base de datos de ajuste de: 6.4 % y de entrenamiento: 5.3 %. En la figura 4.2 se puede ver un histograma de los valores utilizados para la generación del modelo y el umbral de corte que se encuentra graficado con una línea punteada.

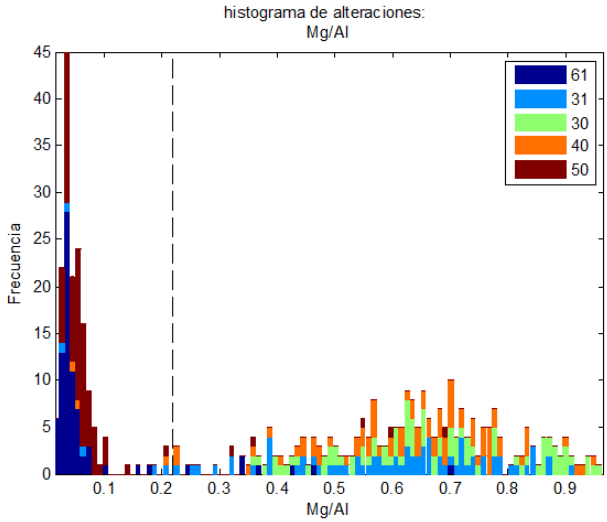


Figura 4.2: Separación Simple 50, 61 vs 30, 31 40 en entrenamiento

En la figura 4.3 se aprecia el modelo desarrollado de separación simple en validación, ingresado información nueva, teniendo un error de validación 7 %.

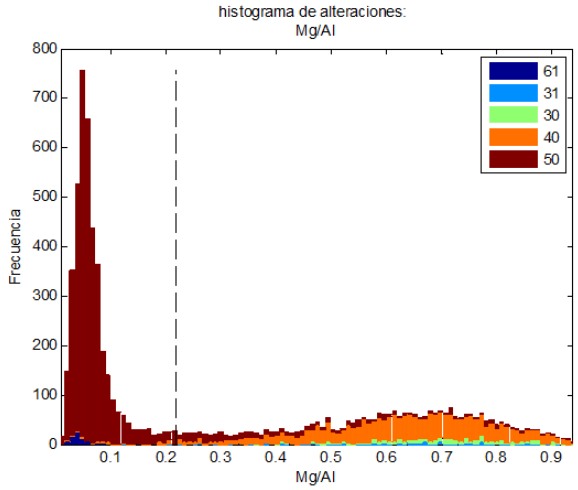


Figura 4.3: Separación Simple 50, 61 vs 30, 31, 40 en validación

4.1.2. Alteración 50 vs 61

El elemento de interés es el aluminio (Al), el cual tiene un umbral de corte menor que 0.6, para pertenecer a las alteraciones tipo 50. Los errores que presentó el modelo fueron para la base de datos de ajuste de: 27.4 % y de entrenamiento: 26.2 %. En la figura 4.4 se puede ver un histograma de los valores utilizados para la generación del modelo y el umbral de corte que se encuentra graficado con una línea punteada.

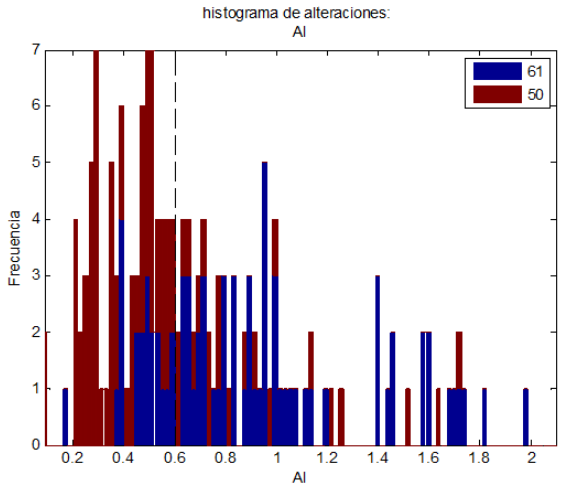


Figura 4.4: Separación Simple 50 vs 61 en entrenamiento

En la figura 4.5 se aprecia el modelo desarrollado de separación simple en validación, ingresado información nueva teniendo un error de validación 25.2 %.

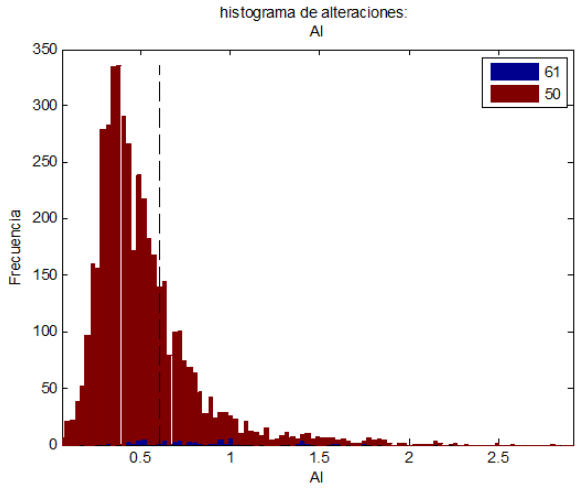


Figura 4.5: Separación simple 50 vs 61 en validación

4.1.3. Alteración 31 vs 30, 40

La variable de interés es: Aluminio (Al), el cual tiene un umbral de corte menor que: 1.3, para pertenecer a la alteración 31 (feldespato K > biotita). Los errores que presentó el modelo fueron para la base de datos de ajuste de: 11.9 % y de entrenamiento: 14 %. En la figura 4.6 se puede ver un histograma de los valores utilizados para la generación del modelo y el umbral de corte, el que se encuentra graficado con una línea punteada.

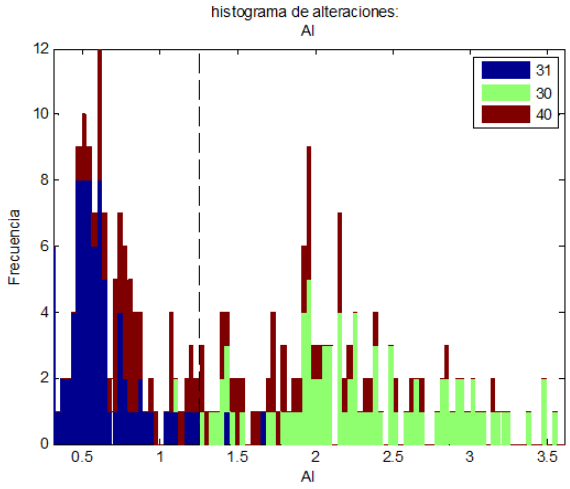


Figura 4.6: Separación simple 31 vs 30, 40 en entrenamiento

En la figura 4.7 se aprecia el modelo desarrollado de separación simple en validación, ingresando información nueva, teniendo un error de validación 11.5 %.

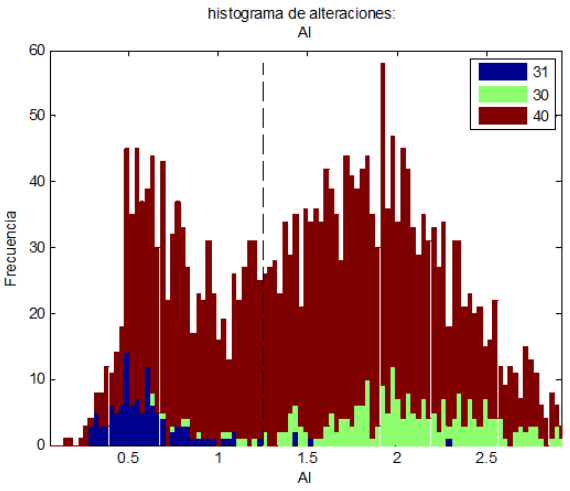


Figura 4.7: Separación simple 31 vs 30, 40 en validación

4.1.4. Alteración 30 vs 40

La variable de interés es el: Rubidio (Rb), el cual tiene un umbral de corte mayor que: 20 para pertenecer a la alteración 30 (biotita secundaria). Los errores que presentó el modelo fueron para la base de datos de Ajuste de: 15.6 % y de entrenamiento: 14 %. En la figura 4.8 se puede ver un histograma de los valores utilizados para la generación del modelo y el umbral de corte que se encuentra graficado con una línea punteada.

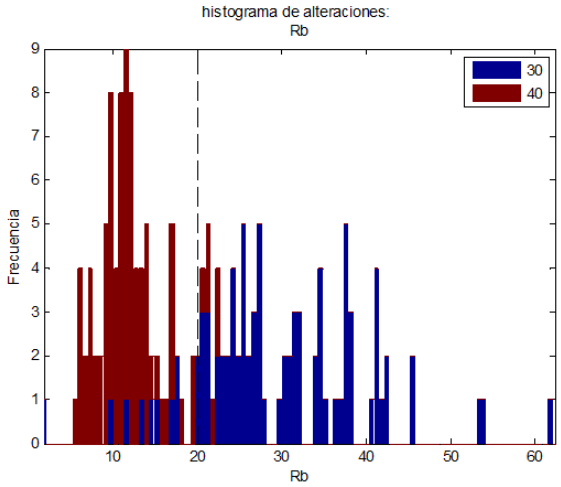


Figura 4.8: Separación simple 30 vs 40 en entrenamiento

En la figura 4.9 se aprecia el modelo desarrollado de separación simple en validación, ingresando información nueva teniendo un error de validación 16.8 %.

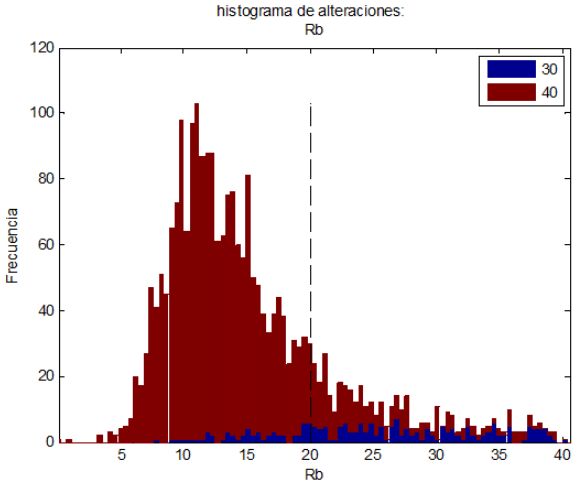


Figura 4.9: Separación simple 30 vs 40 en validación

4.1.5. Modelo Separación Simple

Bajo estos parámetros se obtuvo la tabla 4.1, en la que se puede observar cómo fueron clasificados las mediciones de la base de datos de validación, teniendo un error promedio de 28 %.

Tabla 4.1: Validación separación simple

	Alt. Predicha	30	31	40	50	61
Alt. Mapeada	Cantidad de Datos	508	1263	1277	3160	902
30	261	78%	5%	17%	0%	0%
31	115	4%	95%	0%	2%	0%
40	2169	13%	32%	52%	1%	1%
50	4474	0%	10%	2%	70%	18%
61	91	1%	7%	6%	24%	63%
Acierto Promedio:				72%		

En la figura 4.10 se puede apreciar el error que se generó en cada uno de los nodos del árbol de separación.

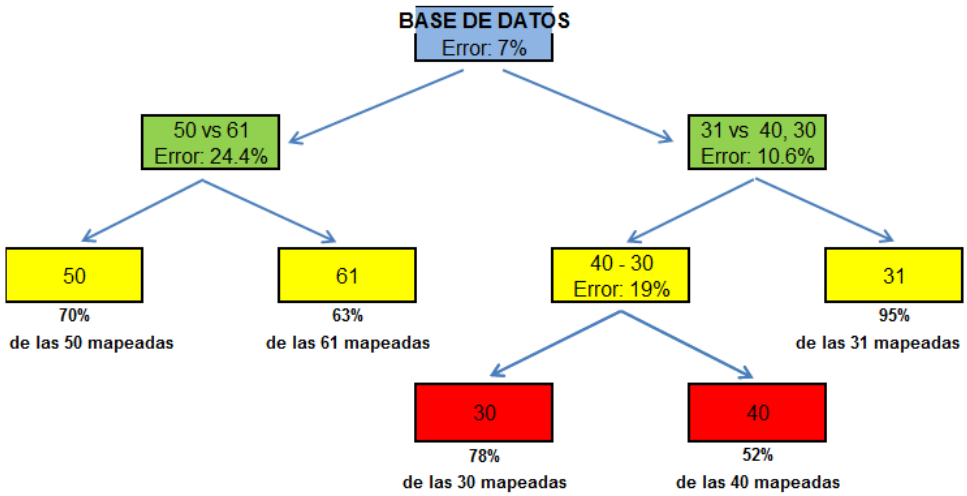


Figura 4.10: Árbol de clasificación separación simple

4.2. K-mean Clustering

Utilizando la selección de variables forward modificado, se realizó la metodología expuesta utilizando el K-mean clustering, con el objetivo de tener una representación visual de estos y ver cómo se encuentran distribuidos.

4.2.1. Alteración 50, 61 vs 30, 31, 40

Para crear la división de las alteraciones 50 y 61 del resto de las alteraciones, a través de la metodología forward, el algoritmo determinó que la mejor variable es: $\frac{Mg}{Al}$ (por sí sola el modelo no mejora en gran medida con más variables), presentando un 6.2 % y 6.7 % de errores en entrenamiento y ajuste respectivamente.

El algoritmo se detuvo con la variable: CuS que presentó errores en entrenamiento de 6.1 % y en ajuste de 6.3 %.

Los centroides fueron situados en la siguiente posición $\frac{Mg}{Al}$:

- Centroide 1: (0.08).
- Centroide 2: (0.67).

La figura 4.11 muestra un modelo creado con la base de datos de entrenamiento, ubicando en ella los centroides.

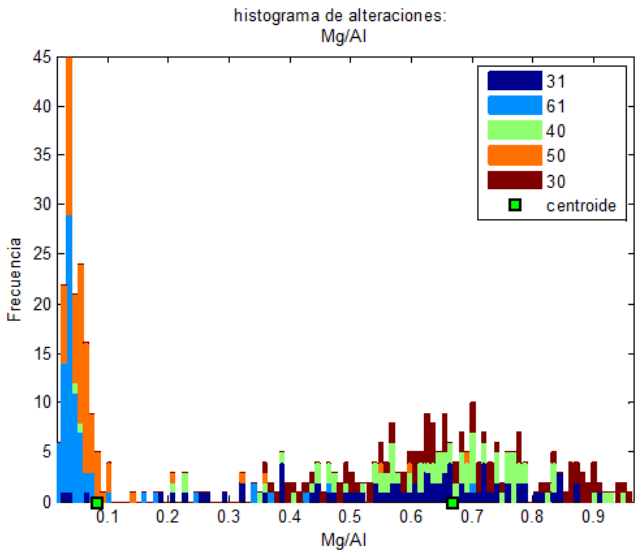


Figura 4.11: k-mean 50, 61 vs 30, 31, 40 en entrenamiento

La figura 4.12 muestra los datos de validación en el modelo con un error de 6.4 % en esta etapa.

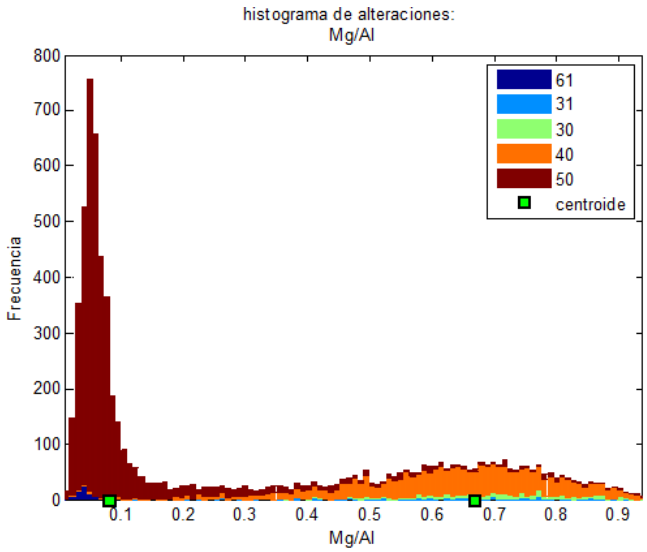


Figura 4.12: k-mean 50, 61 vs 30, 31, 40 en validación

4.2.2. Alteración 50 vs 61

Para crear la división de las alteraciones 50 y 61, a través de la metodología forward, se utilizaron las siguientes variables con sus respectivos errores asociados al momento de su selección:

1. $\frac{(Al+K)}{Na+Ca+Mg}$: Entrenamiento: 35.3 %, Ajuste: 29.7 %.
2. Rb: Entrenamiento: 37.1 %, Ajuste: 28.6 %.
3. $\frac{3*Al}{K+Na}$: Entrenamiento: 34.4 %, Ajuste: 24.8 %.
4. S: Entrenamiento: 33.1 %, Ajuste: 24 %.

El algoritmo se detuvo con la variable Be, la cual presentó los siguientes errores: Entrenamiento: 33.1 %, Ajuste: 24. %.

Los centroides fueron situados en la siguiente posición ($\frac{(Al+K)}{(Na+Ca+Mg)}$, Rb, $\frac{3Al}{(K+Na)}$, S):

- Centroide 1: (8.33, 13.57, 11.01, 2.21).
- Centroide 2: (4.92, 9.55, 6.0, 2.55).

La figura 4.13 muestra un modelo creado con la base de datos de entrenamiento, ubicando en ella los centroides.

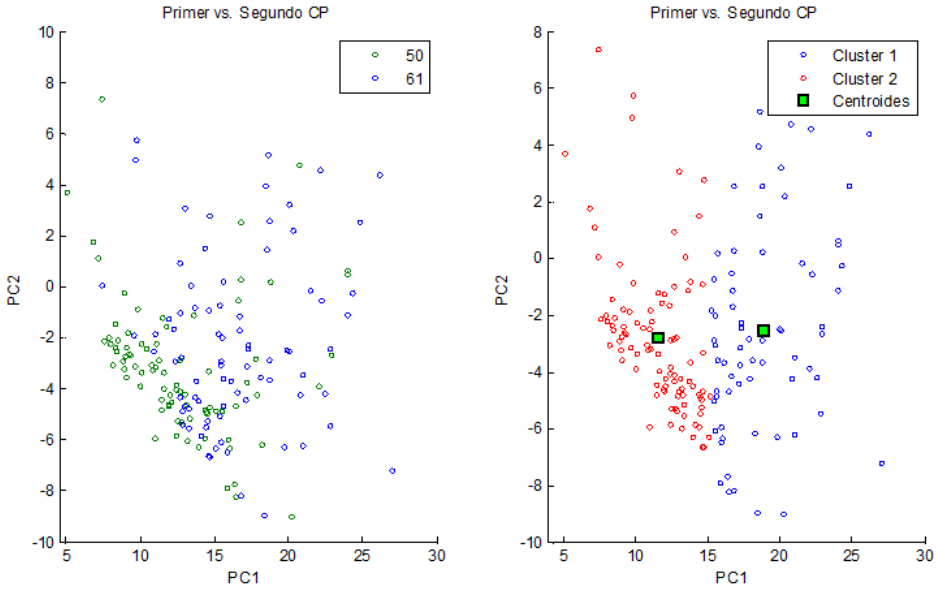


Figura 4.13: k-mean 50 vs 61 en entrenamiento

La figura 4.14 muestra los datos de validación en el modelo con un error de 26.5 % en esta etapa.

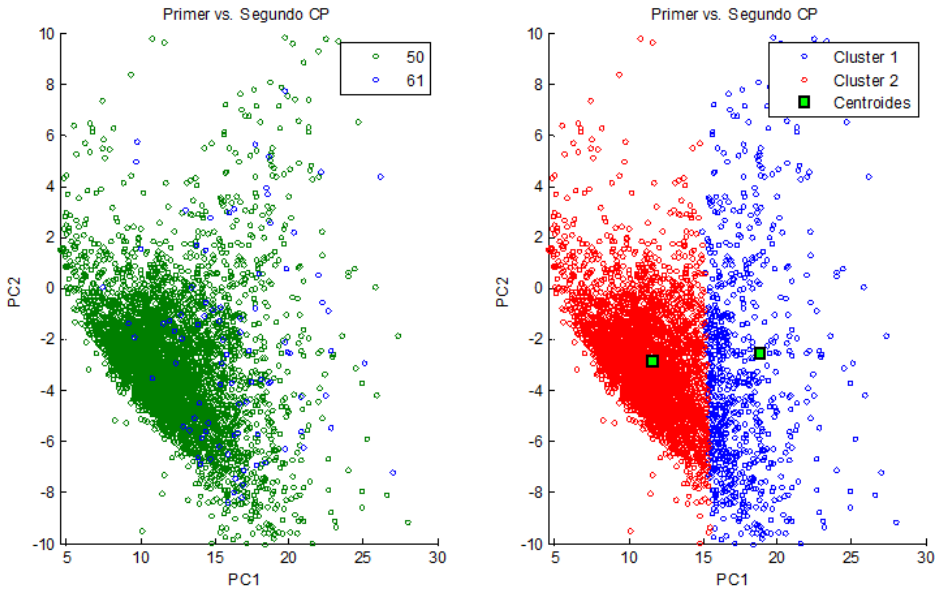


Figura 4.14: k-mean 50 vs 61 en validación

4.2.3. Alteración 31 vs 30, 40

Para crear la división de la alteración 31 de las 30 y 40, a través de la metodología forward, se utilizó las siguientes variables con sus respectivos errores asociados al momento de su selección:

1. Mg: Entrenamiento: 15.8 %, Ajuste: 14.7 %
2. $\frac{K}{Mg}$: Entrenamiento: 15.9 %, Ajuste: 13.1 %
3. $\frac{Ca+Na}{K+Al}$: Entrenamiento: 13.3 %, Ajuste: 12.3 %

El algoritmo se detuvo con la variable $KxAl$ que presentó un 13 % y un 11.9 % de error en entrenamiento y ajuste respectivamente.

Los centroides fueron situados en la siguiente posición (Mg, $\frac{K}{Mg}$, $\frac{Ca+Na}{K+Al}$):

- Centroide 1: (1.39, 0.28, 0.29).
- Centroide 2: (0.34 1.07, 0.62).

La figura 4.15 muestra un modelo creado con la base de datos de entrenamiento, ubicando en ella los centroides.

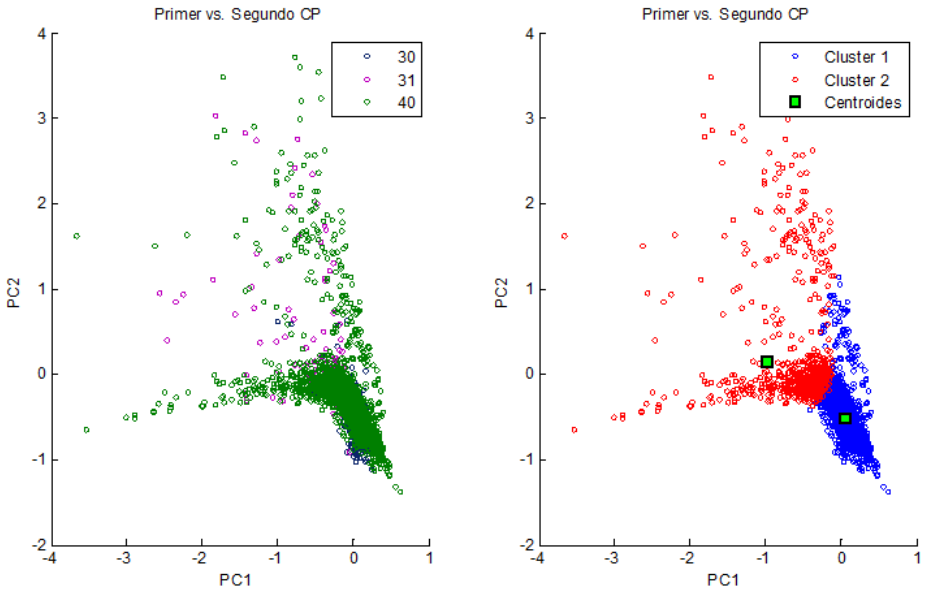


Figura 4.15: k-mean 31 vs 30, 40 en entrenamiento

La figura 4.16 muestra los datos de validación en el modelo con un error de 11.7 % en esta etapa.

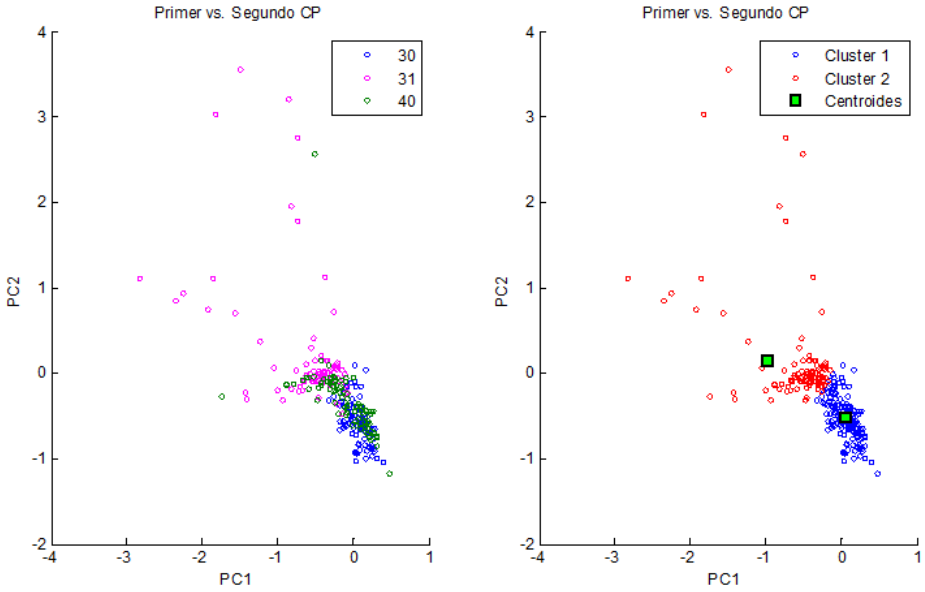


Figura 4.16: k-mean 31 vs 30, 40 en validación

4.2.4. Alteración 30 vs 40

Para crear la división entre las alteraciones 30 y 40, a través de la metodología forward, se utilizaron las siguientes variables, con sus respectivos errores asociados al momento de su selección:

1. Sc: Entrenamiento: 14.2 %, Ajuste: 19.1 %.
2. $\frac{3Al}{K+Na}$: Entrenamiento: 8 %, Ajuste: 16.1 %.
3. Se: Entrenamiento: 8 %, Ajuste: 15.1 %.

El algoritmo se detuvo con la variable: Mg la cual presentó un 8 % y un 14.8 % de error en entrenamiento y ajuste respectivamente.

Los centroides fueron situados en la siguiente posición (Sc, $\frac{3Al}{K+Na}$, Se):

- Centroide 1: (2.3, 13.47, 3.16).
- Centroide 2: (8.84, 11.52, 2.29).

La figura 4.17 muestra el modelo creado con la base de datos de entrenamiento, ubicando en ella los centroides.

La figura 4.18 muestra los datos de validación en el modelo, con un error de 16.9 % en esta etapa.

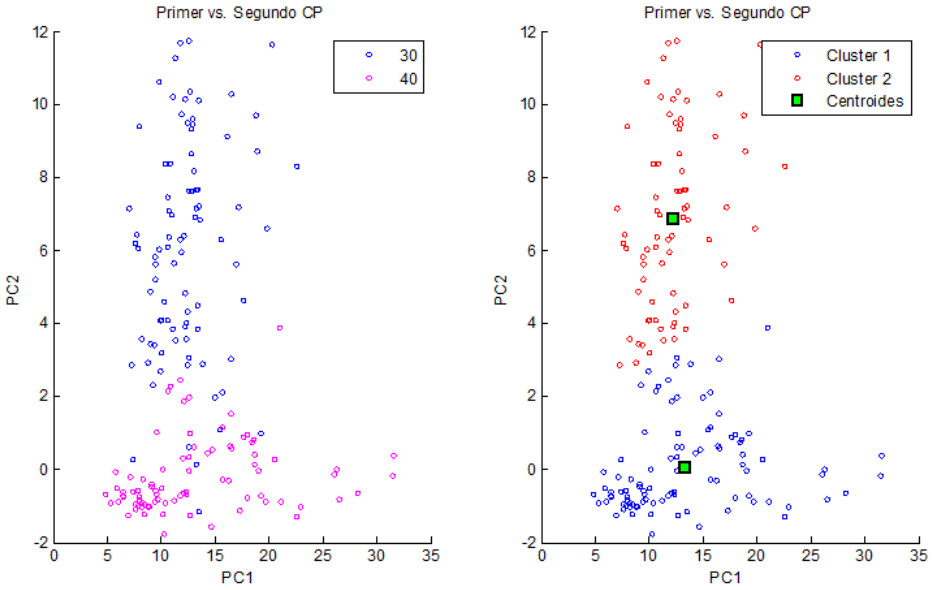


Figura 4.17: k-mean 30 vs 40 en entrenamiento

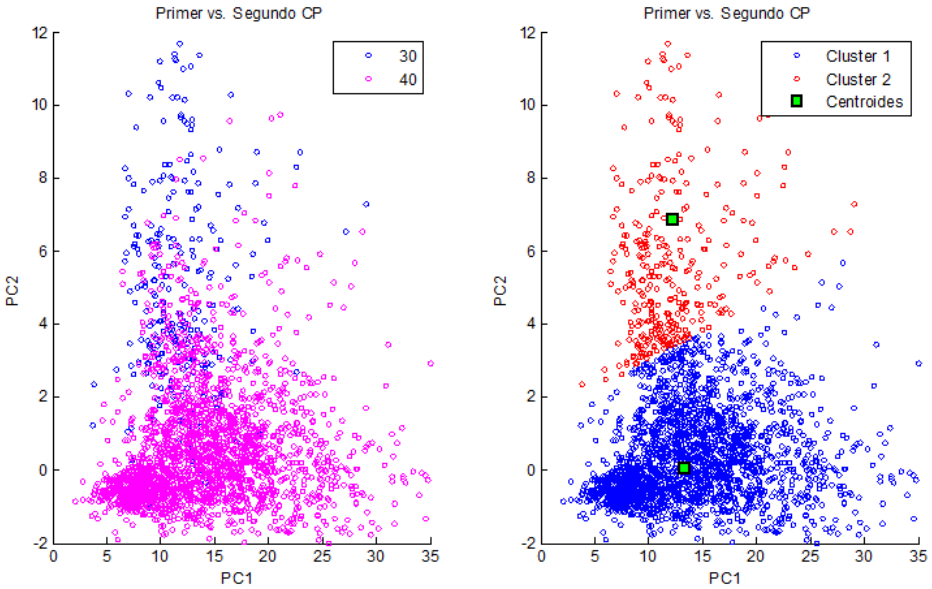


Figura 4.18: k-mean 30 vs 40 en validación

4.2.5. Modelo K-mean clúster

Bajo los modelos anteriormente descritos se clasificó la información de validación siguiendo el árbol de decisión, generándose un error promedio general de 28 % según se aprecia en la tabla 4.2.

En la figura 4.19 se puede apreciar el error que se generó en cada uno de los nodos del árbol de separación.

Tabla 4.2: Validación k-mean

	Alt. Predicha	30	31	40	50	61
Alt. Mapeada	Cantidad de Datos	329	709	1644	3711	717
30	261	72%	2%	25%	0%	0%
31	115	1%	82%	8%	9%	1%
40	2169	6%	19%	67%	6%	2%
50	4474	0%	4%	3%	79%	14%
61	91	0%	2%	4%	34%	59%
Acierto Promedio:				72%		

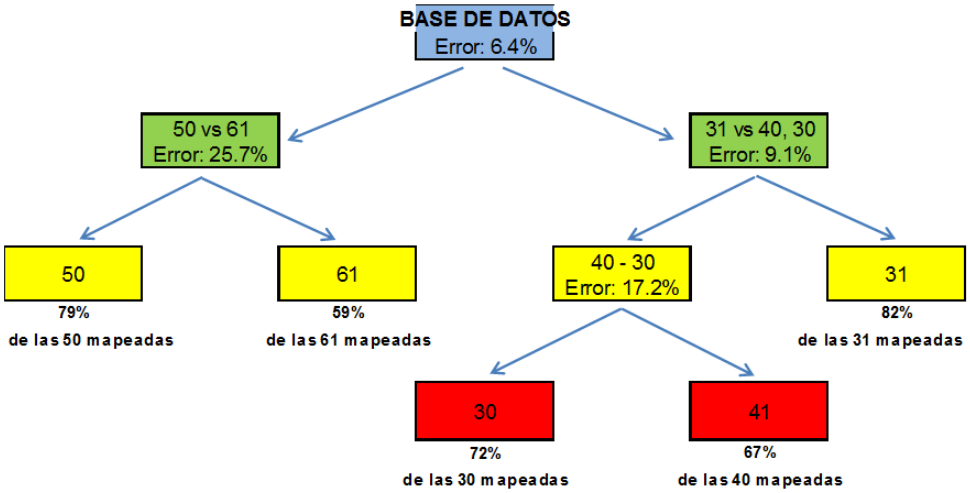


Figura 4.19: Árbol de validación K-mean Clustering

4.3. Regresión Logística

4.3.1. Alteración 50, 61 vs 30, 31, 40

Para crear la división de las alteraciones 50 y 61 del resto de las alteraciones, a través de la metodología forward, se utilizaron las siguientes variables, junto con sus errores asociados al momento de su selección:

1. $\frac{Mg}{Al}$: Entrenamiento: 6 %, Ajuste: 6.6 %
2. K: Entrenamiento: 5.3 %, Ajuste: 5.7 %

El algoritmo se detuvo con la variable: Te, la cual presentó errores de un 5.3 % en entrenamiento y un 5.6 % en ajuste.

El modelo quedó descrito por la siguiente ecuación aproximadamente:

$$ln[\frac{p(x)}{1-p(x)}] = 7.88 - 13.17 * Mg - 17.19 * K \tag{4.1}$$

Quedando la probabilidad de pertenecer a la alteración 50 o 61 ($p(x) \geq 0.5$) definida como:

$$p(x) = \frac{e^{7.88-13.17*Mg-17.19*K}}{1+e^{7.88-13.17*Mg-17.19*K}} \tag{4.2}$$

Finalmente el error de validación es: 6.2 %

4.3.2. Alteración 50 vs 61

Para crear la división entre las alteraciones 50 y 61, a través de la metodología forward, se utilizaron las siguientes variables, junto con sus errores asociados al momento de su selección:

1. $\frac{Na}{Al}$: Entrenamiento: 24.9 %, Ajuste: 27.3 %.
2. CuT: Entrenamiento: 27.3 %, Ajuste: 23.7 %.
3. $\frac{KxNA}{Al}$: Entrenamiento: 24.8 %, Ajuste: 22.7 %.

El algoritmo se detuvo con la variable: Te, la cual presentó los siguientes errores: entrenamiento: 24.8 %, ajuste: 22.2 %.

El modelo quedó descrito por la siguiente ecuación:

$$ln[\frac{p(x)}{1-p(x)}] = -1.93 + 40.6 * Na/Al - 0.4 * CuT - 51.65 * K * Na/Al \tag{4.3}$$

Quedando la probabilidad de pertenecer a la alteración 50 ($p(x) \geq 0.5$) definida como:

$$p(x) = \frac{e^{-1.93+40.6*Na/Al-0.4*CuT-51.65*(K*Na)/Al}}{1+e^{-1.93+40.6*Na/Al-0.4*CuT-51.65*(K*Na)/Al}} \tag{4.4}$$

El error de validación asociado al modelo en validación es: 26.5 %.

4.3.3. Alteración 31 vs 30, 40

Para crear la división de las alteraciones 31 de la 30 y la 40, a través de la metodología forward, se utilizaron las siguientes variables, junto con sus errores asociados al momento de su selección:

1. Mg: Entrenamiento: 13.4 %, Ajuste: 11.8 %
2. Na: Entrenamiento: 10.3 %, Ajuste: 10.9 %
3. Co: Entrenamiento: 11.2 %, Ajuste: 9.4 %

El algoritmo se detuvo con la variable: P, la cual presentó errores de un 10.6 % en entrenamiento y un 9 % en ajuste.

El modelo quedó descrito por la siguiente ecuación:

$$ln[\frac{p(x)}{1-p(x)}] = 2.75 - 4.83 * Mg + 18.81 * Na - 0.26 * Co \tag{4.5}$$

Quedando la probabilidad de pertenecer a la alteración 31 ($p(x) \geq 0.5$) definida como:

$$p(x) = \frac{e^{2.75-4.83*Mg+18.81*Na-0.26*Co}}{1+e^{2.75-4.83*Mg+18.81*Na-0.26*Co}} \tag{4.6}$$

Error que presentó el modelo en validación es de 12.5 %

4.3.4. Alteración 30 vs 40

Para crear la división entre las alteraciones 30 y 40, a través de la metodología forward se utilizaron las siguientes variables, junto con sus errores asociados al momento de su selección:

1. Rb: Entrenamiento: 9.88 %, Ajuste: 17.25 %
2. Ni: Entrenamiento: 7.01 %, Ajuste: 13.72 %
3. Sc: Entrenamiento: 5.82 %, Ajuste: 12.74 %

4. $\frac{K}{Ca+Na}$: Entrenamiento: 4.65 %, Ajuste: 11.59 %
5. Sr: Entrenamiento: 4.65 %, Ajuste: 10.66 %

El algoritmo se detuvo con la variable: W, la cual presentó errores de un 4.7 % en entrenamiento y un 10.6 % en ajuste.

El modelo es descrito por la siguiente ecuación:

$$ln[\frac{p(x)}{1-p(x)}] = -6.8 + 0.34 * Rb + 0.03 * Ni + 0.37 * Sc - 2.14 * K/(Ca+Na) - 0.01 * Sr \tag{4.7}$$

Quedando la probabilidad de pertenecer a la alteración 30 ($p(x) \geq 0.5$) definida como:

$$p(x) = \frac{e^{6.8+0.34*Rb+0.03*Ni+0.37*Sc-2.14*K/(Ca+Na)-.01*Sr}}{1+e^{6.8+0.34*Rb+0.03*Ni+0.37*Sc-2.14*K/Ca+Na-.01*Sr}} \tag{4.8}$$

4.3.5. Modelo regresión logística

En la tabla 4.3 se puede apreciar los resultados obtenidos al realizar el árbol de clasificación con los distintos modelos, anteriormente expuestos, presentando un error promedio general de 26 %.

Tabla 4.3: Validación regresión logística

	Alt. Predicha	**30**	**31**	**40**	**50**	**61**
Alt. Mapeada	Cantidad de Datos	655	573	1611	3000	1271
30	261	92%	1%	7%	0%	0%
31	115	9%	77%	9%	4%	2%
40	2169	18%	14%	63%	3%	2%
50	4474	0%	4%	5%	65%	26%
61	91	0%	3%	6%	20%	71%
Acierto Promedio:				74%		

En la figura 4.20 se puede apreciar el error que se generó en cada uno de los nodos del árbol de separación.

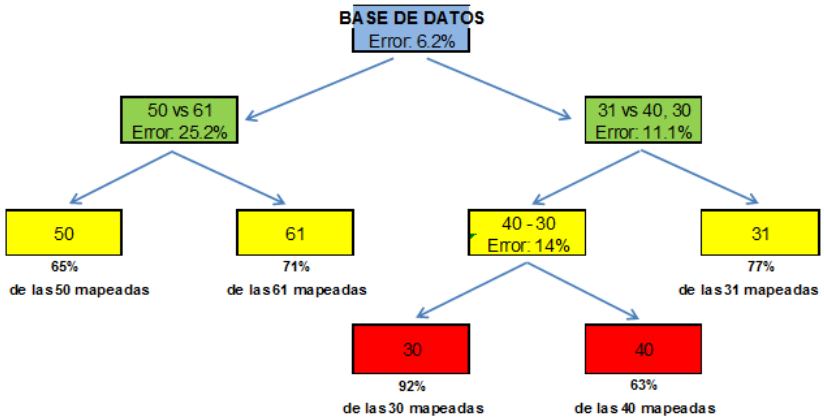


Figura 4.20: Validación árbol regresión logística

4.4. Red neuronal artificial

Bajo el agrupamiento de alteraciones propuesto, se realiza la metodología de redes neuronales con el método forward, explicado en el capítulo anterior.

4.4.1. Alteración 50, 61 vs 30, 31, 40

El modelo de RNA utilizado para la separación de las alteraciones 50 y 61 del resto de las alteraciones, consta de 5 neuronas, con una función de activación con umbral de 0.18 y utilizando las siguientes variables en orden descendente de importancia. Junto a ellas, se encuentran los errores asociados al momento de su selección:

1. $\frac{Mg}{Al}$: Entrenamiento: 5.2 %, Ajuste: 6.6 %.
2. Ca: Entrenamiento: 4.7 %, Ajuste: 5.8 %.

El algoritmo se detuvo con la variable: Na, la cual presentó los siguientes errores: Entrenamiento: 4.8 %, Ajuste: 6.1 %.

La salida de la red neuronal en entrenamiento se ve representada en la figura 4.21.

Para la validación del modelo, este entregó los resultados expuestos en la figura 4.22, en donde el error de Validación es de 5.5 %.

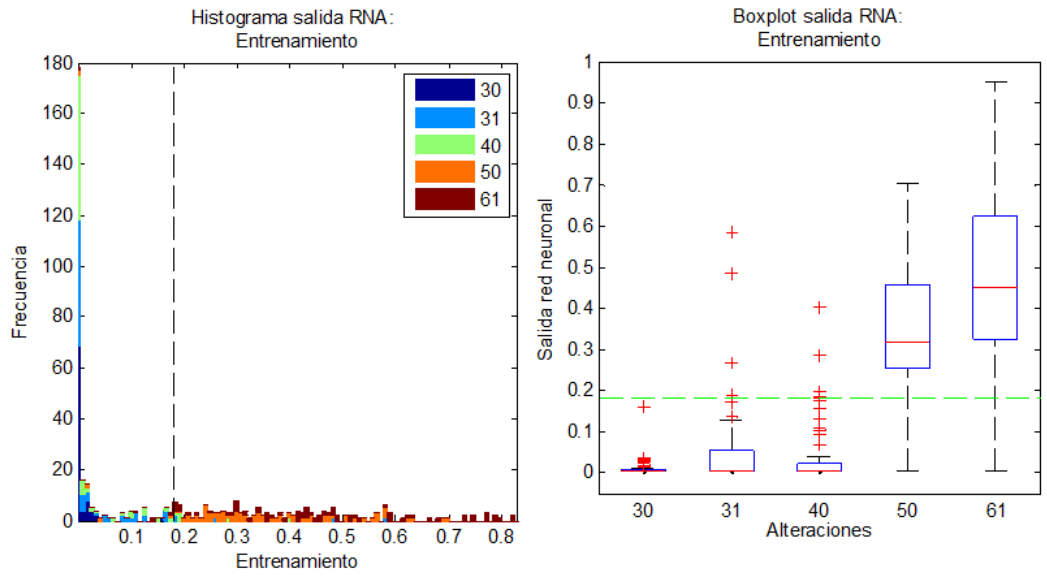


Figura 4.21: Salida RNA entrenamiento 50, 61 vs 30, 31, 40

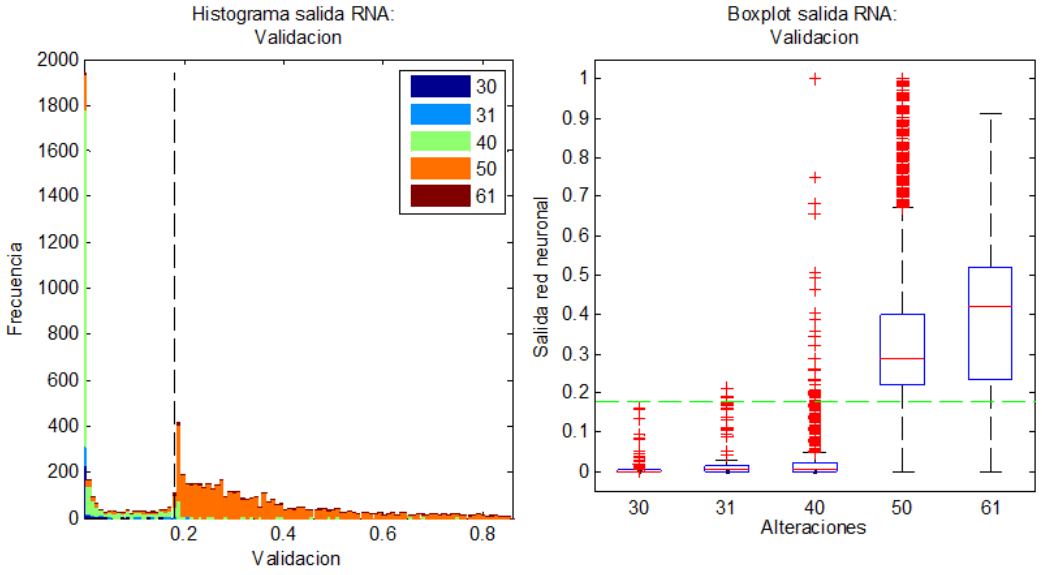


Figura 4.22: Salida RNA validación 50, 61 vs 30, 31, 40

4.4.2. Alteración 50 vs 61

El modelo de RNA utilizado para la separación de las alteraciones tipo 50 de las 60, consta de 4 neuronas, con una función de activación con umbral localizado en 0.51, y utilizando las siguientes variables en orden descendente de importancia. Se encuentran junto a ellas los errores asociados al momento de su selección:

1. Al: Entrenamiento: 25 %, Ajuste: 26.7 %.
2. Be: Entrenamiento: 23 %, Ajuste: 18.5 %.

3. Cu: Entrenamiento: 20.5 %, Ajuste: 16.4 %.
4. Cs: Entrenamiento: 23 %, Ajuste: 14.9 %.
5. S: Entrenamiento: 15.2 %, Ajuste: 14.3 %.

El algoritmo se detuvo con la variable: CuT, la cual presentó en entrenamiento un error de 18.2 % y en ajuste 15.5 %.

La salida de la red neuronal en entrenamiento se ve representada en la figura 4.23.

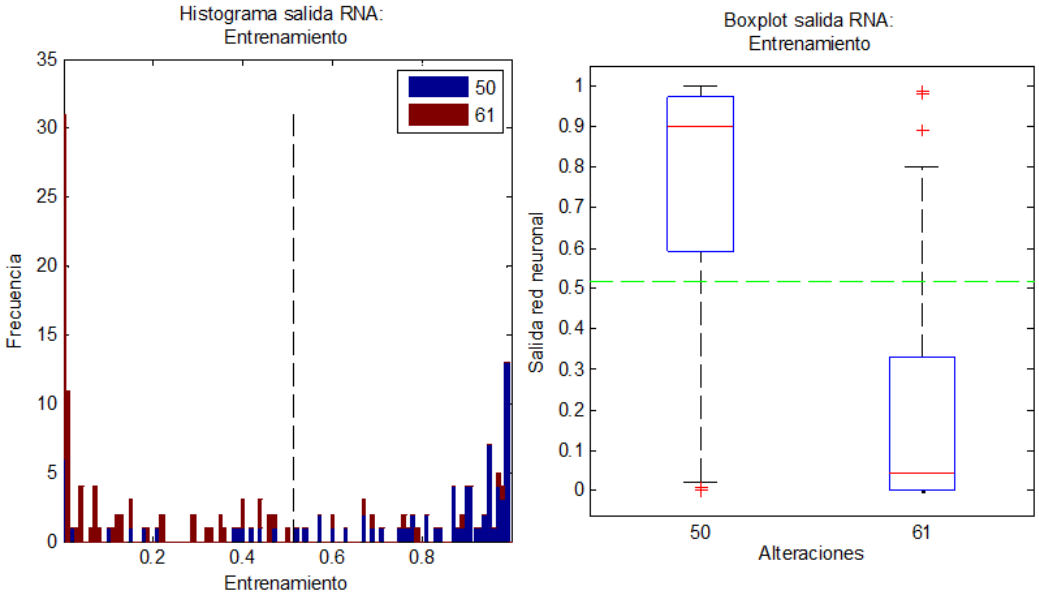


Figura 4.23: Salida RNA entrenamiento, 50 vs 61

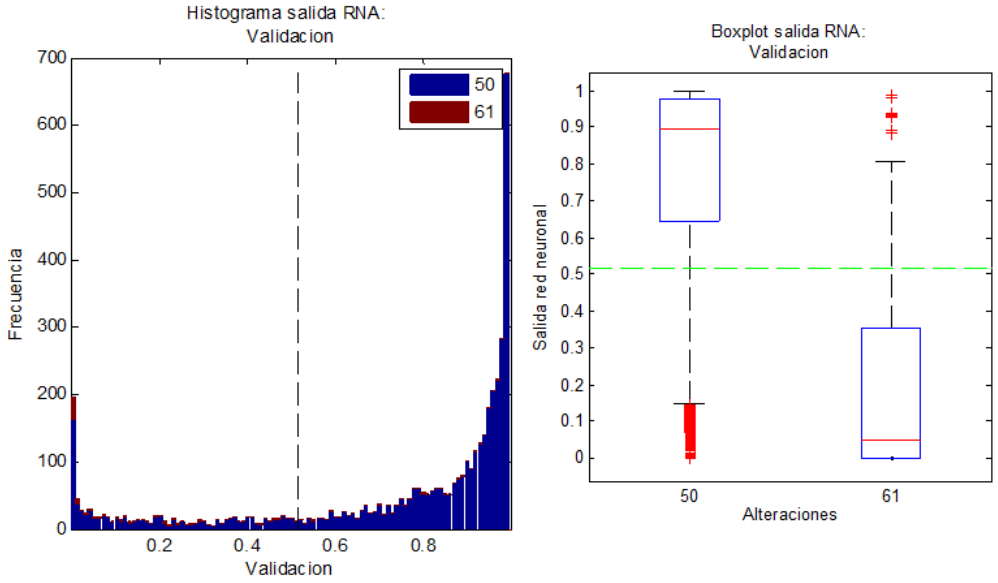


Figura 4.24: Salida RNA validación, 50 vs 61

Para la validación del modelo, este entregó los resultados expuestos en la figura 4.24, en donde el error de Validación es de 17.7 %.

4.4.3. Alteración 31 vs 30, 40

El modelo de RNA utilizado para la separación de las alteraciones tipo 31 de la 30 y la 40, consta de 4 neuronas, con una función de activación con umbral de 0.24, y utilizando las siguientes variables en orden descendente de importancia. Se encuentran junto a ellas los errores asociados al momento de su selección:

1. Al: Entrenamiento: 14.5 %, Ajuste: 11.2 %.
2. Co: Entrenamiento: 12 %, Ajuste: 9.1 %.

El algoritmo se detuvo con la variable: $\frac{3Al}{K+Na}$, la cual presentó en entrenamiento un error de 9 % y en ajuste 8.8 %.

La salida de la red neuronal en entrenamiento se ve representada en la figura 4.25.

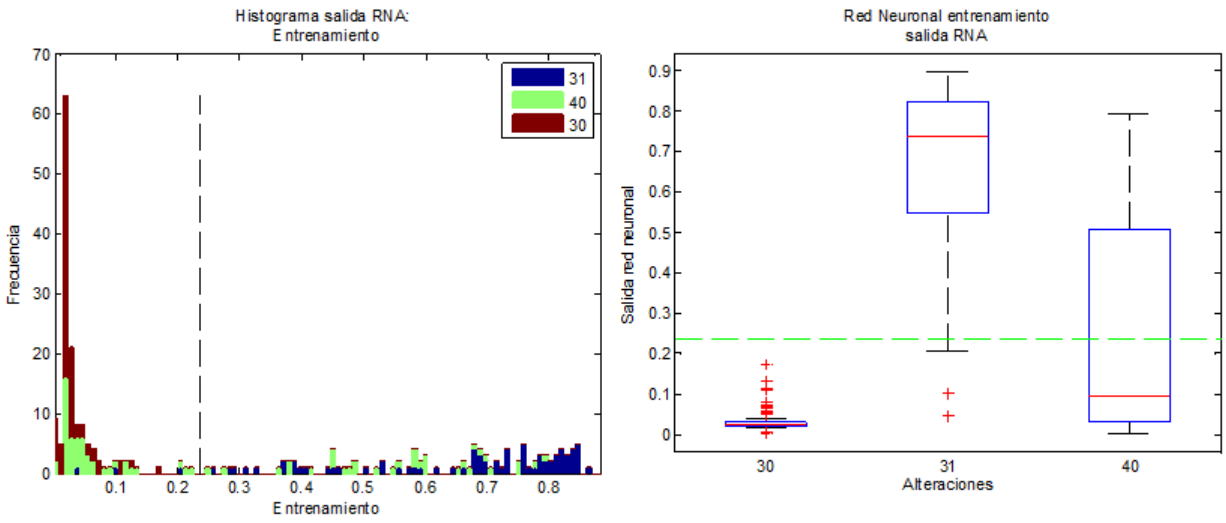


Figura 4.25: Salida RNA entrenamiento

Para la validación del modelo, este entregó los resultados expuestos en la figura 4.26 en donde el error de Validación es de 11 %

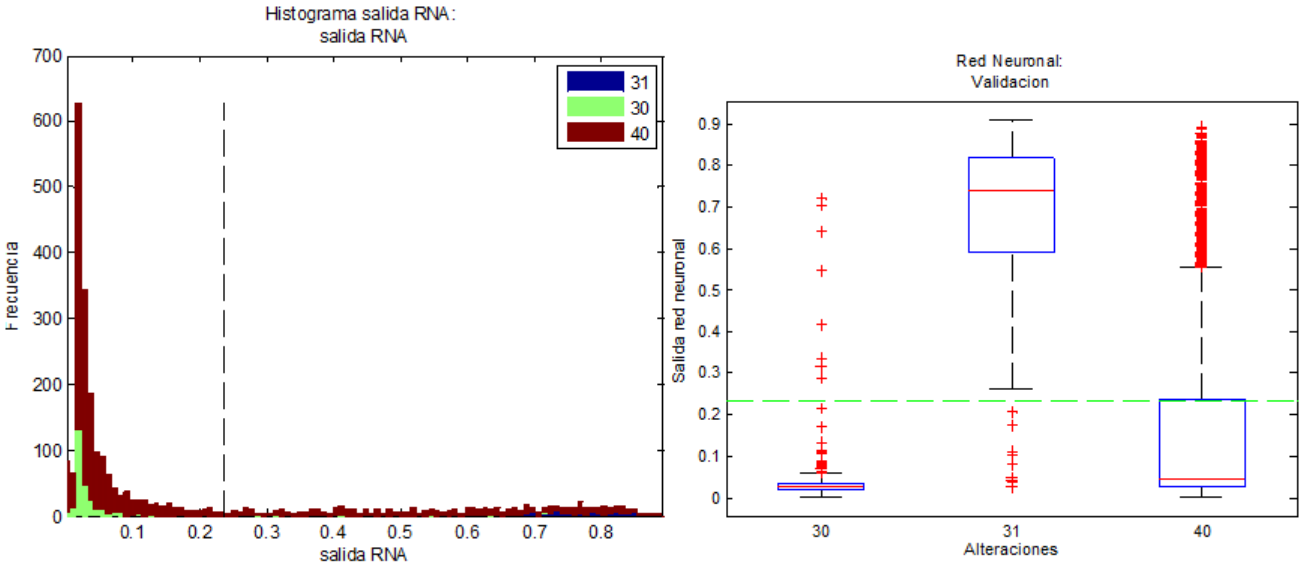


Figura 4.26: Salida RNA validación

4.4.4. Alteración 30 vs 40

El modelo de RNA utilizado para la separación de las alteraciones tipo 30 del resto, consta de 3 neuronas, con una función de activación con umbral de 0.38 y utilizando las siguientes variables en orden descendente de importancia, junto a ellas se encuentran los errores asociados al momento de su selección:

1. Rb: Entrenamiento: 8.7 %, Ajuste: 15.8 %.
2. Na: Entrenamiento: 6.9 %, Ajuste: 14 %.
3. Ni: Entrenamiento: 5.2 %, Ajuste: 12.8 %.
4. Sc: Entrenamiento: 5.2 %, Ajuste: 11.9 %.
5. Sr: Entrenamiento: 3.5 %, Ajuste: 11 %.

El algoritmo se detuvo con la variable: S, la cual presentó en entrenamiento un error de 4 % y en ajuste 8.8 %.

La salida de la red neuronal en entrenamiento se ve representada en la figura 4.27.

Para la validación del modelo, este entregó los resultados expuestos en la figura 4.28 en donde el error de validación es de 13.5 %

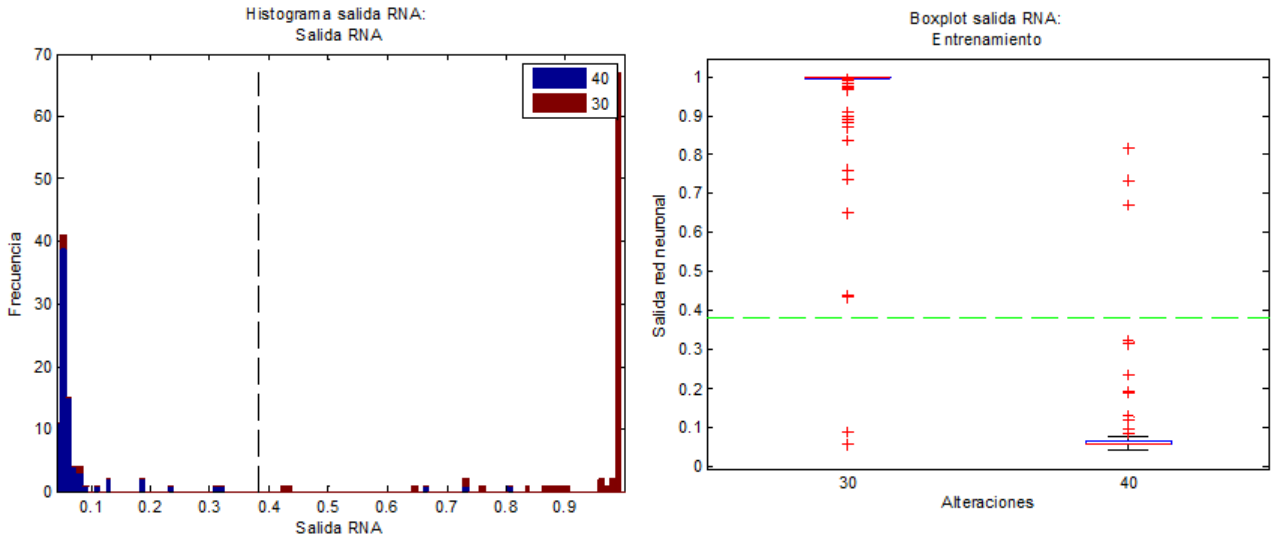


Figura 4.27: Salida RNA entrenamiento 30 vs 40

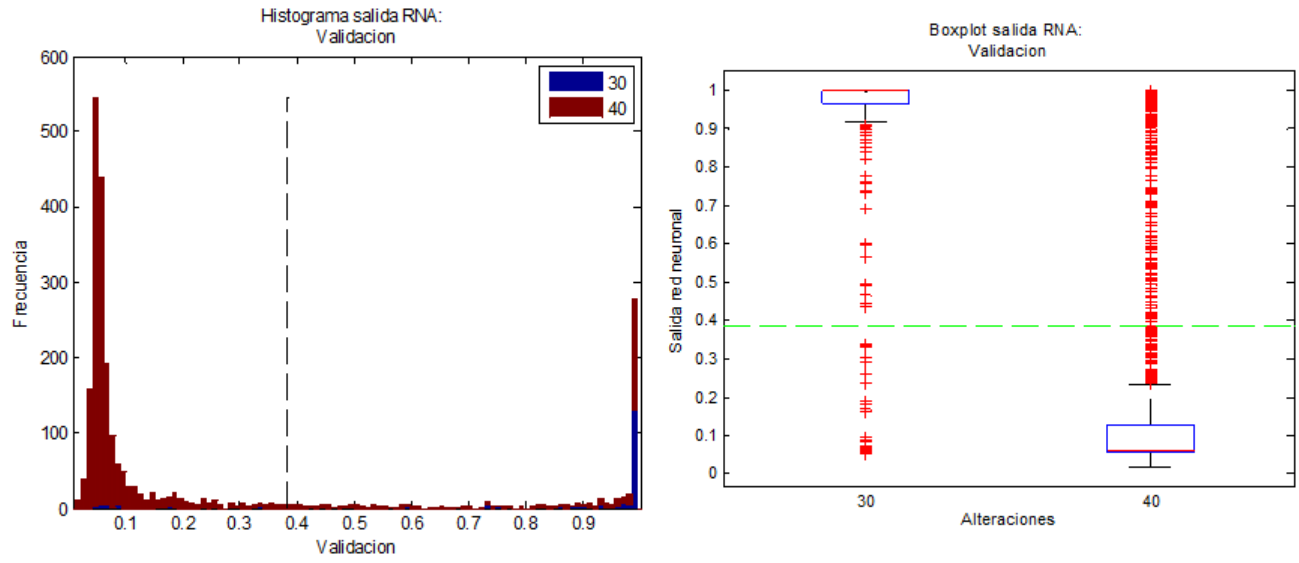


Figura 4.28: Salida RNA validación 30 vs 40

4.4.5. Modelo RNA

En la tabla 4.4 se puede apreciar, de manera compacta, los resultados obtenidos al realizar un árbol binario de clasificación con las distintas redes neuronales creadas, presentando un error promedio de un 22.8 % aproximadamente.

En la figura 4.29 se puede apreciar el error que se generó en cada uno de los nodos del árbol de separación, con respecto a los datos que llegan correctamente a ella.

Tabla 4.4: Validación red neuronal

	Alt. Predicha	30	31	40	50	61
Alt. Mapeada	Cantidad de Datos	575	750	1412	3471	902
30	261	87%	3%	10%	0%	0%
31	115	8%	89%	0%	4%	0%
40	2169	15%	20%	58%	5%	2%
50	4474	0%	4%	3%	75%	18%
61	91	0%	2%	6%	14%	78%
Acierto Promedio:				77.2%		

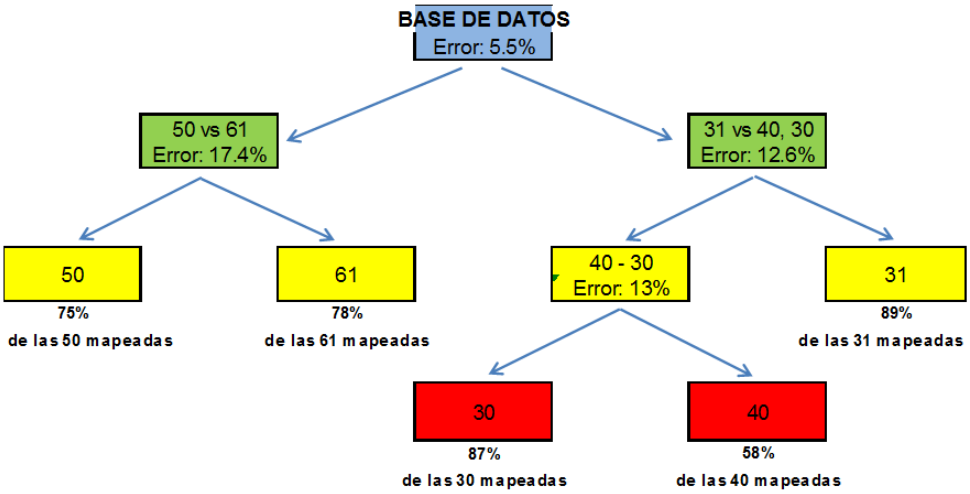


Figura 4.29: Validación árbol RNA

Capítulo 5

Análisis de resultados de los modelos

En el presente capítulo se realizará un análisis a los resultados obtenidos en el capítulo anterior, para cada una de las clasificaciones realizadas según lo expuesto.

5.1. Análisis de Clasificaciones

El análisis se ha realizado en base a cuadros de resumen de los distintas clasificadores y el **«AUC»** (área bajo la curva) generado por las curvas ROC, realizadas a partir de la función construida para Matlab [28]. Para el caso del k-mean cluster, se utilizó la manera paramétrica descrita en el capítulo antecedentes.

Separación según alteraciones 50, 61 de 30, 31, 40

La tabla 5.1 muestra a modo de resumen, las variables utilizadas en este nivel de clasificación, apreciando la importancia de la variable sintética, $\frac{Mg}{Al}$, la que resultó seleccionada en los cuatro algoritmos (como la primera). Esta variable trata de diferenciar las alteraciones supérgenas, debido a contenido de magnesio mayores en las alteraciones 30 (biotita secundaria), la 31 (Bt > Feld K) y las 40s (Clorita-Sericita, cuarzo y arcilla), a causa de la disolución de minerales, como la biotita, la clorita y otras micas.

Es posible ver en la tabla que los modelos no presentan grandes diferenciaciones, con una separación de 1.5 % entre la mejor y la peor (en validación). La figura 5.1 permite analizar la calidad de los clasificadores, permitiendo así, apreciar que no existe una gran diferencia entre ellos, aunque el mejor está constituido por las redes neuronales.

Tabla 5.1: Resumen separación 50, 61 vs 30, 31,40

Metodología	Variables seleccionadas	Error entrenamiento	Error Ajuste	Error Validación
Separación Simple	$\frac{Mg}{Al}$	5.3 %	6.4 %	7.0 %
K-mean Clúster	$\frac{Mg}{Al}$	6.2 %	6.7 %	6.4 %
Regresión Logística	$\frac{Mg}{Al}, K$	5.3 %	5.7 %	6.2 %
Red Neuronal Artificial	$\frac{Mg}{Al}, Ca$	4.7 %	5.8 %	5.5 %

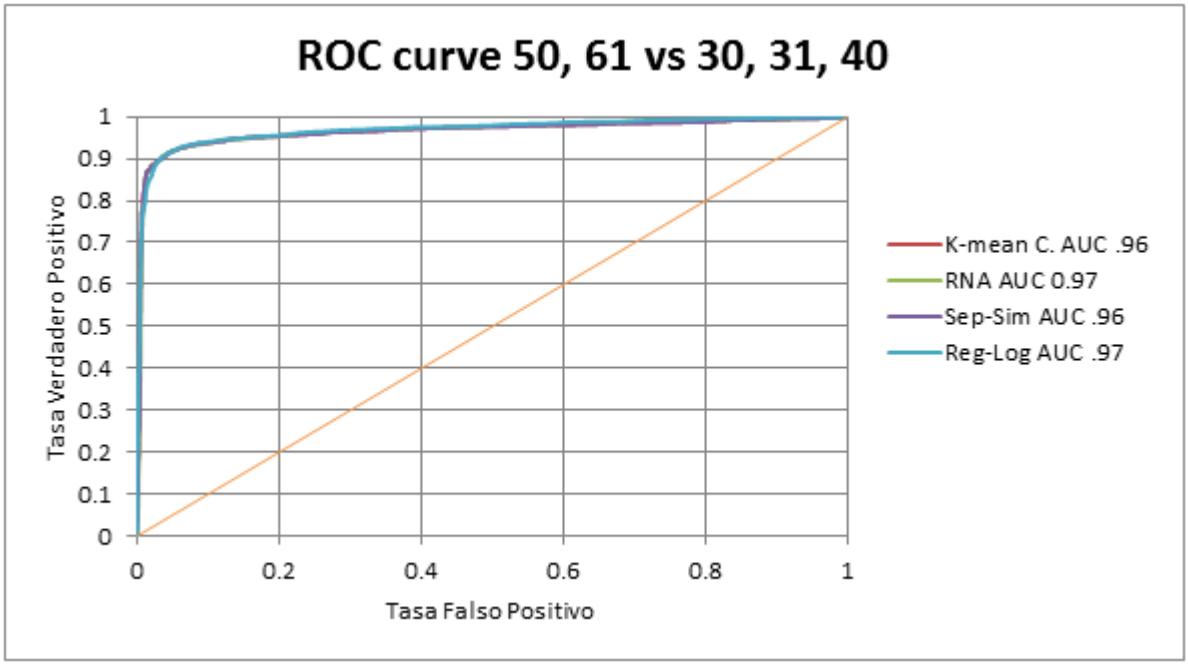


Figura 5.1: Curva ROC, Separación 50, 61 vs 30, 31, 40

5.1.1. Separación de alteración 50 de 61

La tabla 5.2 expone las variables seleccionadas para cada una de las metodologías utilizadas de clasificación. En este nivel de la separación, es posible ver que entre las distintas metodologías existe una importancia del “Aluminio”, el cual está presente en cada una de las variables escogidas en primer lugar, que puede ser reflejo de concentraciones de aluminio remanente luego de la lixiviación en zonas argílicas, presente en minerales como la alunita, gibbsita, esmectita entre otros.

Los errores presentados por las tres primeras metodologías, separación simple, K-mean Clúster y regresión logística no presentan diferencias en la validación presentando un 26.5 %. Siendo el menor error el alcanzado con la utilización de las redes neuronales con un 17.7 %.

La figura 5.2 muestra las curvas ROC para cada una de las metodologías. En esta se aprecia cómo el modelo de redes neuronales presenta el mayor AUC y la metodología k-mean

Tabla 5.2: Resumen separación 50 vs 61

Metodología	Variables seleccionadas	Error entrenamiento	Error Ajuste	Error Validación
Separación Simple	Al	26.2 %	27.4 %	26.5 %
K-mean Clúster	$\frac{Al+K}{Na+Ca+Mg}$, Rb, $\frac{3*Al}{K+Na}$, S	33.1 %	24.0 %	26.5 %
Regresión Logística	$\frac{Na}{Al}$, CuT, $\frac{K*Na}{Al}$	24.8 %	22.7 %	26.5 %
Red Neuronal Artificial	Al, Be, Cu, Cs, S	15.2 %	14.3 %	17.7 %

la menor, no obstante no hay que olvidar que la curva ROC de esta metodología no se realiza de manera directa.

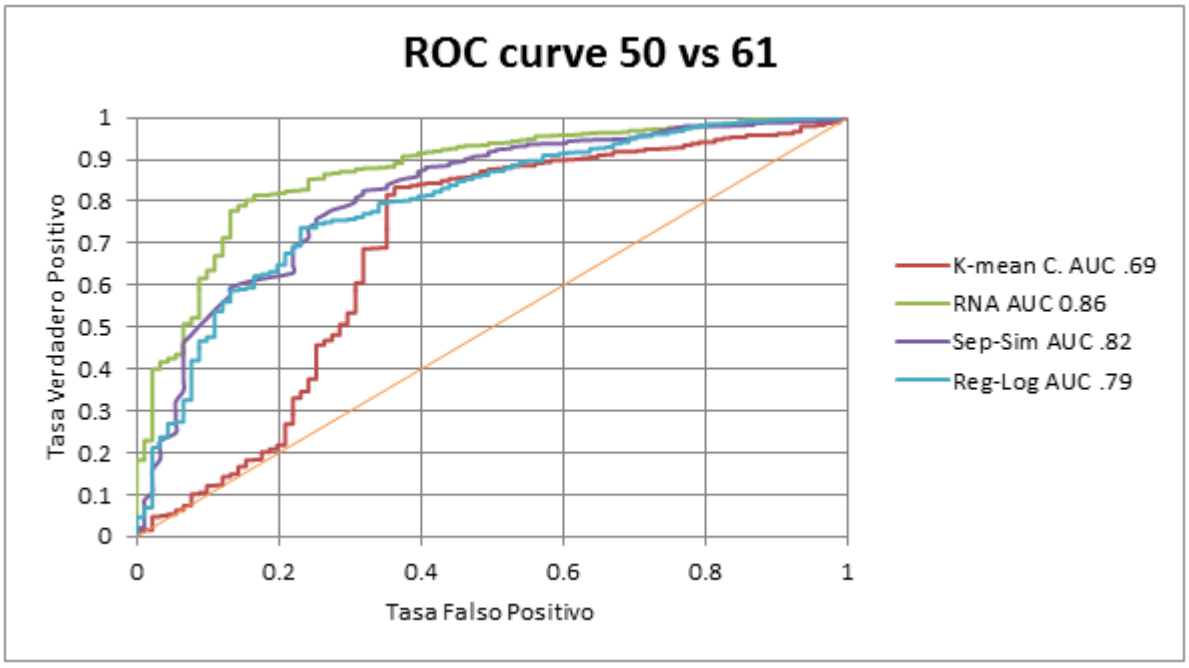


Figura 5.2: Curva ROC, Separación 50 vs 61

5.1.2. Separación de alteración 31 de 30, 40

La tabla 5.3 es un resumen del trabajo realizado para este nivel de clasificación, con la idea de lograr separar las alteraciones 31 (Feld K> biotita). Si bien los modelos no presentan grandes diferencias en cuanto al error, el mejor corresponde al de RNA como se aprecia en la tabla 2.20. Los elementos escogidos como variable principal en los distintos modelos son el aluminio y el magnesio los que se encuentran presentes en menor medida en la alteración 31, debido a su alto contenido de feldespato, ya que son difíciles de disolver en agua regia. En las alteraciones 40s, el aluminio y el magnesio se encuentran presenten en minerales como, la clorita y la sericita. En la alteración 30 estos elementos se encuentran contenidos en las

biotitas.

Tabla 5.3: Resumen separación 31 vs 30, 40

Metodología	Variables seleccionadas	Error entrenamiento	Error Ajuste	Error Validación
Separación Simple	Al	14.0 %	11.9 %	11.5 %
K-mean Clúster	Mg, $\frac{K}{Mg}$, $\frac{Ca+Na}{K+Al}$	13.3 %	12.3 %	11.6 %
Regresión Logística	Mg, Na, Ca	11.2 %	9.4 %	12.5 %
Red Neuronal Artificial	Al, Co	12.0 %	9.1 %	11.0 %

La figura 5.3 muestra las curvas roc graficadas para los distintos modelos, en donde el modelo de RNA y de separación simple alcanzan los mayores valores de área bajo la curva.

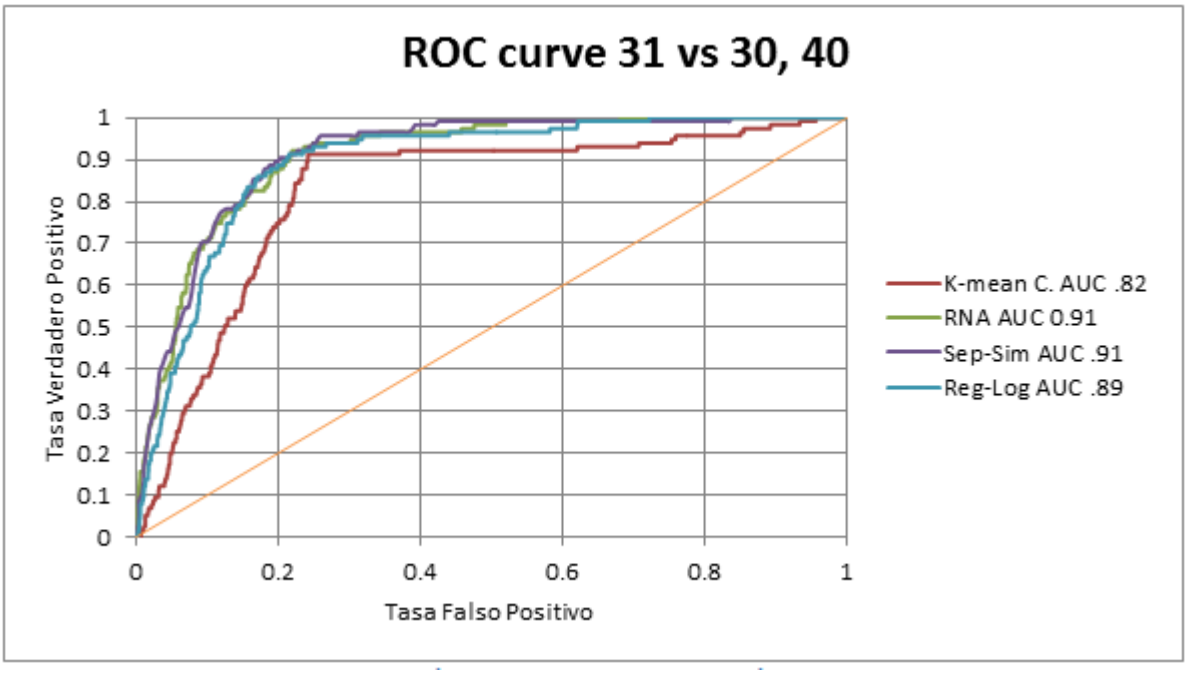


Figura 5.3: Curva ROC, Separación 31 vs 30, 40

5.1.3. Separación alteración 30 de 40

En la tabla 5.4 se observa que el rubidio es un elemento de gran importancia para lograr la separación presente en dos de las cuatro metodologías. Este se encuentra en mayor cantidad en la alteración 30, en donde se cree que entra como catión de intercambio por el potasio de las biotitas, debido a su potencial iónico similar. El escandio tomado por la regresión logística, se encuentra presente en mayores cantidades también en la alteracion 30, ya que al igual que el rubidio, puede reemplazar al Mg^{2+} en la biotita.

Tabla 5.4: Resumen separación 30 vs 40

Metodología	Variables seleccionadas	Error entrenamiento	Error Ajuste	Error Validación
Separación Simple	Rb	15.6 %	14.0 %	16.8 %
K-mean Clúster	Al, $\frac{K}{Mg}$	25.1 %	26.3 %	16.9 %
Regresión Logística	Sc, $\frac{3Al}{K+Na}$, Se	7.9 %	15.1 %	16.9 %
Red Neuronal Artificial	Rb, Na, Ni, Sc, Sr	3.5 %	11.0 %	13.5 %

La figura 5.4 muestras las curvas roc para cada método, en donde el peor de los métodos por un poco es el de separación simple.

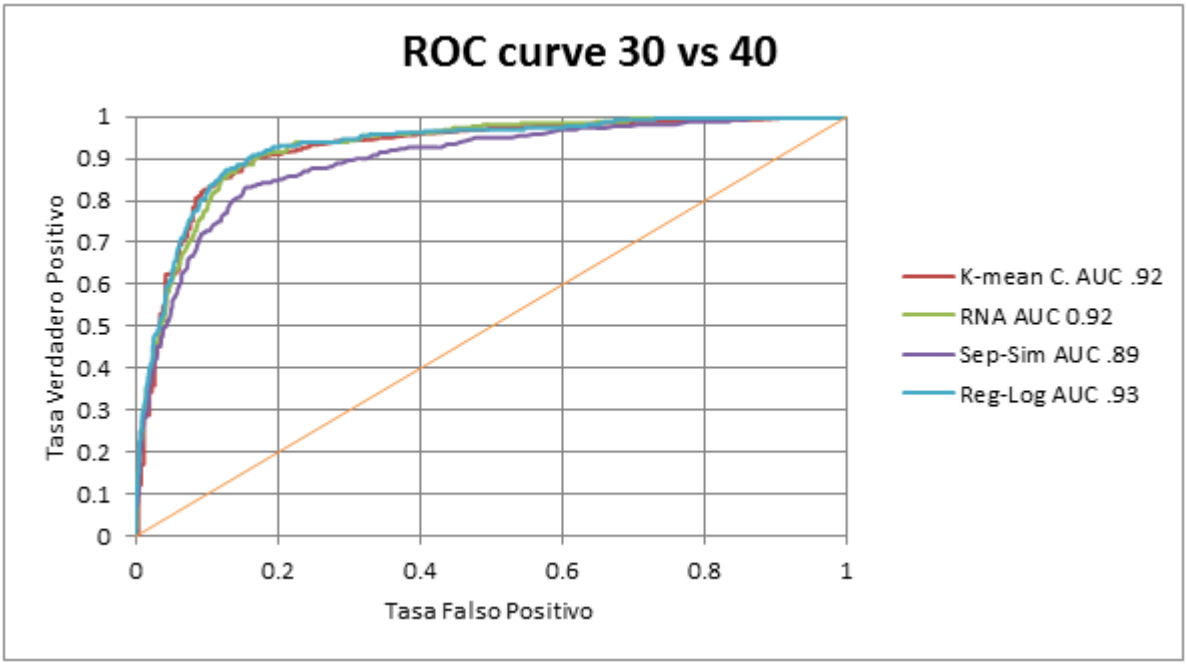


Figura 5.4: Curva ROC, Separación 30 vs 40

5.2. Análisis general

A manera de resumen de los análisis anteriores se tiene la tabla 5.5 en la que se muestra el promedio de los aciertos de cada una de las metodologías.

Tabla 5.5: Resumen desempeño

Metodología	Acierto promedio [%]	Desviación Estándar [%]
Separación Simple	71.5	16.1
K-mean Clúster	71.7	9.1
Regresión Logística	73.6	11.5
Red Neuronal Artificial	77.2	12.4

El modelo que presentó el menor error es el de redes neuronales con un 22.8 %. Este modelo se analizará más a fondo en el próximo capítulo y se estudiará cómo implementar mejoras en él.

Capítulo 6

Análisis de modelo final utilizando RNA

En el presente capítulo se estudiará el modelo generado con redes neuronales a fondo, el cual presentó el menor error final y mejor calidad en general, según el AUC de las curvas ROC en cada uno de los pasos.

6.1. Iteración de modelos

Producto de la aleatoriedad de los pesos iniciales de las RNA, para estudiar los modelos se crearon 100 modelos en cada una de las divisiones realizadas para clasificar la información, con la idea de estudiar su desempeño, analizar su error de entrenamiento, ajuste y validación conjuntamente, lo que permitirá conocer si los rangos de errores con los que trabajan las distintas redes neuronales están en lo correcto. A continuación se expondrán los resultados obtenidos en este paso.

6.1.1. Separación alteración 50, 61 de 30, 31 y 40

El desempeño de los modelos creados en este nivel de separación se puede apreciar en la figura 6.1, donde el modelo final trabaja con un error de separación de 5.5 % para validación. Se puede apreciar que los modelos que tienen menor error de entrenamiento y ajuste, presentan un menor error de validación.

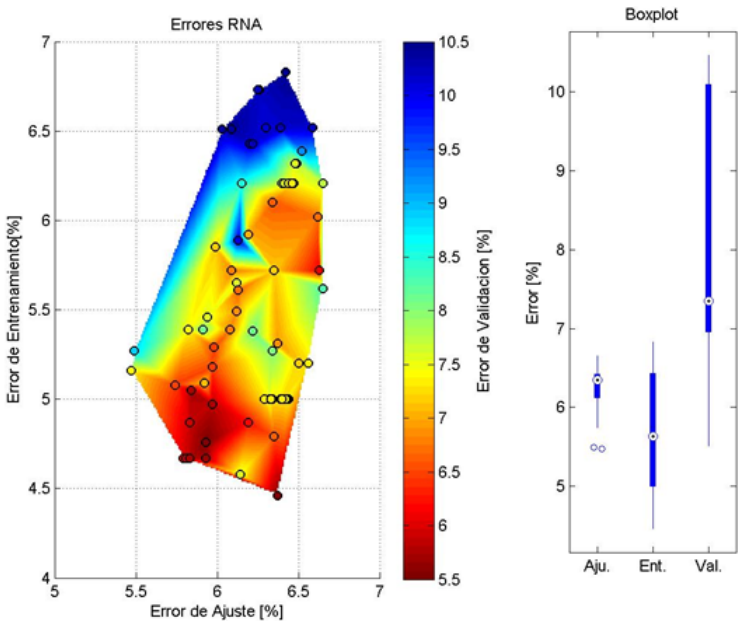


Figura 6.1: Errores Separación 50, 61 de 30, 31 y 40

6.1.2. Separación alteración 50 de 61

El desempeño de los modelos creados en este nivel de separación se puede apreciar en la figura 6.2. El modelo con el cual se trabaja en este nivel tiene un error de 17.7 %, si bien el error es menor que el error promedio (25 %) de las distintas iteraciones, se observa cierto orden en la relación de los errores (a menor errores de entrenamiento y ajuste menor error en validación). Además, esta resulta ser una etapa de suma importancia a nivel de separación debido a que más del 50 % de los datos pertenece a la clase de alteración definida como 50 (Sericita-Cuarzo y Sericita-Cuarzo-Arcilla), y menos del 1 % de las muestras pertenecen a la alteración 61: argilización avanzada, por lo que la decisión de tomar un modelo con el mejor desempeño, se basó en la idea de disminuir el error general promedio de la clasificación de las muestras.

6.1.3. Separación alteración 31 de 30, 40

La figura 6.3 muestra el resultado de los errores de las iteraciones. En la imagen no se observa un orden claro de la organización de los errores. El modelo con el que se trabaja tiene un 11 % de error en validación similar al promedio de estos en los distintos modelos estudiados.

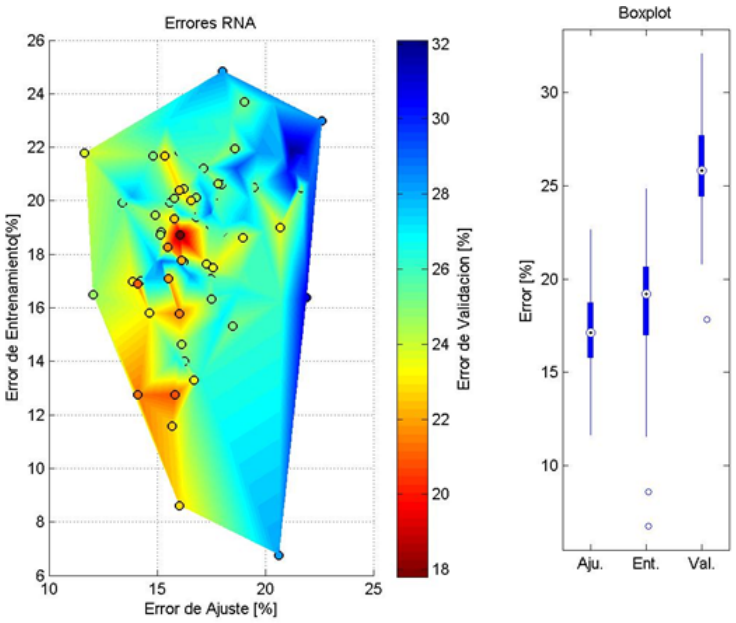


Figura 6.2: Errores Separación 50 de 61

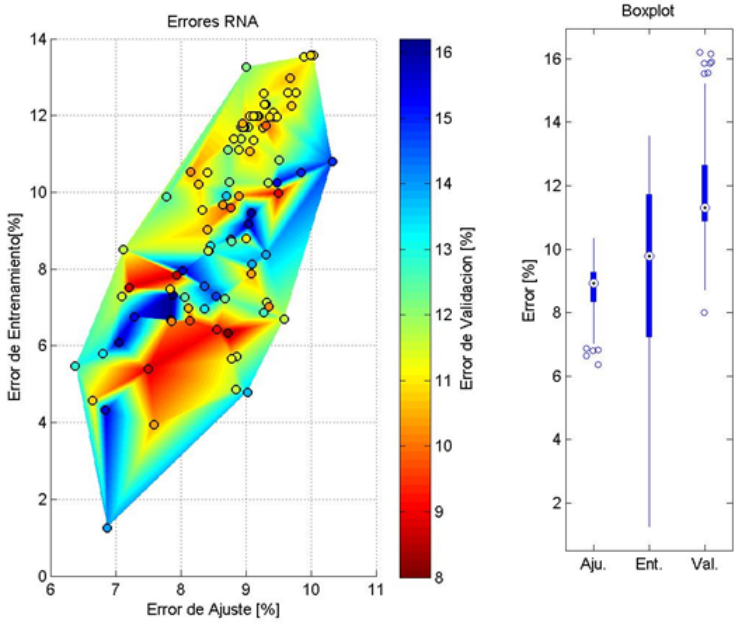


Figura 6.3: Errores Separación 31 de 30 y 40

6.1.4. Separación alteración 30 de 40

En la figura 6.4 se observa los errores producidos al crear los distintos modelos. En la imagen se puede apreciar un orden en cuanto a la magnitud del error, es decir, a pequeños errores de entrenamiento y ajuste, se tienen errores más pequeños en validación, debido a

esto, se trabaja con un modelo que posee un error de 13.5 % para validación.

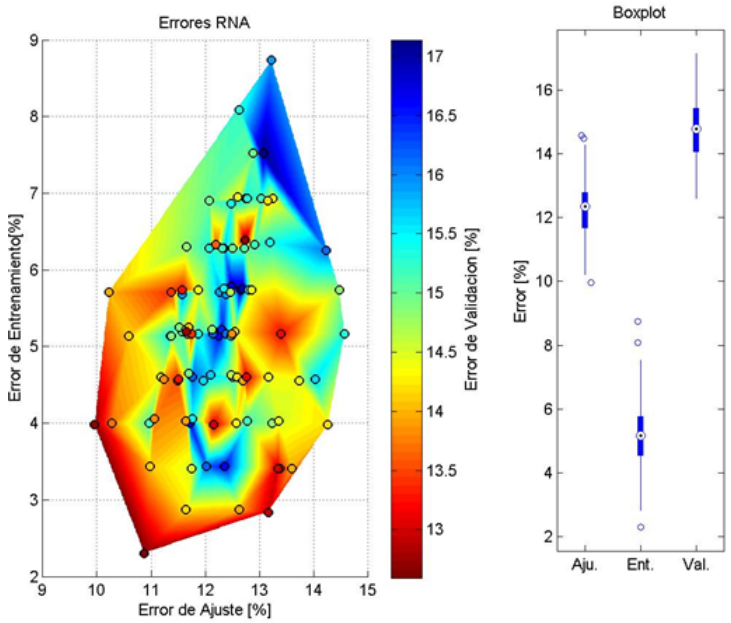


Figura 6.4: Errores Separación 30 de 40

6.2. Análisis de las divisiones del modelo

Al analizar el modelo que se realizó a partir de redes neuronales, se aprecia que la clasificación de las alteraciones tipo 40 (Clorita-Sericita-Arcillas y Clorita-Sericita-Cuarzo) es bastante baja con un 58 % de las muestras bien clasificadas, siendo esta alteración una de las más importante en la base de datos (cerca de un 30 % de los datos utilizados), quedando un 20 % de las muestras clasificadas como alteración 31, un 15 % se clasifica como 30 (esta clasificación es final ya que se clasifica 30 vs 40), y un 7 % se va a la rama de las 50 y 61. Debido a esto se decidió incorporar un paso más en la separación de alteraciones para así separar las alteraciones 31 de las 40, con el objetivo de mejorar la clasificación general de esta.

6.2.1. Separación alteraciones 40 de 31

Modelo de redes neuronales: 40 vs 31

El modelo de RNA utilizado para la separación de las alteraciones tipo 40 de la 31, consta de 4 neuronas, utilizando las siguientes variables en orden descendente de importancia. Se

encuentran junto a ellas los errores asociados al momento de su selección:

1. 3Al/(K+Na): Entrenamiento: 13.4 Ajuste: 16.7 %
2. Be: Entrenamiento: 10.2 %, Ajuste: 12.6 %
3. Fe: Entrenamiento: 8.5 %, Ajuste: 11 %
4. S: Entrenamiento: 6.07 %, Ajuste: 8.46 %

El algoritmo se detuvo con la variable: Sn, la cual presentó en entrenamiento un error de 4.9 % y en ajuste un 9.5 %, deteniendo el algoritmo.

La salida de la red neuronal se ve representada por la figura 6.5 en la que se aprecia en línea punteada el umbral de corte (0.64).

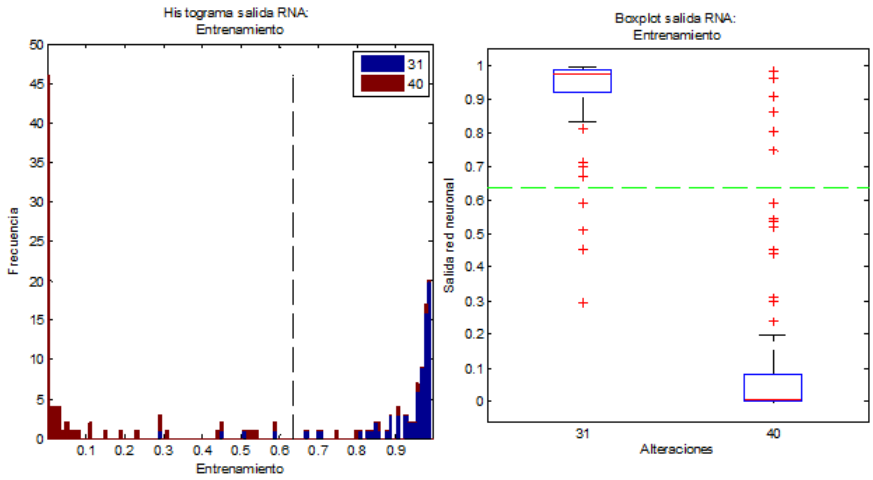


Figura 6.5: Salida RNA entrenamiento 31 vs 40

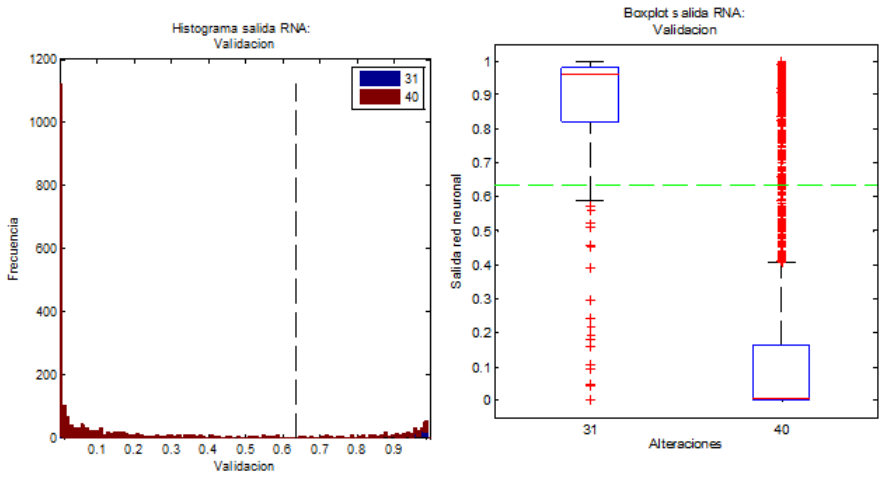


Figura 6.6: Salida RNA validación 31 vs 40

Para la validación del modelo, este entregó los resultados expuestos en la figura 6.6, donde el error de validación es de 15.7 %.

Iteración Separación alteración 40 de 31

La figura 6.7 muestra el resultado de los errores de las iteraciones, en la imagen no se ve un orden muy claro de la organización de los errores. El modelo con el que se trabaja tiene un 15.7 % de error en validación, similar al promedio de los errores estudiados de los distintos modelos.

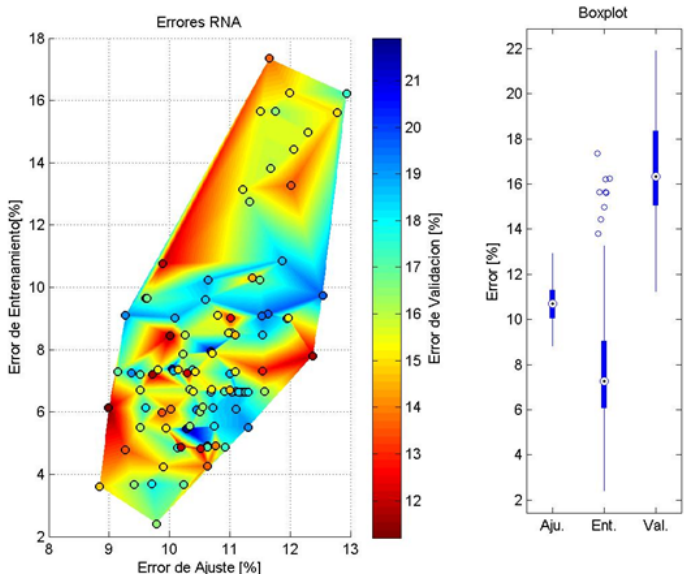


Figura 6.7: Errores Separación 31 de 40

6.2.2. Estudio Separación Alteraciones Agrupadas

En esta sección se estudiará la posible separación de alteraciones agrupadas 51: sericita cuarzo y 52 sericita cuarzo-arcilla en las tipo 50 y 40: clorita sericita-arcillas 41: clorita sericita cuarzo en las tipo 40, cuyo agrupamiento se realizó como resultado de la alta tasa de traslape que tenían como familia. De esta manera, se busca validar el criterio que se tomó en la confección del modelo.

Alteraciones tipo 50

Se estudió la posible separación de las alteraciones agrupadas como 50, la alteración 51: sericita cuarzo y 52: sericita cuarzo arcillas. Para esto se realizó un modelo de clasificación utilizando RNA que ocupa las siguientes variables:

1. Mo

2. Co
3. $\frac{Al + K + Na}{Ca + Mg}$

El modelo presentó un 34 % y 40.2 % de error en entrenamiento y ajuste, respectivamente. La figura 6.8, muestra el modelo creado utilizando la base de datos de entrenamiento.

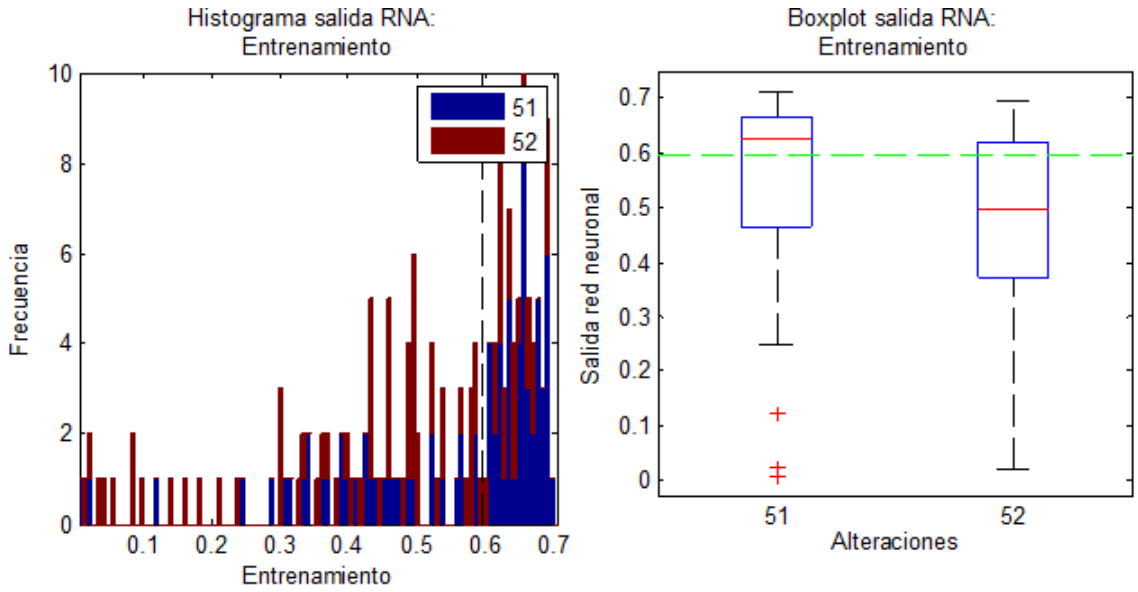


Figura 6.8: Modelo creado 51 vs 52

La figura 6.9 muestra el modelo utilizando la base de datos de validación, presentando un error de 43,5 %.

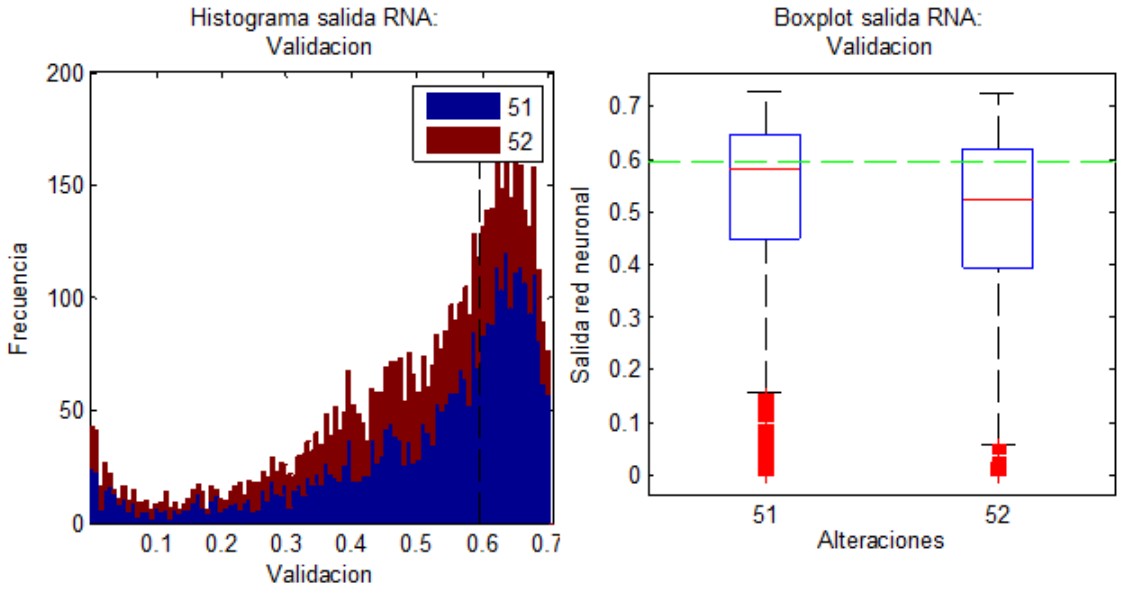


Figura 6.9: Validación del modelo 51 vs 52

La curva ROC del modelo, que se aprecia en la figura 6.10, muestra una calidad de clasificación muy baja con un 0.58 de área bajo la curva.

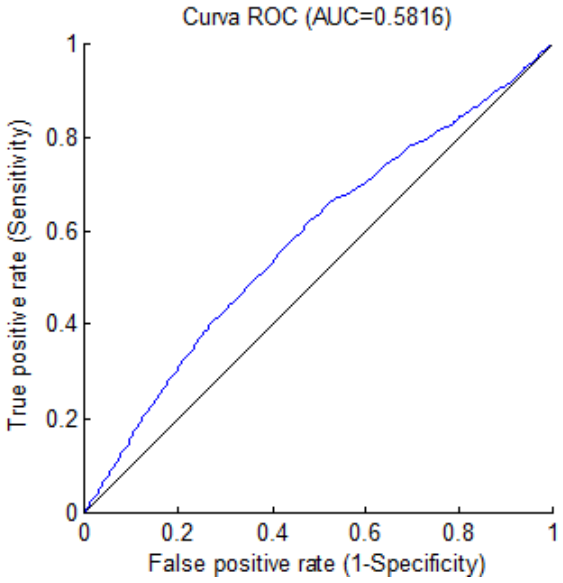


Figura 6.10: Curva roc del modelo 51 vs 52

Alteraciones tipo 40

Con el fin de estudiar una posible división entre esta familia, se realizó un modelo de redes neuronales para observar el comportamiento del modelo, que según se sabe es de muy difícil separación debido al gran traslape de familias que se tenía.

1. Sc
2. Cr

El modelo presentó en entrenamiento un error de 28.1 % y para el ajuste de un 29.5 %. La figura 6.11, muestra la respuesta del clasificador con la base de datos con la que se creó.

La validación del modelo se ve representada en la figura 6.12, donde se probó con nueva información. El modelo creado presenta un error de un 31.1 %.

La figura 6.13 muestra la calidad de clasificación que se obtiene con el modelo, siendo esta bastante escasa.

De esta manera, se confirma la decisión tomada de agrupar estas alteraciones en 2 grupos. El añadir una fase más de separación se traduciría en un modelo menos robusto debido a su difícil separación (también se debe recordar que se añadiría un nivel más a

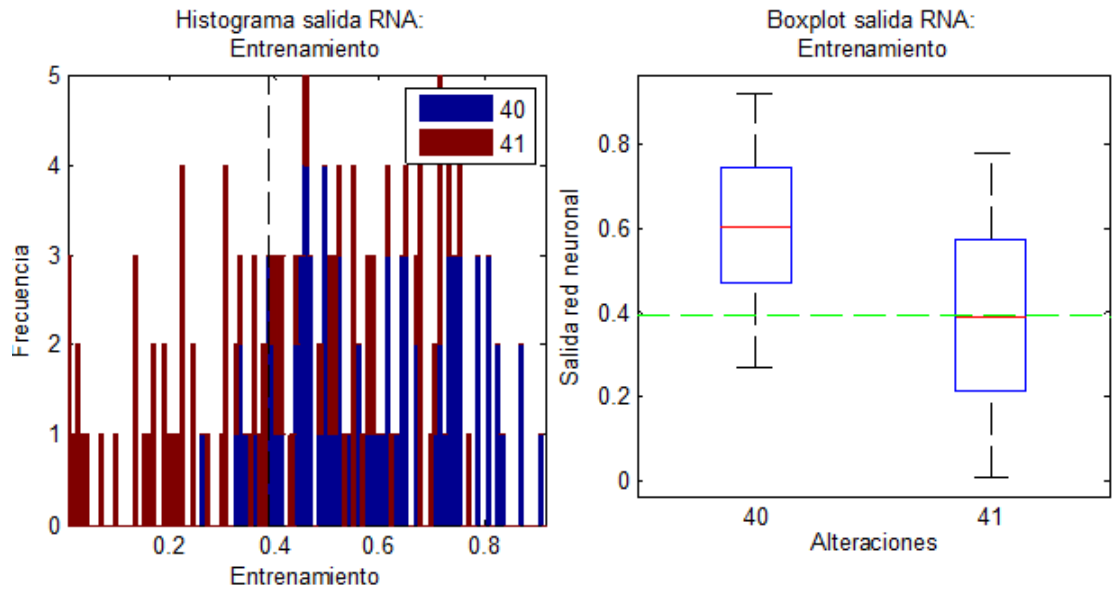


Figura 6.11: Modelo creado 40 vs 41

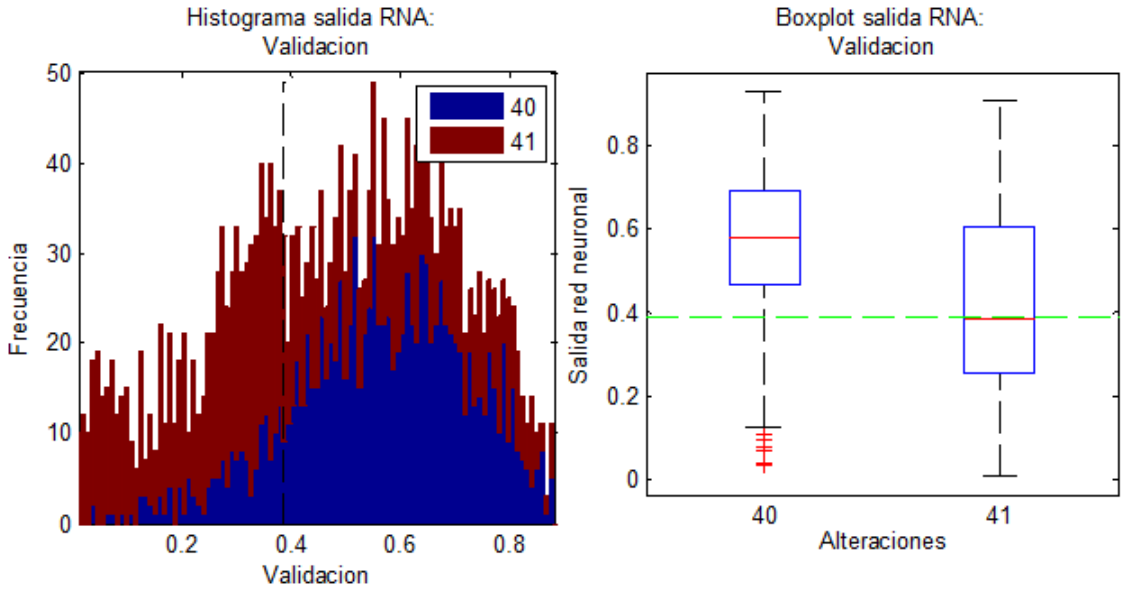


Figura 6.12: Validación modelo 40 vs 41

la clasificación que según lo analizado sería un nivel problemático). Los resultados antes expuestos confirman la calidad de los estadísticos creados para el estudio, traslape de poblaciones y clustering de categorías.

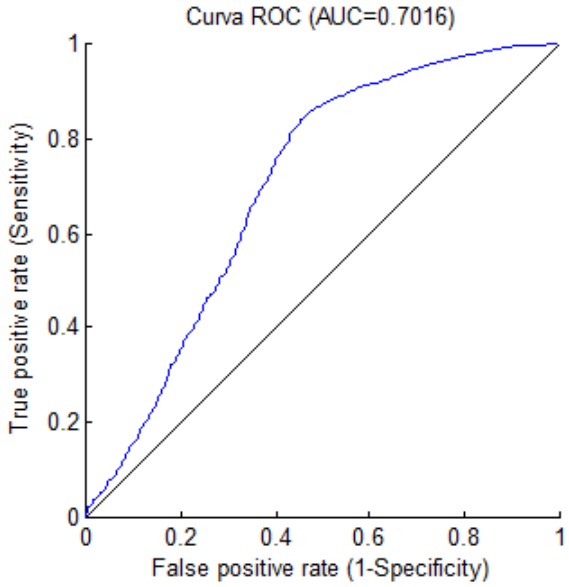


Figura 6.13: Curva roc del modelo 40 vs 41

6.3. Modelo final

En la figura 6.14 se puede apreciar la estructura del modelo final con la nueva separación añadida (separación 40 de 31). Se observa en ésta el porcentaje de los datos que son correctamente clasificados de los originales según clase.

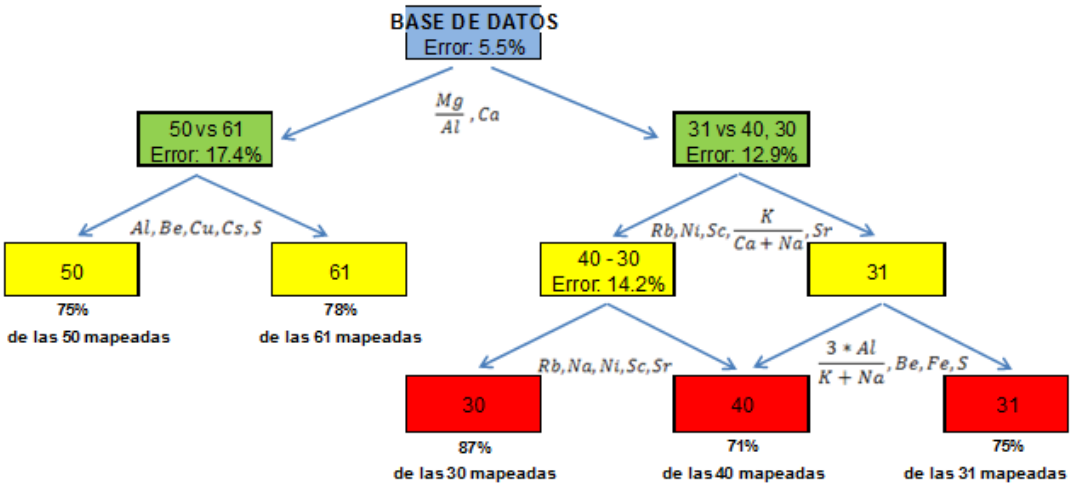


Figura 6.14: Modelo de clasificación final

La tabla 6.1 muestra de manera compacta los resultados finales obtenidos al realizar el árbol de clasificación con los distintos modelos de redes neuronales creados en cada paso, presentando un error promedio de un 22.9 %.

Tabla 6.1: Resultados modelo final

	Alt. Predicha	30	31	40	50	61
Alt. Mapeada	Cantidad de Datos	578	358	1801	3471	902
30	261	87%	3%	10%	0%	0%
31	115	9%	75%	13%	4%	0%
40	2169	15%	7%	71%	5%	2%
50	4474	0%	3%	5%	75%	18%
61	91	0%	0%	8%	14%	78%
Acierto Promedio:				77.1%		

Con los análisis realizados la adición de una fase clasificadora de alteraciones 40 contra 31, resulta beneficiosa para el estudio, al aumentar el porcentaje de datos bien clasificados en la alteración 40, viéndose reflejado en una disminución de la varianza de los errores obtenidos de clasificación final por clase, sin variar en el promedio, como se puede ver en la tabla 6.2.

Tabla 6.2: Comparación Modelos

Alteración	Porcentaje bien clasificado Modelo RNA	Modelo RNA Final
30	87 %	87 %
31	75 %	89 %
40	71 %	58 %
50	75 %	75 %
61	78 %	78 %
Varianza	0.004	0.015
Desviación Estándar	6 %	12 %
Promedio	77.1 %	77.2 %

Capítulo 7

Análisis de resultados modelo final

El presente capítulo, busca explicar los posibles motivos de la clasificación final que realizó el modelo utilizando la estructura de árbol de clasificación, con el objetivo de entregar aspectos geológicos que se puedan ver representados a través de las variables seleccionadas, y cómo las muestras son clasificadas analizando con cuales se tiene una mayor afinidad (dificultad de separación).

7.1. Histograma de Clasificación

Para el estudio de los datos predichos se utilizará la base de datos de validación que fue usada para la corroboración de los modelos, siendo está representada, a través de histogramas sobre las alteraciones predichas según los datos de alteración mapeada.

7.1.1. Alteración 30: Potásico (Sólo Biotita secundaria)

Del total de 326 muestras, las que representan el 3 % de la base de datos, se expone en la figura 7.1 la clasificación correcta de esta alteración, la cual se debe a dos factores; el primero es su "alto" índice de $\frac{Mg}{Al}$, ya que al poseer altas concentraciones de magnesio, contra concentraciones más pequeñas de este elemento en alteraciones más superficiales (tardías : 50s y 61) y el segundo al contenido de rubidio que tiene, que debido a un potencial iónico muy parecido al del potasio, se intercambia con este, en minerales como la biotita, la que está presente en esta alteración. Problemas de clasificación pueden verse originados ya que en la alteración 31 (Feld K > biotita), también hay presencia de biotita, y en alteraciones 40 (Clorita-Sericita: -Arc y Qz), está presente la clorita, que también podría contener rubidio en

su estructura. Por estas razones esta alteración queda clasificada con mayor afinidad con las alteraciones 31 y 40s que con las 50s y 61.

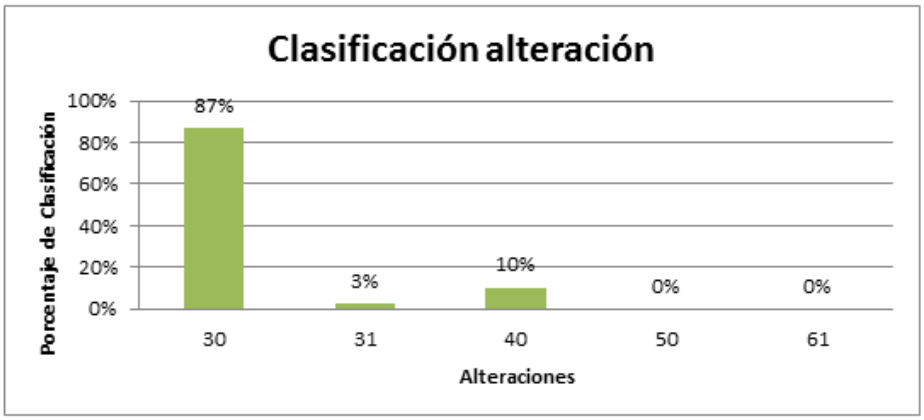


Figura 7.1: Histograma clasificación potásico (Biotita secundaria)

7.1.2. Alteración 31: Feldespato potásico > Biotita

Del total de 164 muestras mapeadas como alteración 31, que representan el 1.6 % de la base de datos, se expone en la figura 7.2 la predicción de estos, según el modelo creado alcanzando un 75 % de acierto. Los motivos de este valor se deben a separaciones realizadas según el índice $\frac{Mg}{Al}$, el cual es alto en esta alteración, al poseer una medianamente baja concentración de magnesio, que en el caso de las alteraciones 50 y 61 es casi cero y una muy baja concentración de aluminio, permiten diferenciarla de estas alteraciones (4 % de las muestras es clasificada como 50 debido a ciertas mediciones que tienen muy poco magnesio), un 9 % de estas se clasifica como alteración 30 debido al contenido de Rubidio que se cree se ve contenido en la biotita como se explicó con anterioridad, y un 13 % es clasificado como 40 debido a muestras que presentaban mayores concentraciones de aluminio.

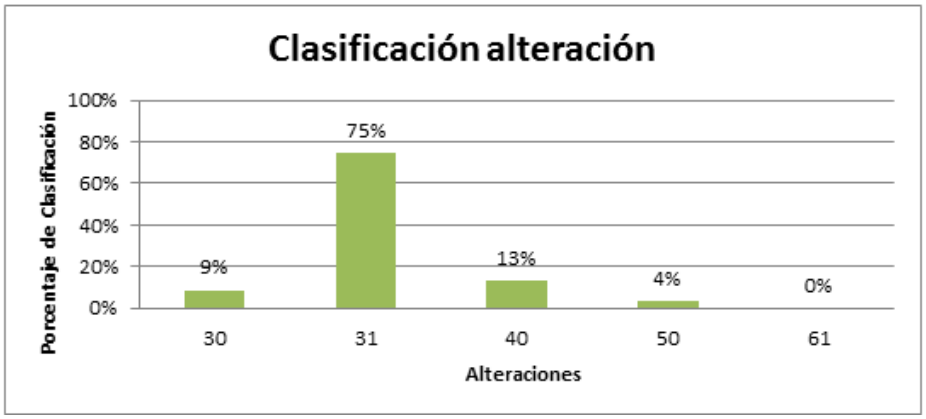


Figura 7.2: Histograma clasificación Feld. K > Biotita

7.1.3. Alteración 40: Clorita-Sericita-Arcillas y Clorita-Sericita-Cuarzo

Del total de 3040 muestras mapeadas como alteración 40 y 41 que representan el 30 % de la base de datos utilizada, se expone en la figura 7.3 la distribución de clasificación con un 71 % de los datos bien catalogados, teniendo como pasos para llegar a él, poseer un “alto” índice de $\frac{Mg}{Al}$ (un promedio 0.7 versus un 0 de alteraciones 50s y 61). El magnesio y aluminio contenido en estas alteraciones proviene de la disolución de clorita y algunas micas en agua regia. La semejanza que presenta con alteraciones como la potásica (biotita secundaria), se debe seguramente a contenidos de rubidio que pueden reemplazar al potasio, además en la alteración clorita-sericita-arcilla se conservan comúnmente relictos de alteración potásica (biotización y feldespato potásico).

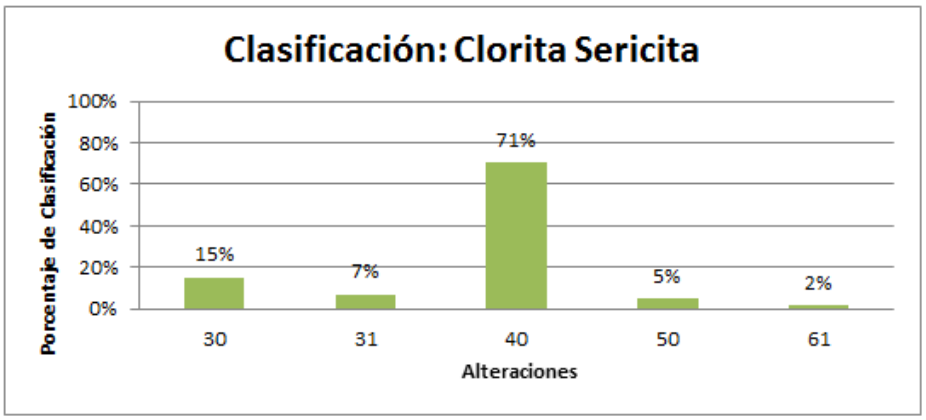


Figura 7.3: Histograma clasificación tipo 40

En la figura 7.4 se puede apreciar la separación más detallada, en donde se puede observar cómo la alteración 41 tiene una mayor semejanza que la 40 con la alteración 31, feldespato K > biotita, lo cual se podría deber al hecho que esta posee niveles más bajos de aluminio, producto que en comparación con la alteración 40, no posee arcillas como la montmorillonita, que son solubles en agua regia liberando aluminio, lo que significaría una mayor semejanza con esta alteración.

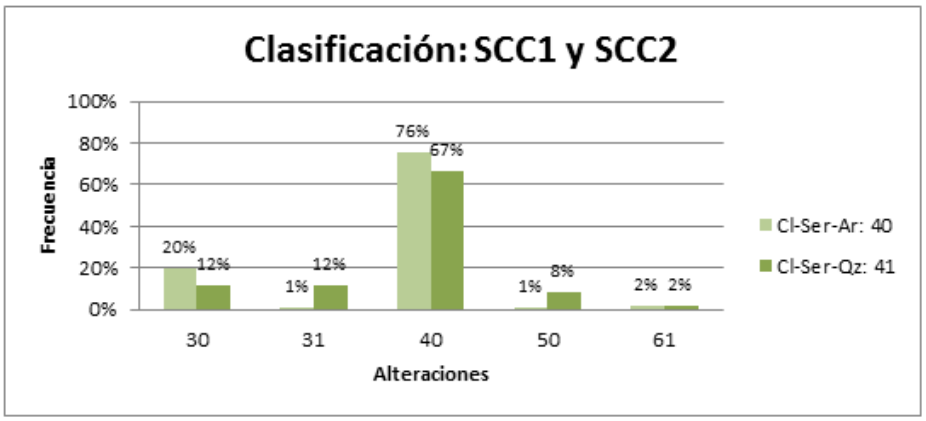


Figura 7.4: Histograma clasificación Cl-Ser: -Arc y -Qz

7.1.4. Alteración 50: Sericita-Cuarzo y Sericita-Cuarzo-Arcilla

Del total de 6640 muestras mapeadas como alteración 51 y 52 que representan el 65 % de la base de datos, se aprecia en la figura 7.5 la distribución de clasificación con un 75 % de los datos bien catalogados como alteración Sericita-Cuarzo (51) o Sericita-Cuarzo-Arcilla (52), quedando un porcentaje no menor predicho como alteración argílica supérgena (61), lo cual se puede explicar por la cantidad similar de aluminio que presentan, al tener esta última minerales como la sericita, alunita, pirofilita y diásporo en una matriz de sílice y sericita.

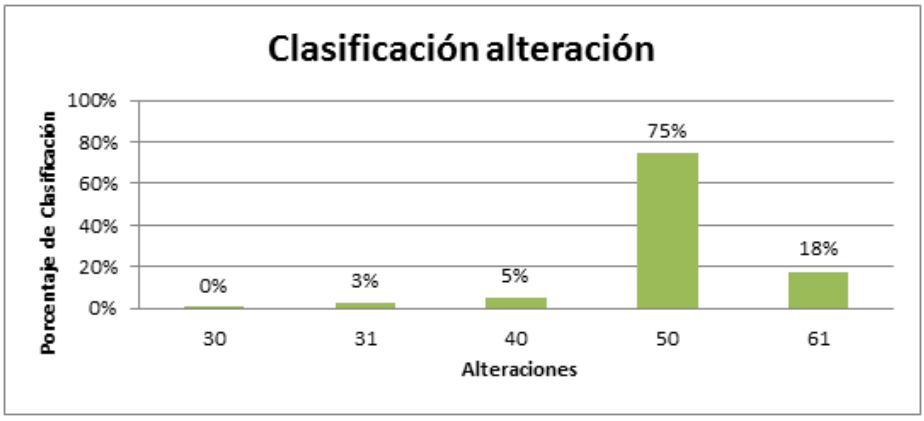


Figura 7.5: Histograma clasificación tipo 50

La figura 7.6 muestra que no existe una gran diferenciación (lo que era esperable por la gran semejanza de estas alteraciones) entre la alteración 51 y 52.

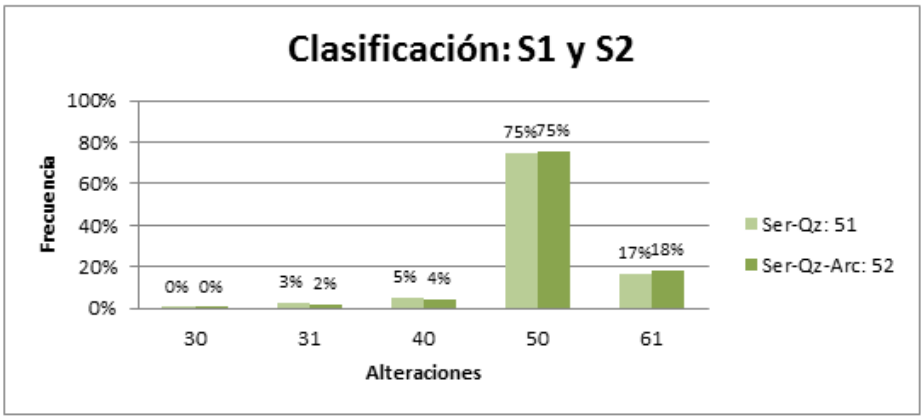


Figura 7.6: Histograma clasificación Ser-Qz y Ser-Qz-Arc

7.1.5. Alteración 61: Argilización supérgena

Del total de 110 muestras mapeadas como alteración argílica supérgena que representan el 1 % de los datos, la figura 7.7 muestra la distribución de clasificación con un 78 % de los datos bien catalogados. Se observa que 14 % de los datos es clasificado como alteración de tipo 50s, por las mismas razones antes expuestas. El 8 % que se encuentra predicho como alteración 40, lo que podría deberse a biotita o clorita remanente de procesos geológicos anteriores.

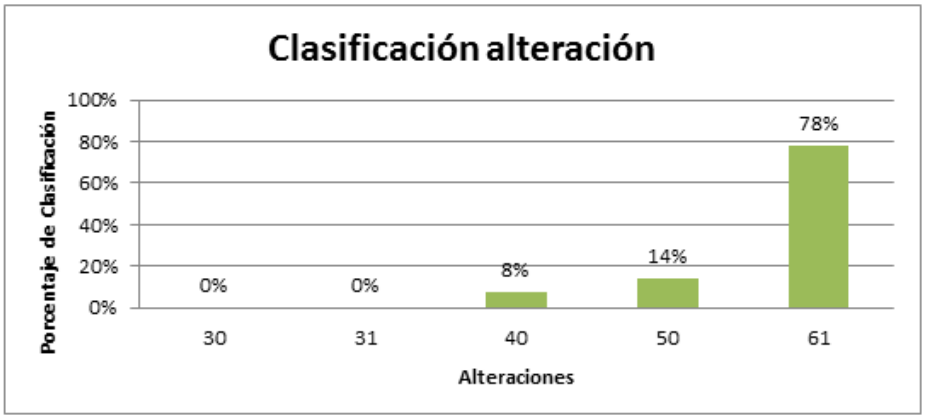


Figura 7.7: Histograma clasificación Argilización Supérgena

7.2. Comparación sondajes mapeados y predichos visualmente

En esta sección se exhibirán los sondajes desplegados contenidos en la base de datos de validación, para lograr contrastar de manera visual, el modelo real mapeado con el modelo

predicho. La comparación se realizará estudiando una vista en planta y cuatro secciones realizadas en el yacimiento, detalladas en la figura 7.8.

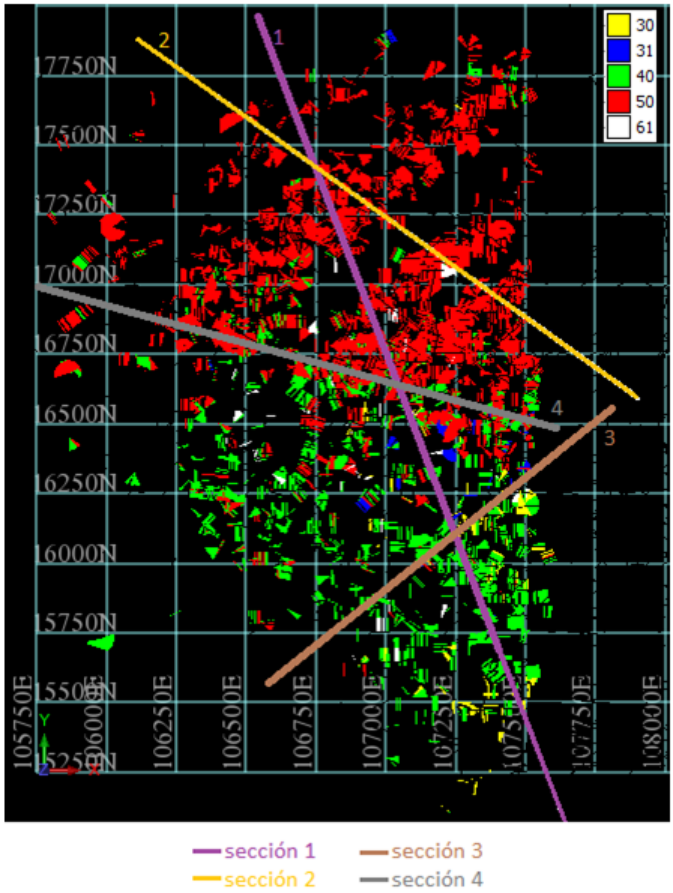


Figura 7.8: Secciones de estudio

En la figura 7.9 se observa a la izquierda los sondajes mapeados desplegados, y en la imagen de la derecha, los sondajes predichos por el modelo de clasificación. En ella se puede apreciar que encuentra de buena manera el límite entre los 2 grandes grupos de clasificación 50-61 con 40-31-30.

En la figura 7.10 se plasma lo comentado anteriormente, además se aprecia una sobreestimación de la alteración argílica supérgena (61) y de la alteración feldespato potásico > biotita (31). La zona donde se mapea esta, es la misma donde se predice que ocurre, pudiendo establecer un nodo feldespático en ese lugar. Es importante ver cómo el modelo predice esta alteración en profundidad, lo que es correcto según los modelos de formación de alteraciones (es una etapa más temprana).

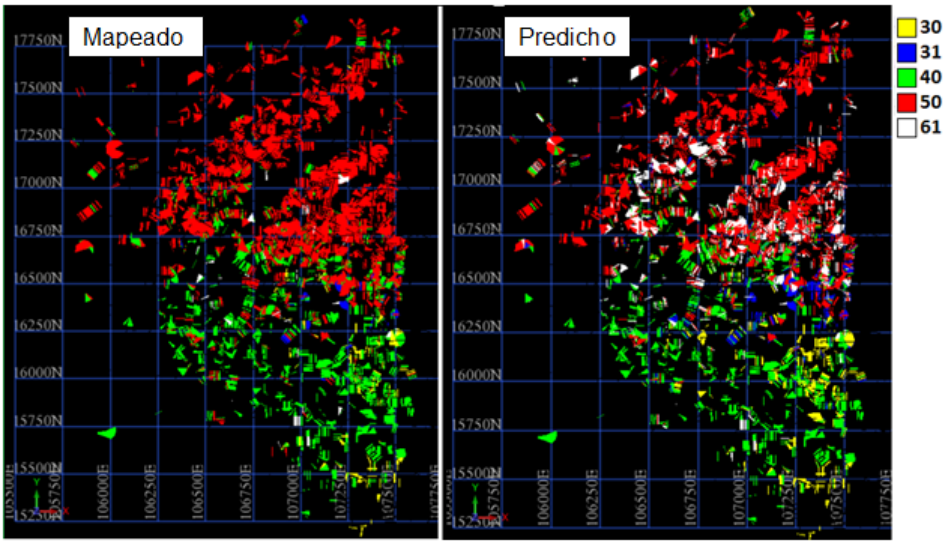


Figura 7.9: Vista en planta sondajes mapeado y predicho

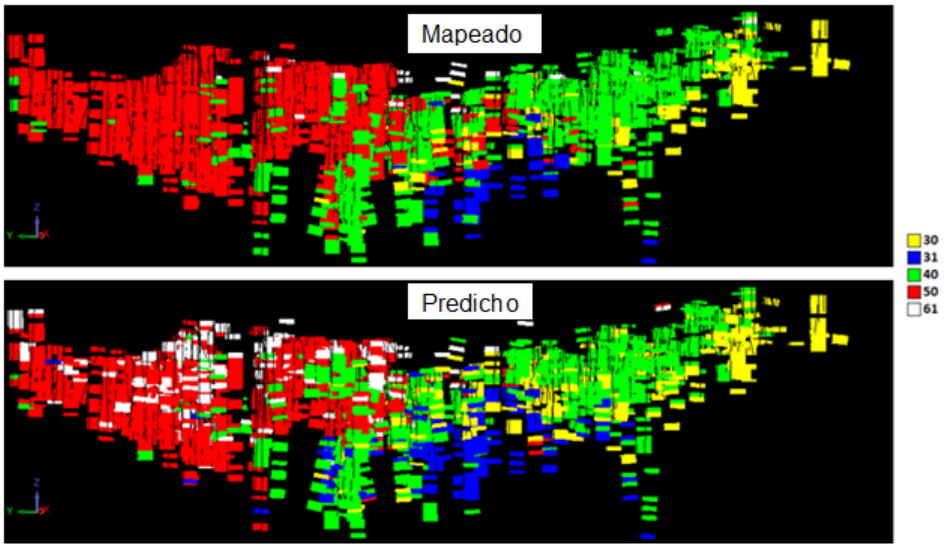


Figura 7.10: Vista de sección 1 modelo mapeado y predicho

En la figura 7.11 se aprecia una sección de la zona dominada por las alteraciones cuarzo-sericita (50s) y argílica supérgena (61). Se puede apreciar cómo el modelo sobreestima esta última, pero es importante notar como en su mayoría se encuentra de manera más superficial que la alteración cuarzo-sericita, siendo relevante ya que se asocia a etapas más tardías de formación y a procesos supérgenos.

En la figura 7.12 se aprecia una sección de la zona dominada por las alteraciones clorita-sericita (40s), solo biotítica secundaria (30) y feldespato mayor que biotita (31), pudiéndose notar sobreestimación de la alteración 30, pero una ubicación espacial parecida a la mapeada.

La figura 7.13 es una sección de la zona de contacto de entre las alteraciones cuarzo-

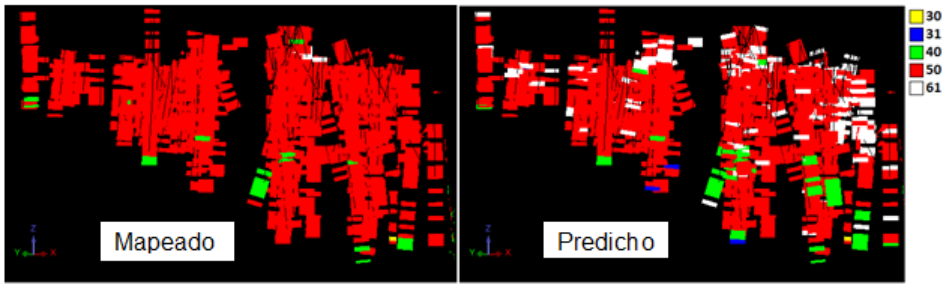


Figura 7.11: Vista de sección 2 modelo mapeado y predicho

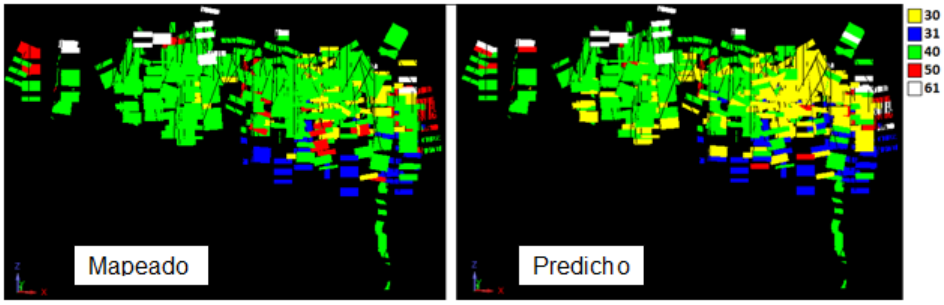


Figura 7.12: Vista de sección 3 modelo mapeado y predicho

sericita y argílica supérgena con las alteraciones potásicas (30 – 31) y clorita-sericita (40). En ella se observa la ubicación del nodo de alteración 31 correspondiente a la zona de alteración mapeada y la alteración argílica supérgena más en superficie.

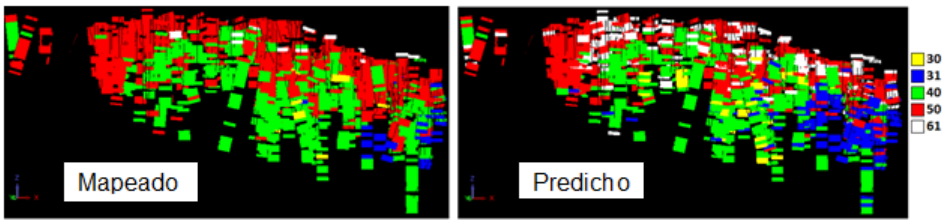


Figura 7.13: Vista de sección 4 modelo mapeado y predicho

Capítulo 8

Conclusiones

A lo largo del estudio se logró definir una metodología de clasificación de alteraciones analizando sus semejanzas, capaz de clasificarlas según sus atributos, construyendo un modelo de discriminación en base a la información facilitada por Minera Escondida. Herramientas como el clustering de categorías a partir del traslape de familias, desarrollados en la investigación, sirvieron como base para la construcción de un modelo, logrando simplificar el análisis y facilitando su comparación (variables categóricas).

La metodología presentada, trata de crear un modelo de manera semiautomática e interactiva, en donde las decisiones que se toman sean apoyadas por un set de herramientas geológicas que tienen la intención de guiar el trabajo bajo referencias entregadas por herramientas estadísticas, dejando grados de libertad para el usuario, lo cual es muy importante, ya que el modelador tiene instancias en las cuales puede aplicar sus conocimientos sobre el yacimiento en el proceso, teniendo que verificar si las asociaciones halladas tienen un significado, siendo posible su modificación. La metodología trata de describir una manera de trabajo guiada y simple para el usuario, resumida en 6 pasos:

1. Selección de variables y estudio de Co-linealidad de variables.
2. Aplicación de traslape de poblaciones para estudiar semejanzas de categorías y variables.
3. Aplicar Clustering de categorías y estudiar el dendrograma asociado a este.
4. Definir el modelo y separabilidad de las categorías.
5. Definir algoritmo para realizar el modelo y su selección de variables.
6. Estudiar el modelo y sus resultados.

Es importante recalcar que la creación de los modelos de clasificación se realiza sobre datos mapeados, los que tienen un error intrínseco asociado a ellos (debido a la subjetividad

del geólogo, la superposición de alteración, aparición de nuevas unidades geológicas entre otras) por esta razón la etapa de modelamiento necesita de flexibilidad para su creación.

El traslape de poblaciones resultó ser una buena herramienta para estudiar la semejanza de las alteraciones. Como se vio en el caso de las alteraciones de tipo 50s y 40s, para las cuales en un inicio se concluyó que eran de difícil separación, lo que fue confirmado en el estudio.

En la metodología, al realizar la comparación de resultados, cambiando el orden de clasificación en el proyecto, separando primero la alteración 30 antes que la 31 (ver anexos C), se observó un aumento del error promedio final bajando el desempeño del modelo. Se puede inferir que el modelo de clasificación utilizando el clúster de categorías es útil, y es una herramienta eficaz ya que al hacer clasificaciones basadas en un árbol de decisión, se trata que los niveles superiores tengan el mínimo error para así aumentar la cantidad de clasificación correcta (un error elevado en las primeras etapas significaría arrastrarlo a los niveles inferiores).

Crear modelos con cuatro distintas técnicas de clasificación permitió realizar una comparación entre ellas, ya que a priori es imposible saber cuál de ellas presentará un mejor desempeño. La técnica de redes neuronales resultó ser la mejor en cuanto al error de clasificación. Técnicas probadas como la regresión logística (la segunda mejor) son también de bastante utilidad debido principalmente a sus bajos tiempos de ejecución (la técnica utilizando RNA es más lenta).

Utilizando las herramientas antes descritas se forman dos familias altamente diferenciables entre sí (95 %), las cuales están constituidas por las alteraciones más tardías argílica supérgena y cuarzo - sericita contra las alteraciones más tempranas como lo son las potásicas y clorita-sericita. Finalmente, en total se lograron diferenciar cinco familias de alteraciones de las siete que se tenían. Estas constituían cerca del 99 % de la información (dejando fuera a la alteración histórica 50: blancas que ya no se mapea), logrando alrededor de 77 % en cada una de ellas; cabe destacar la gran semejanza que existe entre las alteraciones sericita - cuarzo (50s) y la alteración argílica supérgena (61) geoquímicamente, siendo estas dos alteraciones distintas, ya que seguramente se disuelven micas y arcillas presentes en ambas. Esto puede deberse también a la poca cantidad de datos que se tiene de esta última alteración (1 % de los datos corresponden a alteración 61).

Es importante recalcar cómo los modelos tienen dificultades con las categorías que se diferencian según las arcillas según lo que se ha podido estudiar, complicando la separación como en el caso de las clasificaciones 50-61, 51-52, 40-41.

Los elementos más importantes encontrados en la realización del proyecto son: el

magnesio, el aluminio, y el índice entre estos $\frac{Mg}{Al}$, el escandio, y el rubidio, los cuales son elementos que se encuentran seguramente ligados a las micas y arcillas solubles en agua regia que constituyen cada tipo de alteración, siendo de gran importancia a la hora de discriminar; ya que según alteración aparecen mayor o menor cantidad de minerales como la biotita, clorita y sericita. De esta manera se puede concluir que es la base de datos de geoquímica en donde se encuentran las variables más importantes, dejando en segundo lugar a la de leyes: CuT, CuS y Fe.

En el cálculo de errores, la idea inicial de tomar un error cuadrático general del modelo no es lo más recomendable en situaciones en las que existe una gran diferencia en la cantidad de datos que contiene cada categoría, debido a que si un 99 % de los datos perteneciere a una clase y un 1 % de ellos a otros, al clasificar todo como la primera se tendría un error de un 1 %, siendo que la segunda clase está totalmente mal clasificada, por lo cual la definición del error utilizado busca maximizar la clasificación por alteración y no replicar los volúmenes de datos asociado a cada uno.

La utilización de algoritmos de búsqueda de las principales variables, juega un gran rol dentro de un sistema complejo en donde se toma una gran cantidad de variables, no todas de importancia. Se debe sumar la dificultad de ver en tres o más dimensiones las variables desplegadas. Búsquedas como la utilizada en esta investigación resultan de gran ayuda debido a su simpleza y rapidez en la exploración de la base de datos hallando tuplas de elementos que ayuden a la discriminación por categorías. El limitar el porcentaje de cambio o el número de variables máximas que se incorporan en el sistema resulta muy útil, ya que se vio que luego de cierto punto los modelos al incorporar una mayor cantidad de variables se van complejizando aportando en muchos casos ruido y sobre ajustando el modelo, entregando peores resultados en validación al perder precisión y desempeño. Se estudiaron modelos con criterios de detención distintos como «parar cuando no mejore el modelo», donde cambios porcentuales muy pequeños incorporaban variables que no eran muy útiles.

De esta manera se concluye que el proceso que conlleva el descubrir conocimiento en las bases de información a través de la minería de datos, es de gran ayuda y valor, debido a las grandes cantidades de datos que no están siendo procesados ni analizados, conteniendo información útil, la que está esperando a ser encontrada, para poder beneficiar la construcción de los modelos geológicos y los siguientes procesos, cumpliendo así los objetivos del trabajo.

8.1. Recomendaciones

Se recomienda realizar estudios de disolución de minerales en agua regia con el fin de tratar de inferir la proveniencia de los elementos de la base de datos.

El trabajo se vio situado en una cierta zona de la mina, sería muy interesante replicarlo en su totalidad, continuando con la metodología y aumentando la cantidad de datos de alteraciones en las cuales se tenían muy pocos, para tener una mayor representatividad de estos.

Herramientas como el LabSpec que están siendo utilizadas hoy en terreno en Minera Escondida, pueden ser de gran ayuda a la hora de aportar información semi-cuantitativa, al estudiar concentraciones de la moscovita, caolinita, alunita, clorita, albita y biotita que son minerales típicos de diferentes alteraciones. Aplicando a esta información técnicas de análisis de datos, estas pueden ser un nuevo referente en la clasificación, ya que según el estudio las alteraciones que contienen arcillas son las de más difícil separación (son muy similares de su par sin arcillas).

Se sugiere estudiar a fondo la similitud de las alteraciones 50s con las alteraciones 61 que según lo analizado son muy similares entre ellas, siendo ambas conformadas por distintos tipos de minerales, que no se ven reflejados en los análisis.

Testear el modelo con más información y nuevos sondajes permitirá estudiar su desempeño frente a nuevas zonas de la mina (hay que recordar que solo se trabajó con los datos de una parte de la mina).

Se podría también trabajar con datos que contengan la información de superposición de alteraciones (al momento de mapear es posible ingresar dos alteraciones una como dominante y otra como subordinada), ya que los cambios de alteración son graduales y no repentinos, por lo que puede existir más de una alteración en una misma muestra. Esto se debería estudiar para la realización y análisis de los modelos. Existen lugares en donde se da la tendencia de tener ciertas alteraciones ubicadas dentro del yacimiento (recordar imágenes del capítulo anterior). Esto podría ocurrir debido a superposiciones en las alteraciones donde en un tramo convivan más de una, siendo información que no fue tomada en cuenta, definiendo límites duros. El incluir la posición de las muestras podría ayudar en su análisis, lo que se traduciría en la idea de unir el estudio con el modelo geológico.

Finalmente, se recomienda añadir una fase de validación con los sondajes reales, remapeando los tramos en donde se presentaron problemas con el modelo, con el objetivo de encontrar las razones de posibles equivocaciones del modelo y poder solucionarlas.

Posibles cambios del modelo

A continuación se hablará de las posibles modificaciones que podrían llevarse a cabo en la metodología descrita, a modo de estudiar cambios en el desempeño de un nuevo modelo.

El traslape de categorías se podría realizar modelando las funciones de distribución de las distintas variables según sea su categoría, calculando el área bajo la curva que se encuentre contenida entre la intersección de estas dos líneas, entregando un índice de similitud. Esta idea se ve representada en la figura 8.1.

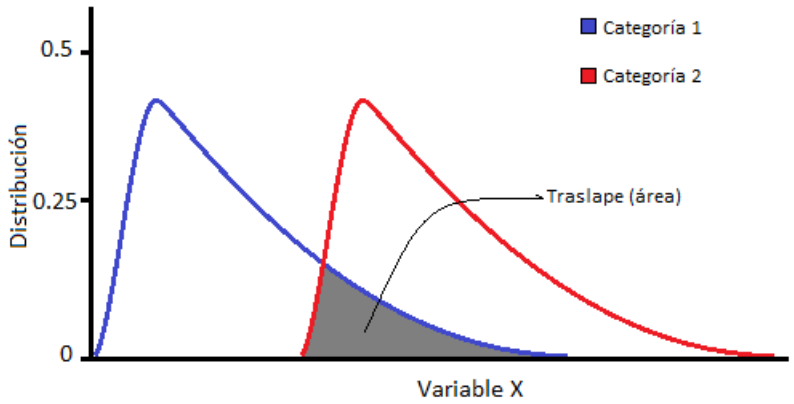


Figura 8.1: Ejemplo similitud de categorías

La clasificación también se podría generar en base a respuesta de varios clasificadores que identifique si la alteración pertenece a una alteración única, de tal manera que se tendría una respuesta vectorial del estilo $[0, 0, 1, 1]$, por ejemplo si existiesen 4 categorías. Con esta información se podrían ir creando modelos que tomen estas combinaciones y permitan discriminar de esta manera.

Sería interesante hacer los modelos de clasificación utilizando la información ya clasificada en etapas superiores, es decir, al crear el discriminador en el nivel 2 del modelo, este se realice en base a la información que fue clasificada correctamente en el nivel 1 (ver figura 8.2), ya que sería posible que desaparezcan y/o aparezcan nuevos patrones.

Es posible en el modelo incorporar una fase final de clasificación, más allá del nivel final del modelo (ver figura 8.2), en donde se incorporen clasificadores que estudien discriminar 1 sola categoría versus todas las demás, de esta manera seria mucho más fácil asignarle grados de certeza a la medición.

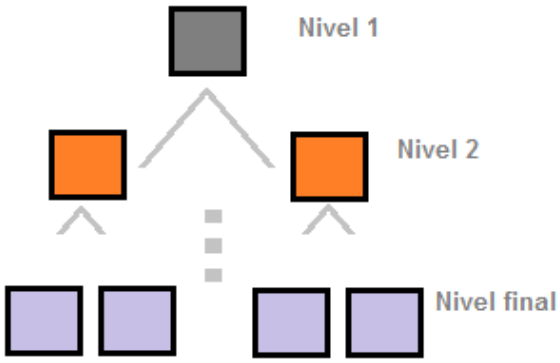


Figura 8.2: Niveles de Clasificación

8.2. Nuevos usos y utilizaciones de la metodología

La metodología permite discriminar distintos tipos de variables geológicas categóricas, como lo son la litología y/o el minzone de la roca. Además es posible incorporar al análisis otras variables categóricas utilizadas en minería como por ejemplo, línea de procesamiento, tipos de maquinaria, etc. Al ser una metodología general es aplicable a distintos escenarios.

Al utilizar las redes neuronales en la metodología, se incorpora una gran cualidad en lo que respecta a la minería de datos, que es la entrega de una respuesta de salida continua, a diferencia de las otras técnicas de análisis que entregan una categórica. En el estudio se le incorporó un umbral de corte (límite duro) para discriminar entre 2 categorías, eliminando este paso se podrían crear modelos utilizando RNA de variables continuas en minería que también dependan de muchas variables medidas, como por ejemplo la recuperación y la dureza, tratando de buscar las variables más importantes y generar modelos predictivos a partir de la información recolectada, que en este caso sería la disolución en agua regia. Además se podría intercambiar la función de redes neuronales por una regresión multilineal y comparar el resultado de ambos.

De esta manera, para finalizar, la minería de datos representa un área del conocimiento que tiene el potencial de entregar un mayor valor al proceso minero. En base a herramientas estadísticas es posible entender la alteración que presenta una roca en función de la probabilidad de ocurrencia de estas, siendo de gran ayuda al existir en el yacimiento zonas de superposición de alteraciones que son difíciles de mapear existiendo variabilidad de mezclas en ellas.

Bibliografía

[1] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery: an overview,” in *Advances in knowledge discovery and data mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996, pp. 1–34. [Online]. Available: http://portal.acm.org/citation.cfm?id=257942

[2] S. Carmona, “Análisis exploratorio de relaciones geometalúrgicas en sulfuros de Radomiro Tomic,” Master’s thesis, Universidad de Chile, Beaucheff 850, Santiago, Chile, jul 2009.

[3] R. Sillitoe, “Porphyry copper systems,” *Society of Economic Geologists, Inc*, vol. 105, pp. 3–41, 2010.

[4] J. D. Winter, “Chapter 9: Trace elements,” may 2014. [Online]. Available: http://www.whitman.edu/geology/winter/Petrology/Ch%2009%20Trace%20Elements%20and%20Isotopes.ppt"

[5] V. Maksaev, “Procesos supergenos,” jan 2014. [Online]. Available: http://www.cec.uchile.cl/~vmaksaev/PROCESOS%20SUPERGENOS.pdf

[6] S. Bulatovic, D. Wyslouzil, and C. Kant, “Effect of clay slimes on copper, molybdenum flotation from porphyry ores,” *Copper 99-Cobre 99 International Enviroment Conference*, vol. 2, no. 1-3, pp. 95– 112, 1999.

[7] D. Langmuir, *Aqueous Environmental Geochemistry*. New jersey: Prentice-Hall, Inc., 1997.

[8] BHP-Billiton, “Marco geologico Minera Escondida,” 2009, Informe interno.

[9] W. K. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, 2nd ed. Springer, Sep. 2007. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/3540722432

[10] A. J. Izenman, *Modern Multivariate Statistical Techniques*. Philadelphia, Pa 19122: Springer, 2008.

[11] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[12] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. New York, NY 10013: Springer, 2008.

[13] C.-L. Chen, D. B. Kaber, and P. G. Dempsey, "A new approach to applying feedforward neural networks to the prediction of musculoskeletal disorder risk," *Applied Ergonomics*, vol. 31, no. 3, pp. 269 – 282, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003687099000551

[14] I. V. Tetko, D. J. Livingstone, and A. I. Luik, "Neural network studies, 1. comparison of overfitting and overtraining." *Journal of Chemical Information and Computer Sciences*, vol. 35, no. 5, pp. 826–833, 1995. [Online]. Available: http://dblp.uni-trier.de/db/journals/jcisd/jcisd35.html#TetkoLL95

[15] J. V. Tu, "Advantages and disadvantages of using artifical neural networks versus logistic regression for rediction for medical outcomes," *Journal of Clinical Epidemiology*, vol. 49, no. 11, pp. 1225–1231, 1996.

[16] C. M. Cuadras, *Nuevos Métodos de Análisis Multivariante*. Barcelona, España: CMC Editions, 2012.

[17] M. Mures, A. García, and M. Vallejo, "Aplicación del análisis discriminante y regresión logística en el estudio de la morosidad en las entidades financieras : comparación de resultados," *Pecvnia : Revista de la Facultad de Ciencias Económicas y Empresariales, Universidad de León*, vol. 0, no. 1, 2005. [Online]. Available: http://revpubli.unileon.es/ojs/index.php/Pecvnia/article/view/746

[18] C.-L. Chen, D. B. Kaber, and P. G. Dempsey, "Using feedforward neural networks and forward selection of input variables for an ergonomics data classification problem," *Hum. Factor. Ergon. Manuf.*, vol. 14, no. 1, pp. 31–49, Jan. 2004. [Online]. Available: http://dx.doi.org/10.1002/hfm.v14:1

[19] D. Broadhurst, R. Goodacre, A. Jones, J. J. Rowland, and D. B. Kell, "Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry," *Analytica Chimica Acta*, vol. 348, no. 1-3, pp. 71–86, aug 1997. [Online]. Available: http://dx.doi.org/10.1016/s0003-2670(97)00065-2

[20] I. V. Tetko, “Neural network studies, 4. introduction to associative neural networks.” *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 3, pp. 717–728, 2002. [Online]. Available: http://dblp.uni-trier.de/db/journals/jcisd/jcisd42.html#Tetko02

[21] T. Fawcett, “Roc graphs: Notes and practical considerations for researchers,” Intelligent Enterprise Technologies Laboratory, Tech. Rep., 2004.

[22] G. N. Lee and H. Fujita, “K-means clustering for classifying unlabelled mri data.” in *DICTA*. IEEE Computer Society, 2007, pp. 92–98. [Online]. Available: http://dblp.uni-trier.de/db/conf/dicta/dicta2007.html#LeeF07

[23] J. Boisvert, M. E. Rossi, and C. V. Deutsch, “Prediction, hierarchical multivariate regression for mineral recovery and performance,” *CCG annual Report*, vol. 11, p. 302, 2009.

[24] J. P. Bernal and L. B. Railsback, “Introducción a la tabla periódica de los elementos y sus iones para ciencias de la tierra,” *Revista Mexicana de Ciencias Geológicas*, vol. 25, no. 2, pp. 238–246, 2008.

[25] B. Townley, M. Muñoz, and R. Luca, “Modelamiento geoquímico distrito escondida: Discriminación de ambientes hidrotermales,” Geo AV, Santiago, Chile, Informe privado, 2011.

[26] T. Masters, *Practical Neural Network Recipes in C++*. New York: Academic Press, 1993.

[27] H. N. Koivo, “Neuronal networks: Basics using matlab. manual neuronal network toolbox,” Finland: Alto University, Tech. Rep., 2008.

[28] G. Cardillo, “Mathworks: Roc curve: compute a receiver operating characteristics curve,” may 2008. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/19950

Apéndice A

Anexos: Programas realizados

En la secion «A» de los anexos expondrán y explicaran los códigos hechos para la realización de la tesis presentando los distintos programas a través de seudo-códigos.

A.1. Fusión de bases de datos

Algoritmo 1: Fusión base de datos Continua

Input: 2 bases de datos de sondajes, con la segunda compuesta por variables continuas
Output: Base de datos unida

```
1  BD1 ← valores de database 1
2  BD2 ← valores de database 2
3  Hole_Name1 ← holeID de database1
4  Hole_Name2 ← holeID de database2
5  for i ← 0 to Length : Hole_Name1 do
6      Var ← Matriz variables de que están presentes
7      Por ← el porcentaje que está presente
8      for ii ← 0 to Length : Hole_Name1 do
9
10         if HoleID(i) = HoleID(ii) then
               /* 4 posibles condiciones para estar contenido          */
11             if From1>From2 & To1>To2 and From1<To2: then
                   /* Caso en el cual está contenido la parte inferior */
12                 Var ← almacena variable
13                 Por ← almacena porcentaje
14             if From1<From2 and To1<To2 and From2 <=To1 then
                   /* Caso en el cual está contenido la parte superior */
15                 Var ← almacena variable
16                 Por ← almacena porcentaje
17             if (From1 >= From2 and To2 >= To1 and (From1 ! = To2) and (From2 ! =
               To1)): then
                   /* Caso en el cual está contenido completamente y no excede
                      al otro sondajee                                  */
18                 Var ← almacena variable
19                 Por ← almacena porcentaje
20             if (From1 < From2 and To1 >= To2) or (From1 == From2 and To1 > To2) or
               (From1 < From2 and To1 == To2) then
                   /* Caso en el cual está contenido completamente y excede al
                      otro sondajee                                     */
21                 Var ← almacena variable
22                 Por ← almacena porcentaje
23
24     Variables_prom ← Var*Por
25     Porc_sondaje ← sum(Por)
26     Escribir línea Base de datos 1+ Variables_prom + Porc_sondaje
```

En la figura A.1 se aprecia una imagen de cómo se ve la aplicación.

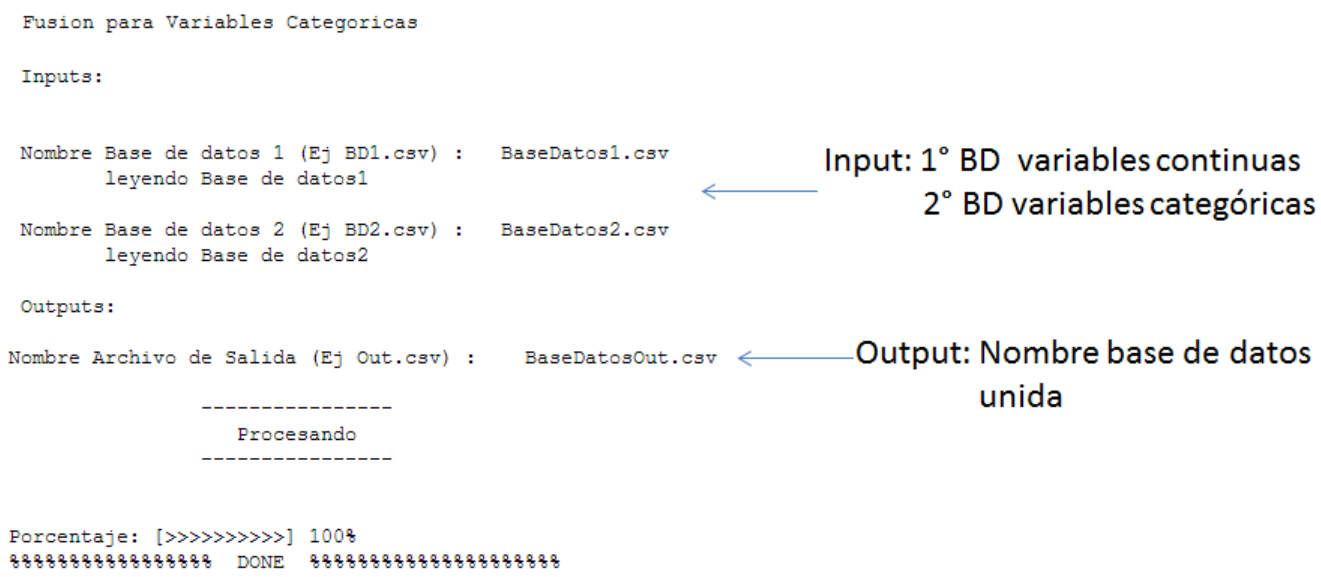
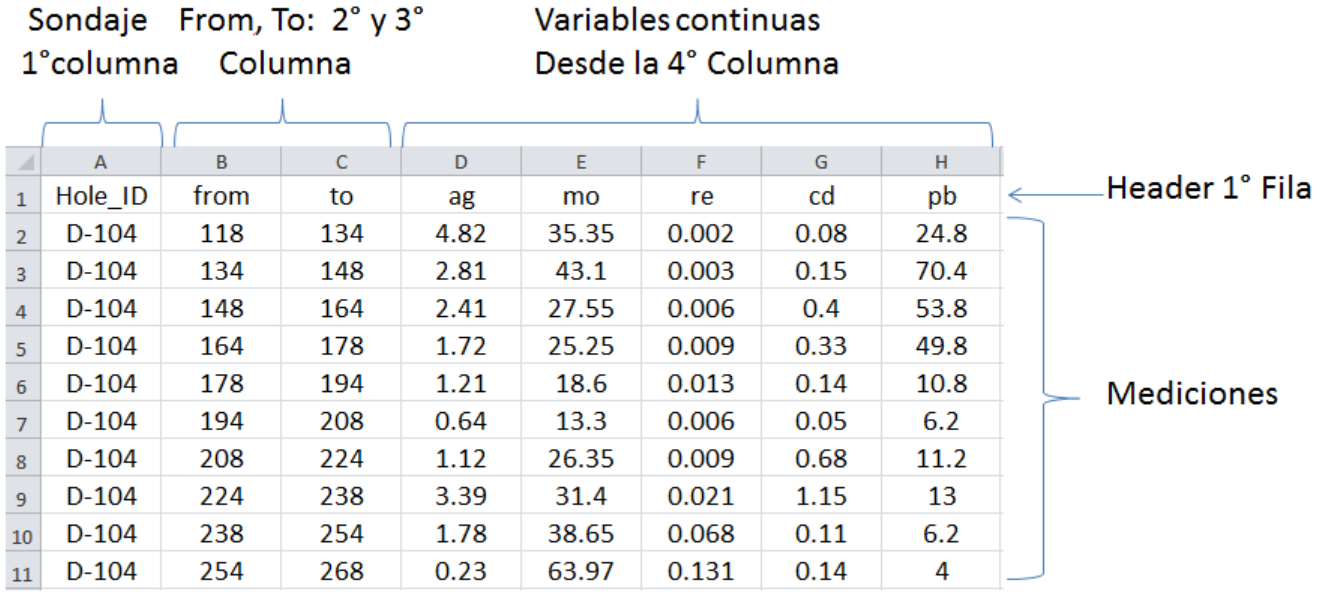


Figura A.1: Fusión continua

En la figura A.2 se aprecia el formato que debe tener el csv que se ingresa en la aplicación.



	A	B	C	D	E	F	G	H
1	Hole_ID	from	to	ag	mo	re	cd	pb
2	D-104	118	134	4.82	35.35	0.002	0.08	24.8
3	D-104	134	148	2.81	43.1	0.003	0.15	70.4
4	D-104	148	164	2.41	27.55	0.006	0.4	53.8
5	D-104	164	178	1.72	25.25	0.009	0.33	49.8
6	D-104	178	194	1.21	18.6	0.013	0.14	10.8
7	D-104	194	208	0.64	13.3	0.006	0.05	6.2
8	D-104	208	224	1.12	26.35	0.009	0.68	11.2
9	D-104	224	238	3.39	31.4	0.021	1.15	13
10	D-104	238	254	1.78	38.65	0.068	0.11	6.2
11	D-104	254	268	0.23	63.97	0.131	0.14	4

Figura A.2: Formato Base de datos continua

Algoritmo 2: Fusión base de datos Continua

Input: 2 bases de datos de sondajes, con la segunda compuesta por variables categóricas
Output: Base de datos unida

```
1  BD1 ← valores de database 1
2  BD2 ← valores de database 2
3  Hole_Name1 ← holeID de database1
4  Hole_Name2 ← holeID de database2
5  for i ← 0 to Length : Hole_Name1 do
6      Var ← Matriz variables de que están presentes
7      Por ← el porcentaje que está presente
8      for ii ← 0 to Length : Hole_Name1 do
9
10         if HoleID(i) = HoleID(ii) then
               /* 4 posibles condiciones para estar contenido                */
11             if From1>From2 & To1>To2 and From1<To2: then
                   /* Caso en el cual está contenido la parte inferior       */
12                 Var ← almacena variable
13                 Por ← almacena porcentaje
14             if From1<From2 and To1<To2 and From2 <=To1 then
                   /* Caso en el cual está contenido la parte superior       */
15                 Var ← almacena variable
16                 Por ← almacena porcentaje
17             if (From1 >= From2 and To2 >= To1 and (From1 ! = To2) and (From2 ! =
               To1)): then
                   /* Caso en el cual está contenido completamente y no excede
                      al otro sondajee                                        */
18                 Var ← almacena variable
19                 Por ← almacena porcentaje
20             if (From1 < From2 and To1 >= To2) or (From1 == From2 and To1 > To2) or
               (From1 < From2 and To1 == To2) then
                   /* Caso en el cual está contenido completamente y excede al
                      otro sondajee                                           */
21                 Var ← almacena variable
22                 Por ← almacena porcentaje
23
       /* ordenar de mayor a menor los valores de las muestras según su
          ocurrencia                                                          */
24     Porc_ordenado ← sort(Por)
25     Indices_orden(i) ← Por.index of (Porc_ordenado(i)
       /* caso limite cuando se encuentran a la mitad 50% y 50%              */
26     if (pororden[-1]==0.5) and (pororden[-2]==0.5) then
27         Escribir linea Base de datos 1+Var(pororden[-1]), Var(pororden[-2]), porcentajes
           (50 %)
28     Escribir linea Base de datos 1+Var(Indices_orden[-1]), Var(Indices_orden[-2]),
       pororden[-1],pororden[-2];
```

En la figura A.3 se aprecia una imagen de como se ve la aplicación.

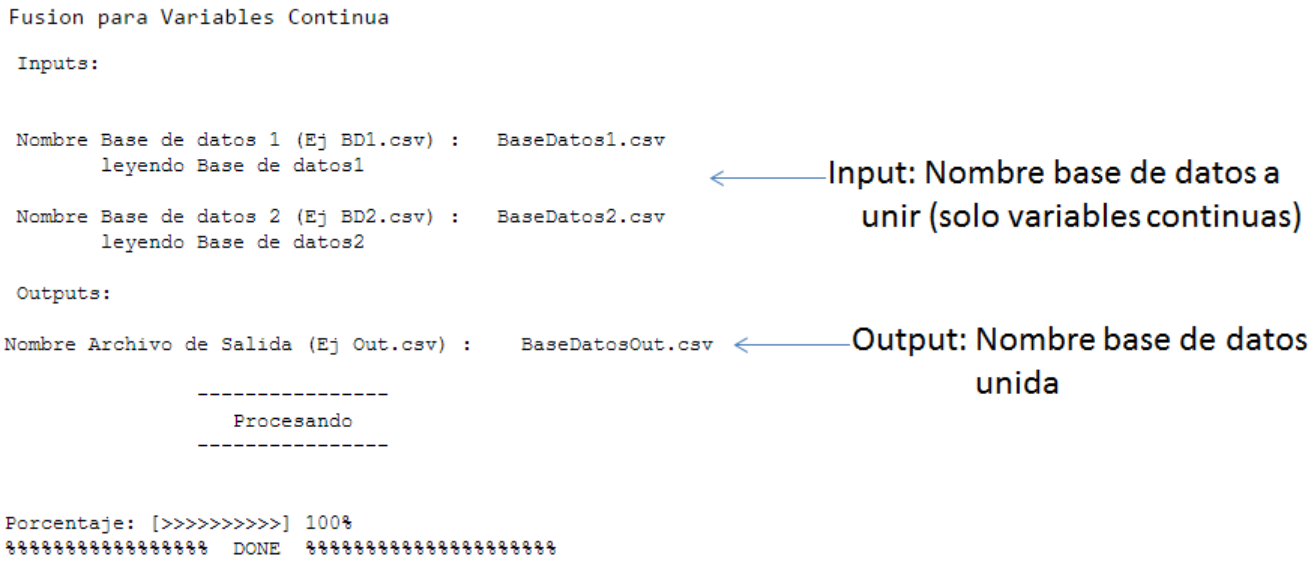
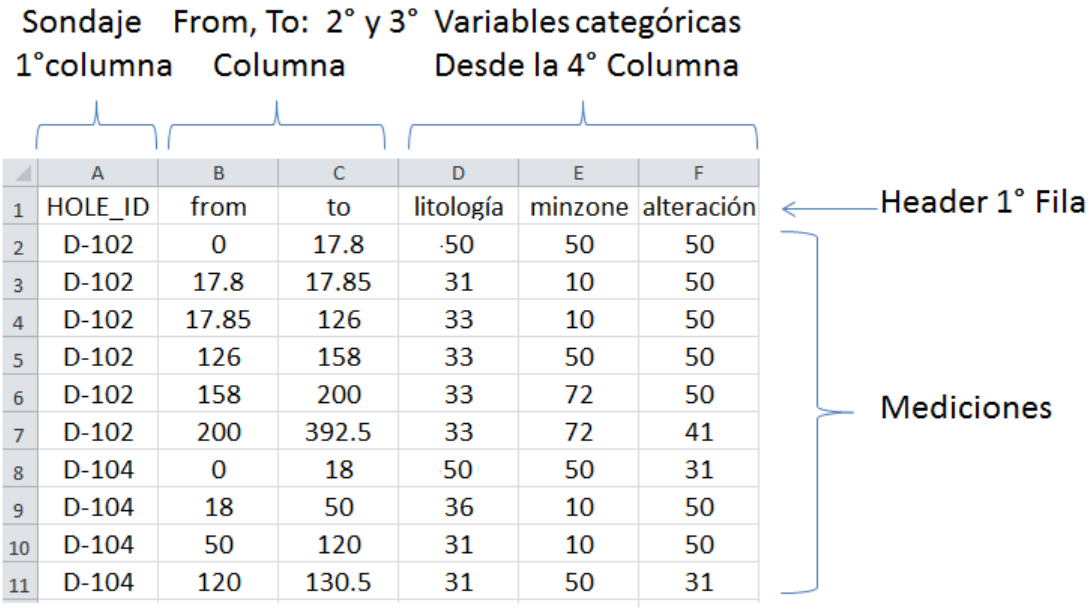


Figura A.3: Fusión categórica

En la figura A.4 se aprecia el formato que debe tener el csv que se ingresa en la aplicación.



	A	B	C	D	E	F
1	HOLE_ID	from	to	litología	minzone	alteración
2	D-102	0	17.8	50	50	50
3	D-102	17.8	17.85	31	10	50
4	D-102	17.85	126	33	10	50
5	D-102	126	158	33	50	50
6	D-102	158	200	33	72	50
7	D-102	200	392.5	33	72	41
8	D-104	0	18	50	50	31
9	D-104	18	50	36	10	50
10	D-104	50	120	31	10	50
11	D-104	120	130.5	31	50	31

Figura A.4: Formato Base de datos categórica

A.2. Metodología Forward modificada

Algoritmo 3: Metodología Forward Parte 1

Input: Bases datos (Entrenamiento, validación y ajuste). Cantidad de categorías a separar. Categorías denotadas como 1. Saber si normaliza

Output: Modelo, curva roc, errores asociados

```
/* Se normaliza con respecto a la de ajuste ya que al ser datos sacados
   al azar consevan la media y la varianza del yacimiento            */
```

1 **if** *Normaliza=1* **then**

```
    /* Cacular media y varianza de BD. ajuste y aplicar estos valores
       para normalizar las demás bases de datos.                     */
```

2 **for** *i←1 to cantidad de variables* **do**

3 BD_Entrena[i,:]={BD_Entrena[i,:]-Var(BD_Ajuste[i,:])}//mean(BD_Ajuste[i,:])

4 Pasar datos a formato de RNA en Matlab (traspuesto)

```
/* Definir los error para almacenar los de la iteración actual con la
   anterior y lograr compararlos, el error 2 debe ser menor que el 1  */
```

5 error1 ← 1

6 error2 ← 0.9

7 X ← cantidad de variables

8 N ← 1 (cantidad de iteraciones) elementos ← [] (variables que van a ser utilizadas)

9 i ← 0 (índice de la posición de la variable)

En la figura A.5 se aprecia una imagen de como se ve la aplicación.

```
------------------------------------------------------------------
  ____________________________________________________________
  |                                                           |
  |      -Metodologia Mod Forward : Redes Neuronales-         |
  |___________________________________________________________|

Inpunts del sistema:

Nombre base de datos de Entrenamiento:  Ent_5061.xlsx

Nombre base de datos de Ajuste: Tunning3.csv

Desea normalizar con respecto a tunning?  1 (si) , 0 (no): 0

Codigos de Alteraciones Presentes :
    30    31    40    50    61

Cantidad de alteraciones como 1:  2
          > COD de Alteracion:  50
          > COD de Alteracion:  61

Buscando Variables .......
```

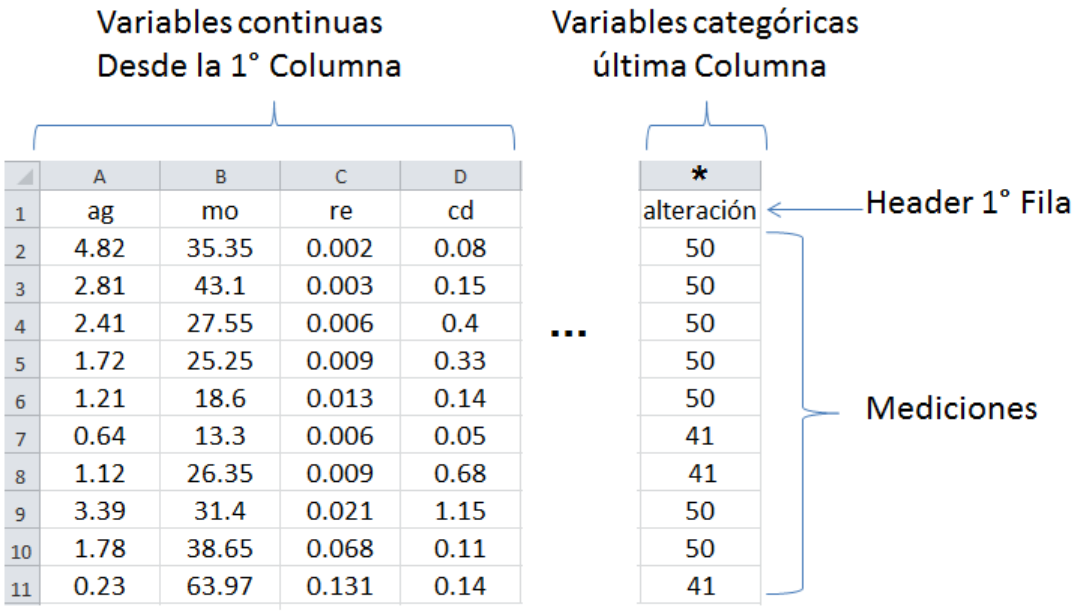
Input: Nombre base de datos Entrenamiento y Ajuste.

Opción de normalizar.

Input: Numero de Alteraciones que serán catalogadas como 1 y cuales son.

Figura A.5: input algoritmo

En la figura A.6 se aprecia el formato que debe tener el csv que se ingresa en la aplicación.



	A	B	C	D	*
1	ag	mo	re	cd	alteración
2	4.82	35.35	0.002	0.08	50
3	2.81	43.1	0.003	0.15	50
4	2.41	27.55	0.006	0.4	50
5	1.72	25.25	0.009	0.33	50
6	1.21	18.6	0.013	0.14	50
7	0.64	13.3	0.006	0.05	41
8	1.12	26.35	0.009	0.68	41
9	3.39	31.4	0.021	1.15	50
10	1.78	38.65	0.068	0.11	50
11	0.23	63.97	0.131	0.14	41

Figura A.6: Formato Base de datos metodología forward

Algoritmo 4: Metodología Forward Parte 2: iteración

```
   /* diferencia entre los error mayor a 0.5% para seguir:                    */
1  while (error1 − error2) > 0.005 do
       /* si se seleccionan todos los elementos:                              */
2      if N=X then
3          break
4      i=i+1;
5      for c ← 1 to length(elementos) do
           /* si ya los tiene todos                                           */
6          Break;
           /* saltar la variable si ya está en el grupo:                      */
7          if find(elementos==i)>0 then
               /* Caso en que el último elemento es el que se seleccionó      */
8              if i==x then
                   /* informa que trabajo con el último elemento              */
9                  extremo=+1;
                   /* bajar la cantidad de variables total a utilizar         */
10                 X=X-1
11             tccsi la variable fue escogida decirle que tiene 100 % de error para no ser
               seleccionada y que pase a la próxima errores(i,1)=1;
12             i=i+1
13     if i >= x then
           /* entra si ya se dio una vuelta                                   */
14         i=1 (reinicia el i);
15         vueltas=vueltas+1 (cuenta las vueltas);
16     if extremo ≠ 1 then
           /* outFuncion corresponde a la salida dependiendo del algoritmo
              que se utilice separación simple, Cluster, RNA, Regresión
              Logística (ver las funciones más adelantes)                     */
17         [error(i), out_train(i),outFuncion(i)]= Algoritmos( Variables de entrada)
           /* error(i)= outFuncion almacena basicamente los modelos que se
              generan                                                         */
18     if i>=x-1 and extremo=1 then
19         extremo=0
20         Pos_minimo_Error=find(errores=min(errores))
21         if cont_vueltas==1 then
               /* si el contador de vueltas es igual a uno significa que ya
                  recorrió todos los elementos y realizo los modelos con ellos
                  por lo cual hay que guardar el resultado del mejor          */
22             elementos(cont_elemento)=Pos_minimo_Error
23             cont_elemento=cont_elemento+1 error1=error2
24             Modelo=OutFuncion(Pos_minimo_Error)
25         else
26             error1=error2
27             error2=errores(Pos_minimo_Error);
               elementos(cont_elemento)=Pos_minimo_Error
28             cont_elemento=cont_elemento+1
29             errores(Pos_minimo_Error)=1
30             Modelo=OutFuncion(Pos_minimo_Error)
31         print [errores, Variable Elegida]
```

Algoritmo 5: Redes Neuronales

Input: entrenamiento(variables de entrenamiento),target_train (Variable Objetivo entrenamiento codificada 01),y,Alt_inicial(Variable Objetivo entrenamiento),inAlt(categorias como 1),validacion(variables de ajuste),target_val(Variable Objetivo ajuste), uc(categorias presentes)

Output: out_train(Variable salida entrenamiento codificada 01) ,out_net(red neuronal),x_val_(valor de corte), fval_e (error ajuste),neuronas(cantidad de neuronar capa oculta), error_ENT (error entrenamiento)

```
target_trainCOD=target_train
(y, x)=size(target_train)
Q_alt1=length(inAlt)
target_train=[]
for i=1 To length(target_train) do
    for ii=1:length(inAlt) do
        if target_trainCOD(i,1)==inAlt(ii) then
            target_train(1,i)=1
        else
            target_train(1,i)=0;
[Ninput x1]=tamaño(entrenamiento);
[Noutput x1]=tamaño(target_train); /* cantidad de neuronas en la capa oculta        */
neuronas= truncar((Ninput * Noutput)^0.5);
if neuronas <2 then
    neuronas=2
/* variables para ir comparando el caso anterior con el actual                */
error0=-1
error1=0
xval0=0
xval1=0
while error0 < error1 do
    for C=1 To 5 do
        /* iniciar red neuronal con neuronas (cantidad)                       */
        net_c = newff(neuronas) /* Entrenar red neuronal                      */
        net_c = train(net_c,(entrenamiento),(target_train)); /* salida entrenamiento */
        out _train_c = sim(net_c, entrenamiento); (x_val(c),fval(c)) = encontrar el óptimo valor x_val
            que sea el umbral de corte /* salida Ajuste                      */
        out_val_c = sim(net1, validacion'); /* errores de Entrenamiento y ajuste */
        fval(c)=Error_Sep_RNA(x_val,target_val ,out_val1 ,uc,inAlt);
        error_ENT_(c)=Error_Sep_RNA(x_val, target_trainCOD ,out_train1 ,uc,inAlt);
    posMIN=find(fval==min(fval));
    out_val=out_val_posMIN
    x_val_min=x_val(posMIN)
    out_train=out_train_posMIN
    net=net_posMIN
    out_val=out_val_posMIN
    fval_min=fval(posMIN)
    error_ENT=error_ENT_(posMIN)
    /* almacena los valores de la iteración y los cambia por los nuevos valore
        */
    xval0=xval1
    xval1=x_val_min
    error0=error1
    error1=fval_min
    if error0 >= error1 then
        break
    out_net=net
    neuronas=neuronas+1;
out_train = sim(out_net, entrenamiento)
(x_val_ ,fval_e) = fminbnd(@(x) Error_Sep_RNA(x,Alt_inicial ,out_train ,uc,inAlt),-1,1); out_val =
sim(out_net, validacion');
fval_e = Error_Sep_RNA(x_val_ , target_val, out_val ,uc,inAlt);
error_ENT = Error_Sep_RNA(x_val_ , target_trainCOD ,out_train ,uc,inAlt);
```

Algoritmo 6: Funcion K-mean Clustering

Input: entrenamiento(variables de entrenamiento),target_train (Variable Objetivo entrenamiento codificada 01),y,Alt_inicial(Variable Objetivo entrenamiento),inAlt(categorias como 1),validacion(variables de ajuste),target_val(Variable Objetivo ajuste), uc(categorias presentes)

Output: out_train(Variable salida entrenamiento codificada 01) ,out_net(red neuronal),x_val_(valor de corte), fval_e (error ajuste),neuronas(cantidad de neuronar capa oculta), error_ENT (error entrenamiento)

```
   /* se realiza el clustering 50 veces por si cae en un outlier y forma un
      cluster falso                                                          */
1  for i=1:50 do
2      [IDX,ctr]=kmeans(entrenamiento',2,'emptyaction','drop')
       /* Función Matlab: IDX, es la salida del cluster como son 2 son 1 , 2
          y ctr son los centroides del cluster                               */
       /* Almacenar los centroides:                                         */
3      ctr_MATRIX1(i,:)=ctr(1,:); ctr_MATRIX2(i,:)=ctr(2,:);
       /* se calcula el error cuadrático medio para saber que cluster fue
          llamado como 1 y 2                                                 */
4      result= sum((target_train'-(IDX-1)).^2)/y
       /* n es para ver si los números están cambiados                      */
5      n=0 /* si el error gral. es mayor que 0.5 significa que esta al revés
          el número del cluster                                              */
6      if result>0.5 then
7          n=1;
8      if n=1 then
9          Cambiar los 0 por 1 y 1 por 0
       /* calcular el error general medio del clustering (ver función error)
          */
10     resultados(i,1)=errorg([variables validación])
   /* buscar la posición del error mínimo                                   */
11 Min_find=find(resultados==min(resultados));
   /* archivos de salida error entrenamiento y centroides                   */
12 errorEnt=resultados(Min_find)
13 ctr_out=[ctr_MATRIX1(Min_find,:);ctr_MATRIX2(Min_find,:)]
   /* fin entrenamiento e inicio de ajuste                                  */
14 [out_val]=ClasificationKmean( ctr_out, Validacion);
   /* salida error de validación                                            */
15 error =errorg([variables]);
```

Algoritmo 7: Clasificador K-mean Clustering

Input: centroides, matriz_de_puntos
Output: class (identificación de la cual cluster pertenece el punto)

```
for i=1 To Número de puntos do
    Distance1= (sum((Centroide_1-Matriz(i,:)).^2))^.05
    Distance0=(sum((Centroide_2(2,:)-Matriz(i,:)).^2))^.05
    if Distance1>Distance0 then
        class(i,1)=0
    else
        class(i,1)=1
```

Algoritmo 8: Regresión Logística

Input: entrenamiento(variables de entrenamiento),target_train (Variable Objetivo entrenamiento codificada 01),y,Alt_inicial(Variable Objetivo entrenamiento),inAlt(categorías como 1),validacion(variables de ajuste),target_val(Variable Objetivo ajuste), uc(categorías presentes)
Output: out_train(Variable salida entrenamiento codificada 01) ,out_net(red neuronal),x_val_(valor de corte), fval_e (error ajuste),neuronas(cantidad de neuronas capa oculta), error_ENT (error entrenamiento)

```
/* Función de MATLAB «glmfit» para definir la función de la regresión
   logistica, con sus pesos b                                          */
[b,dev,stats] = glmfit(entrenamiento ,target_train,'binomial','link','logit')
/* Función de MATLAB «glmval» para evaluar el modelo con los distintos
   puntos                                                              */
out_train1 = glmval(b, entrenamiento,'logit')
tccpasar a 1 o a 0 si es mayor o menor que 0.5 for i=1:número de datos do
    out_train1(i,1)> 0.5 out_train(i,1)=1
    out_train(i,1)=0
/* calcular el error asociado a entrenamiento                          */
error_ENT=errorg( Alt_inicial, out_train , uc, inAlt);
/* iniciar análisis de datos de ajuste                                 */
M= variables que se utilizan
/* evaluar                                                             */
expXB=(exp(M*b))
out_val=expXB./(expXB+1)
for i=1:número de datos ajuste do
    , if out_val(i,1)> 0.5 then
        out_val(i,1)=1
    else
        out_val(i,1)=0;
/* calcular error de ajuste                                            */
error=errorg( M_val(:,xval), out_val , uc, inAlt)
```

Algoritmo 9: Programa: Separación Simple

Input: Base de Datos de Entrenamiento, Base de Datos de ajuste, base de datos de validación, variables

Output: curvas roc (datos y gráficos)

```
/* esto se hace para cada uno de los modelos de clasificación:[50,61 vs
   30,31,40][50 vs 61][31 vs 30,40][30 vs 40]                          */
```

1 , uc= categorías presentes para generar el modelo

2 inAlt= alteraciones que serán definidas como 1

```
/* las alteraciones que están en uc y no en inAlt son 0               */
```

3 [separadorXX error NXX valorXX entrada_binaria error_ENT]= Sep_Sim(BaseDatosEnt,BaseDatosVal, uc, inAlt)

```
/* separador(variable seleccionada para separar), error(error asosiado a
   la clasificación de ajuste), valor(para ver si es mayor que o menor
   que), valor( valor del umbral de corte), entrada_binaria(vector con
   las variables de alteración de entrada categorizadas como 1 o
   0),error_ENT(error de entrenamiento) (la XX significa inAlt)       */
```

4 Show(error en ajuste)

5 Show(error en entrenamiento)

```
/* generar curva roc y puntos de ellas (yroc xroc)                    */
```

6 [yroc xroc]=roc(inAlt,[variable de separación, alteración codificada])

7 Guardar(yroc xroc como Roc$_{C}urve_ssXX.xls$)

```
/* se utilizan los modelos y se clasifica como un árbol (si mayor que
   este entonces 1 si no 0 y así sucesivamente)                       */
```

8 Guardar(clasificación de los puntos de validación en salidaSS.xls) Guardar(matriz de confusión en matriz_confusion_SS.xls)

Algoritmo 10: Función: Separacion Simple(Sep_sim)

Input: BaseDatosEnt,BaseDatosVal,uc, inAlt

Output: separador(variable seleccionada para separar), error(error asociado a la clasificación de ajuste), signo (para ver si es mayor que o menor que), valor(valor del umbral de corte), entrada_binaria(vector con las variables de alteración de entrada categorizadas como 1 o 0),error_ENT(error de entrenamiento)

```
1  for i=1 To número de variables do
       /* Existen dos casos uno donde las variables son catalogadas como 1 y
          las demás como 0 y el caso en que sucede al revés, de no
          especificar esto bien la selección comete errores por lo cual se
          hacen los dos casos y luego se escoje                              */
2      [x_val, fval] = Min(Error de separación[mayor que] )
       /* Valores que cumple minimizan el error minimizando la función error
          de separación, caso mayor que. x_val es valor que minimiza la
          función (el umbral) y fval es el error minimizado                  */
       /* almacenar los valores de en cada iteración                         */
3      separador_0(i,1)=x_val
4      Error_sep_0_ENT(i,1)=fval
       /* Valores que cumple minimizan el error minimizando la función error
          de separación, caso menor que.                                     */
5      [x_val, fval] = Min(Error de separación[menor que] )
       /* almacenar los valores de en cada iteración                         */
6      separador_1(i,1)=x_val
7      Error_sep_1_ENT(i,1)=fval
   /* encontrar la posición del mínimo de los vectores                       */
8  min_0=find(Error_sep_0==min(Error_sep_0));
   min_1=find(Error_sep_1==min(Error_sep_1));
9  if Error_sep_0(min_0)>Error_sep_1(min_1) then
10     signo=0
11     separador=min_1
12     valor=separador_1(min_1)
13     error=Error_sep_1(min_1)
14     error_ENT=Error_sep_1_ENT(min_1)
15 else
16     signo=1; separador=min_0
17     valor=separador_0(min_0)
18     error=Error_sep_0(min_0)
19     error_ENT=Error_sep_0_ENT(min_0)
```

Algoritmo 11: Función Error Medio

Input: Target_val, out_val , uc, inAlt
Output: error, errores

```
Q_alt=length(uc)
Q_alt1=length(inAlt)
(y x)=size(Target_val)
Target_01 =zeros(y,1)
for i=1:y do
    for ii=1:Q_alt1 do
        if Target_val(i,1)==inAlt(ii) then
            Target_01(i,1)=1
errores=zeros(Q_alt,1);
for i=1:Q_alt do
    L=find(Target_val==uc(i))
    errores(i,1)=sum(((Target_01(L) − (out_val(L))).^2))/length(L)
Error1=0;
for i=1:Q_alt do
    Error1=Error1+errores(i,1);
error=(Error1)/Q_alt
```

Algoritmo 12: Traslape de poblaciones

Input: Base de datos, numero de variables a tomar en cuenta, porcentaje de investigacion.
Output: Traslape por categoría, variables utilizadas, matriz de traslape, Dendrogramas : single, average y complete.

```
/* Separando por familias segun elementos                                    */
for i ← 1 To Cantidad de Variables do
    for ii ← 1Cantidad de alteraciones do
        ElexCat_i_ii ← buscar indice y crear variable
/* crear capeta para guardar resultados                                      */
Crear Carpeta traslape_por_familia_porc_X % Matriz_traslape_final ← matriz para almacenar valores finales de la matriz de traslape.
/* Loop para calcular los cuantiles                                          */
for alteracion ← 1 To Cantidad de alteraciones do
    for i ← 1 to Cantidad variables do
        Quan_AltEst(i,1)= cuantil Superior varible i de la alteracion
        Quan_AltEst(i,2)=cuantil inferior varible i de la alteracion
    /* Loop para tomar cada cuantil                                          */
    for ii ← 1 To Cantidad de variables do
        /* Loop p compararlos entre si los cuantiles                         */
        for i← 1 To Cantidad de alteraciones-1) do
            /* se evalúan los casos igual que si fueran sondaje para ver porcentaje de traslape */
            EleMax ← cuantil superior del elemento ii de la categoria i
            EleMin ← cuantil inferior del elemento ii de la categoria i
            M_porc gets Matriz para almacenar los valores de los traslapes de familias
            /* caso que no se toquen los intervalos                          */
            if (EleMax < Quan_AltEst(ii-1,2) ) or ( EleMin >= Quan_AltEst(ii-1,1) then
                M_porc(ii,i) ← 0
            /* caso en el cual contiene al otro intervalo                    */
            else if EleMax ≥ Quan_AltEst(ii-1,1)  EleMin ≤ Quan_AltEst(ii-1,2) then
                M_porc(ii,i)← 1
            /* caso donde está contenido                                     */
            else if EleMax ≤ Quan_AltEst(ii-1,1)  EleMin ≥ Quan_AltEst(ii-1,2) then
                M_porc(ii,i) ← 1
            /* caso estudio está arriba el intervalo                         */
            else if Quan_AltEst(ii-1,1) > EleMax then
                M_porc(ii,i)=(EleMax-Quan_AltEst(ii-1,2))/(Quan_AltEst(ii-1,1)-Quan_AltEst(ii-1,2))
            /* caso estudio está más abajo el intervalo                      */
            else if Quan_AltEst(ii-1,1) < EleMax then
                M_porc(ii,i)=(Quan_AltEst(ii-1,1)-EleMin)/(Quan_AltEst(ii-1,1)-Quan_AltEst(ii-1,2))
    /* Grabar los valores de M_porc en la carpeta «carpeta_traslape» con el nombre traslape_por_familia_a(alteracionX)_%  */
    save(M_porc)
    /* in variable auxiliar para establecer que paso por la diagonal al escribir la matriz de traslape */
    in=0 /* Loop para comparar una familia con las otras                     */
    for j ←1 to Cantidad alteraciones-1 do
        /* menos uno porque hay una que se está comparando                   */
        Ordenar de mayor a mayor los traslapes de almacenados en M_port, por culumna tras=0 for i ← 1 to n (Numero seleccionado de variables de mayor diferencia) do
            tras=tras+[Sort(M_porc(:,j)](i,1))
        tras=tras/n
        /* la diagonal se completa con 1(If j=alteración)                    */
        /* recordar que la matriz de traslape no es simétrica                */
        Matriz_traslape_final(Altestudiar,j+in)=tras
/* para dejar la matriz traslape simétrica                                   */
M ← Matriz que contiene los valores simétricos
for i ← 1 to cantidad de alteraciones do
    for ii ← 1 to cantidad de alteraciones do
        M(ii,i)=(Matriz_traslape_final(ii,i)+Matriz_traslape_final(i,ii))/2 M(i,ii)=M(ii,i)
/* aplicar la función dendrograma para generarlos                            */
Dendrogram_Func( M, <nombre archivos> );
```

En la figura A.7 se aprecia una imagen de como se ve la aplicación.

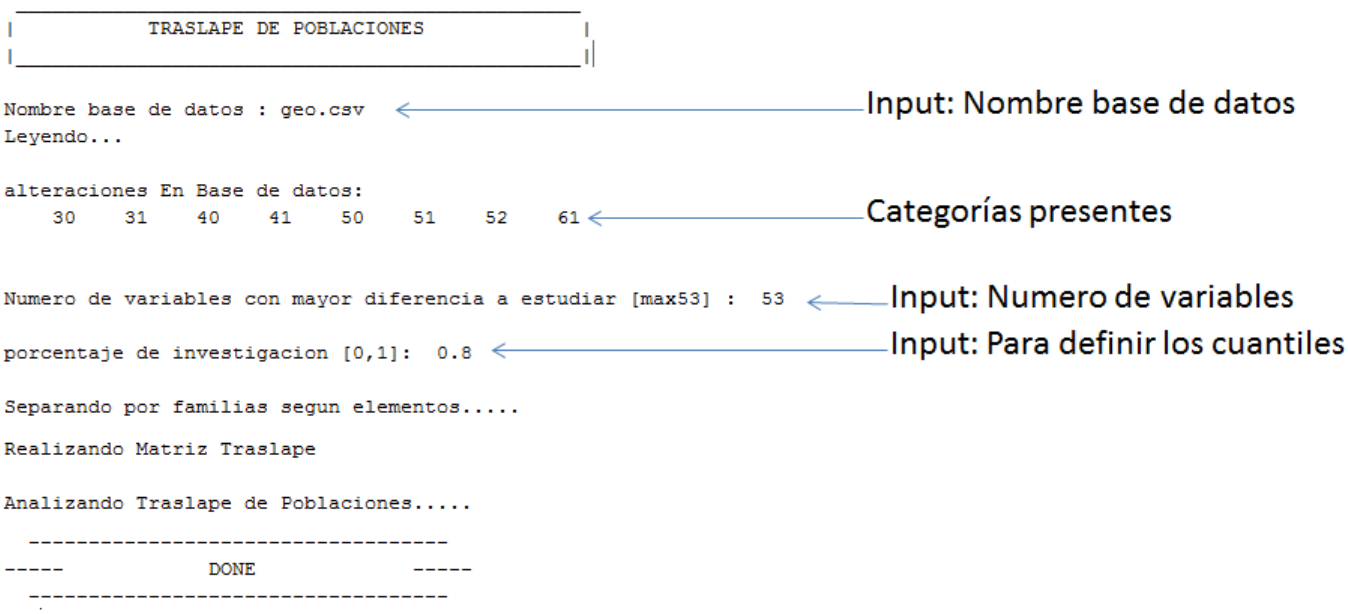


Figura A.7: Input traslape de poblaciones

El formato con el que se ingresa la información es el mismo de la figura A.6

Algoritmo 13: Funcion Dendrograma de Matriz de traslape

Input: Matriz de traslape simétrica, nombre de archivos
Output: 3 Dendrogramas Simple, Complete y Average

```
/* crear carpeta dendograma_ + nombre de archivos, para salvar los
   archivos                                                        */
Crear Capeta(dedrograma)
/* transformar la matriz de traslape a un vector                   */
for i ← 1 to Cantidad de categorias do
    for j ← 1 to Cantidad de categorias do
        if i < j then
            /* convertir el traslape a distancia usando d = 1 - el traslape
               que existe entre las familias                       */
            /* son las combinaciones de distancia que pueden haber entre 2
               categorías, si son 8 categorías sería un vector de largo
               binomio de 8 sobre 2 =28 (deben de ir ordenados)    */
            Distancia((i-1)*(M-i/2)+j-i) ← 1 - Matriz_traslape(i, j)
/* se crea un límite de para colorear familias                     */
colorlim=1-mean(1-Distancia);
/* se aplica la función linkage al vector de Distancia de esta manera se
   ubican las categorías que se van uniendo su punto y los nuevos
   cluster que se van formando a través de una matriz que será analizada
   por la función dendogram, ingresando que modo se desea para unir los
   puntos                                                          */
Link_categorias_single= linkage(Distancia, <single>)
Link_categorias_average= linkage(Distancia, <average>)
Link_categorias_complete= linkage(Distancia, <complete>)
/* generar los dendogramas                                         */
Figura1=Dendogram(Link_categorias_simple)
Figura2=Dendogram(Link_categorias_average)
Figura3=Dendogram(Link_categorias_complete)
/* Guardar los dendrogramas en la carpeta creada                   */
Save(figura1)
Save(figura2)
Save(figura3)
```

Apéndice B

Anexos: Datos ocupados en el estudio

La sección «B» de los anexos incluye la información de las bases de datos antes de la creación de los modelos (antes de un procesamiento), en este se estudian las relaciones entre las variables, sus distribuciones y correlaciones. Básicamente esta sección contiene la información utilizada para comenzar el estudio.

MINERA ESCONDIDA LIMITADA

HOJA CODIFICACION DE SONDAJES INFILL ESCONDIDA-ESCONDIDA NORTE FY12

SONDAJE :

Codificación

Leyes: [ ] si [ ] no

PtXt: [ ] si [ ] no

	Nombre	Fecha
Codificado Por		
Revisado por (1)		
Aprobado por (2)		

Coordenadas	
N	
E	
Cota	
Az	
Incl.	
Largo	

Valid. Bancos: [ ] si [ ] no

Mapeado Por:

[ ] si [ ] no	desde		hasta	
[ ] si [ ] no	desde		hasta	
[ ] si [ ] no	desde		hasta	
[ ] si [ ] no	desde		hasta	

Aprobación MEL

	Nombre	Fecha
Litología		
Alteración		
Minzone		

EXPANSIÓN	N7
SDD	NA

COTA Sondaje						COTA Sección					
TOX		TDS		TDB		TOX		TDS		TDB	
Box		TDCpy		BE		Box		TDCpy		BE	
TS(1)		TDCv				TS(1)		TDCv			
Tyeso		T Anhidrita+ yeso		Techo Anhidrita							

LITOLOGIA	Abrev.	Código	Desde	Hasta	Código	Desde	Hasta	Código
Pórfido Feldespático	PF	31						
Pórfido Riolítico (ENorte)	PR	32						
Pórfido Cuarcífero (Mina)	PC	33						
Pórfido Grueso (ENorte)	PG	34						
Pórfido Dacítico (tardia)	DT	35						
Porfido Feldespático Tardio	PFT	70						
Pórfido Negro	PNO	71						
Porfido Dacitico de Anfibol	PDA	72						
Porfido Temprano	PT	73						
Andesita	AN	50						
Autobrecha (Volcánica)	AB	51						
Unidad Volcano-Sedimentaria	UVS	81						
Tobas Daciticas	TD	53						
Tobas Riolíticas	TR	54						
Brecha Hidrotermal	BH	21						
Brecha Tectónica	BT	22						
Brecha Magmática-Ignea-Contacto	BI	23						
Brecha Pebbles Dikes	BP	24						
Grava	GR	6						
Diorita	DR	40						
Relleno Artificial	RA	1						
Relleno Acopio	RS	2						
Arenisca Volcánica	AV	64						
Roca Moteada	RM	10						
Dique (en el sentido amplio)	D	11						

ZONAS MINERALOGICAS	Abrev.	Código	Desde	Hasta	Código	Desde	Hasta	Código
Lixiviado	LX	10						
Oxidos Verdes (Sulfatos)	OXV	21						
Oxidos Azules (Silicatos)	OXA	22						
Oxidos Negros	OXN	23						
Cuprita + Ox Cu	CPOX	24						
Oxidos Verdes+Negro+Arcillas Cupríferas	OXVN	25						
Cuprita / Cobre Nativo	CPCu	26						
Fosfatos + Oxidos Verdes	POXV	27						
Parcial Lixiviado (Oxidación Parcial)	Plx	31						
Parcial Lixiviado (Limonitas Exóticas)	PlxE	32						
Mezcla OxCu + Sulf	MX	40						
Mixto + Cuprita y/o Nativo	CPMX	41						
Súlfuro + Cuprita y/o Nativo	CPCCPY	42						
Cc + Py	HE1	50						
Cc + Cv + Py	HE2	51						
Cv + Py	HE3	52						
Cc + Cpy + Py	LE1	60						
Cc + Cv + Cpy + Py	LE2	61						
Cv + Cpy + Py	LE3	62						
Cv + Cpy + Bo (+/- Cc, +/- Py)	LE4	63						
Py	PR1	72						
Cpy - Py	PR2	70						
Bn - Cpy (+/- Py)	PR3	71						
Lixiviación Hipógena (Mg-Esp)	LxH	73						
Sin Mineralización	SM	80						
skarnificación	sk	90						

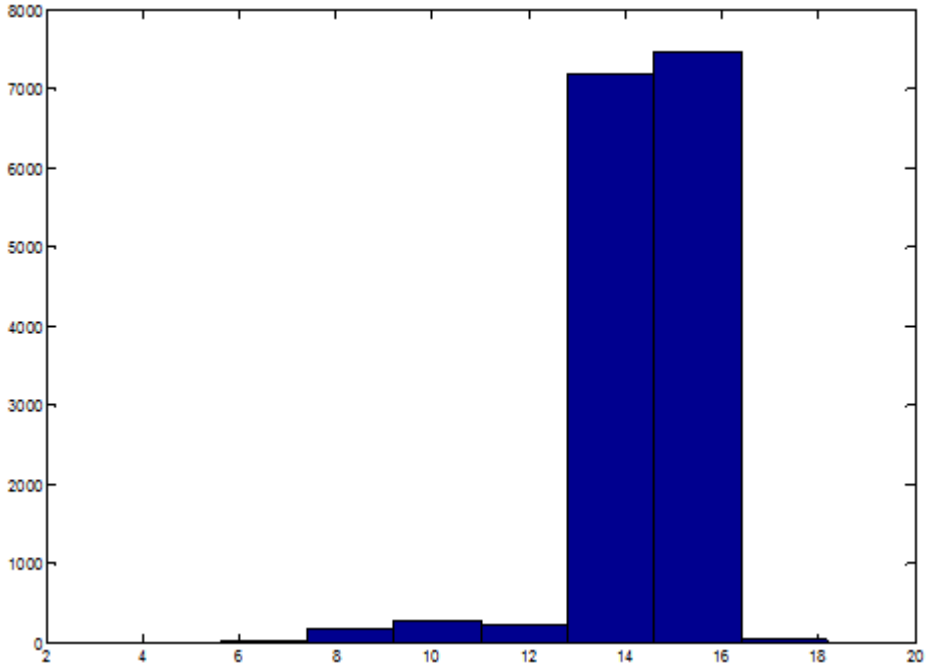
ALTERACION	Abrev.	Código	Desde	Hasta	Código	Desde	Hasta	Código
Fresco: sin Alteración	F	10						
Propilítico	P	20						
Potásico (Sólo Bt secundaria)	K1	30						
Feld K > Bt	K2	31						
Feld K < Bt	K3	32						
(Bt sec +/- Feld K) > Cl	K4	33						
Clorita-Sericita-Arcillas	SCC1	40						
Clorita-Sericita-Cuarzo	SCC2	41						
Clorita-Biotita +/- Feld K	SCC3	42						
Clorita	SCC4	43						
Sericita-Cuarzo	S1	51						
Sericita-Cuarzo-Arcilla	S2	52						
Sericita Gris Verde	S4	54						
Argilización Supérgena	AA	61						
Argilización Avanzada	AAV	62						
Silicificación	Q	70						

Paso a Sección: [ ] SI [ ] No

Ingresado Base Datos: [ ] SI Fecha [ ]

	ag	mo	re	cd	pb	zn	te	bi	sb	hg	co	ni	se	al	as	b	ba	be	ca	ce	cr	cs	cu	ga	ge	hf	in	k	la	li	mg	mn	na	nb	p	rb	s	sc	sn	sr	ta	th	ti	tl	u	v	w	y	zr
ag	1.00																																															
mo	0.08	1.00																																														
re	0.14	0.27	1.00																																													
cd	0.20	0.04	0.20	1.00																																												
pb	0.43	0.03	-0.02	0.30	1.00																																										
zn	0.39	0.01	0.00	0.35	0.61	1.00																																									
te	0.24	-0.03	-0.01	0.01	0.04	0.00	1.00																																									
bi	0.62	0.04	0.15	0.10	0.13	0.19	0.39	1.00																																								
sb	0.44	0.02	0.03	0.08	0.29	0.18	0.29	0.28	1.00																																							
hg	0.42	0.07	0.16	0.15	0.32	0.25	0.18	0.30	0.48	1.00																																						
co	-0.02	0.01	0.17	0.10	0.03	0.08	0.01	0.02	-0.01	0.03	1.00																																					
ni	0.06	0.04	-0.02	0.06	0.13	0.14	-0.01	0.03	0.03	0.03	0.08	1.00																																				
se	0.40	0.18	0.19	0.07	0.09	0.09	0.28	0.38	0.17	0.20	0.09	-0.01	1.00																																			
al	-0.07	0.02	-0.11	0.09	0.12	0.29	-0.17	-0.17	-0.10	-0.09	0.19	0.15	-0.19	1.00																																		
as	0.40	0.01	0.07	0.03	0.06	0.05	0.30	0.37	0.56	0.45	-0.01	-0.02	0.17	-0.15	1.00																																	
b	0.01	0.04	-0.01	-0.02	-0.02	-0.04	0.07	0.02	0.01	0.02	-0.02	-0.01	0.07	-0.06	0.02	1.00																																
ba	-0.11	-0.06	-0.14	-0.02	0.03	0.06	-0.16	-0.17	-0.08	-0.09	-0.13	0.03	-0.21	0.25	-0.10	-0.01	1.00																															
be	-0.02	0.01	-0.17	0.09	0.16	0.36	-0.16	-0.12	-0.07	-0.09	0.06	0.09	-0.14	0.74	-0.13	-0.04	0.26	1.00																														
ca	-0.03	0.00	-0.13	-0.01	0.06	0.10	-0.04	-0.06	-0.01	-0.02	0.01	0.03	-0.05	0.28	-0.04	-0.02	0.13	0.25	1.00																													
ce	0.02	0.09	-0.06	0.10	0.14	0.35	-0.23	-0.10	-0.10	-0.10	-0.07	0.06	-0.11	0.47	-0.16	-0.07	0.28	0.61	0.09	1.00																												
cr	0.11	-0.04	0.04	-0.01	0.00	0.00	0.12	0.14	0.07	0.03	-0.08	-0.09	0.15	-0.23	0.06	0.01	0.00	-0.13	-0.01	-0.05	1.00																										
cs	0.00	-0.07	-0.13	0.08	0.14	0.25	-0.08	-0.03	-0.04	-0.04	0.06	0.02	-0.24	0.47	-0.06	-0.02	0.20	0.50	0.20	0.37	-0.12	1.00																									
cu	0.33	0.20	0.20	0.08	-0.01	-0.01	0.09	0.21	0.08	0.11	-0.07	-0.01	0.48	-0.11	0.12	0.03	-0.17	-0.15	-0.15	-0.03	0.03	-0.21	1.00																								
ga	0.00	-0.02	-0.16	0.09	0.19	0.41	-0.14	-0.10	-0.05	-0.05	0.13	0.15	-0.15	0.91	-0.12	-0.05	0.28	0.76	0.32	0.55	-0.14	0.48	-0.16	1.00																								
ge	0.18	0.04	-0.05	0.01	0.07	0.10	0.42	0.27	0.21	0.13	0.13	0.00	0.33	0.13	0.28	0.03	-0.06	0.13	0.11	0.01	0.04	0.09	-0.01	0.19	1.00																							
hf	-0.03	0.01	-0.03	0.06	0.02	0.05	-0.02	-0.02	-0.01	-0.02	0.02	0.04	-0.04	0.20	-0.03	0.02	0.10	0.18	0.07	0.12	-0.04	0.14	-0.02	0.20	0.03	1.00																					
in	0.43	0.06	0.31	0.27	0.22	0.45	0.13	0.37	0.15	0.26	0.09	0.02	0.22	-0.05	0.14	-0.01	-0.10	-0.02	0.00	0.03	0.07	0.07	0.15	0.00	0.09	-0.02	1.00																			
k	-0.01	0.05	-0.07	0.02	0.08	0.15	-0.16	-0.11	-0.07	-0.04	0.05	0.13	-0.14	0.55	-0.09	-0.05	0.43	0.35	0.23	0.28	-0.16	0.33	-0.08	0.53	0.03	0.11	-0.03	1.00																		
la	0.03	0.10	-0.04	0.10	0.13	0.33	-0.23	-0.09	-0.10	-0.09	-0.09	0.04	-0.10	0.40	-0.16	-0.07	0.27	0.56	0.06	0.98	-0.03	0.35	-0.02	0.48	0.00	0.10	0.04	0.25	1.00																	
li	-0.03	0.02	-0.13	0.10	0.17	0.39	-0.14	-0.13	-0.06	-0.08	0.16	0.15	-0.14	0.87	-0.13	-0.04	0.18	0.75	0.29	0.55	-0.16	0.47	-0.13	0.89	0.17	0.18	-0.06	0.42	0.49	1.00															
mg	-0.04	0.02	-0.15	0.07	0.15	0.37	-0.13	-0.13	-0.06	-0.09	0.15	0.17	-0.14	0.90	-0.13	-0.04	0.18	0.75	0.34	0.53	-0.16	0.44	-0.14	0.94	0.18	0.19	-0.06	0.46	0.46	0.94	1.00															
mn	0.08	-0.01	-0.13	0.08	0.31	0.50	-0.07	-0.02	-0.01	-0.03	0.07	0.13	-0.16	0.56	-0.08	-0.04	0.21	0.56	0.30	0.47	-0.07	0.49	-0.20	0.67	0.12	0.11	0.05	0.28	0.43	0.64	0.67	1.00													
na	0.01	0.00	-0.09	0.08	0.17	0.32	-0.13	-0.08	-0.04	-0.03	0.10	0.21	-0.10	0.66	-0.11	-0.01	0.29	0.53	0.15	0.41	-0.10	0.31	-0.10	0.70	0.10	0.13	-0.02	0.42	0.37	0.60	0.63	0.43	1.00												
nb	0.02	-0.04	-0.07	-0.02	0.03	0.02	0.11	0.05	0.04	0.02	0.01	0.00	0.05	0.08	0.01	-0.01	0.07	0.05	0.30	0.01	0.09	0.07	-0.12	0.20	0.28	0.07	-0.01	0.13	0.00	0.12	0.15	0.13	0.12	1.00											
p	0.01	0.04	-0.15	0.11	0.21	0.39	-0.13	-0.11	-0.04	-0.06	0.10	0.16	-0.14	0.77	-0.11	-0.05	0.18	0.75	0.35	0.61	-0.16	0.43	-0.14	0.78	0.15	0.15	-0.02	0.36	0.54	0.83	0.81	0.62	0.56	0.12	1.00										
rb	-0.01	0.03	-0.04	0.03	0.09	0.20	-0.14	-0.08	-0.06	-0.05	0.06	0.10	-0.13	0.57	-0.09	-0.05	0.38	0.41	0.24	0.36	-0.13	0.47	-0.08	0.63	0.09	0.14	-0.02	0.89	0.33	0.51	0.55	0.37	0.45	0.22	0.42	1.00									
s	0.25	-0.03	0.04	0.03	0.08	0.01	0.46	0.42	0.24	0.21	0.26	0.01	0.38	-0.28	0.27	0.05	-0.34	-0.28	0.19	-0.46	0.09	-0.12	0.02	-0.28	0.36	-0.05	0.23	-0.21	-0.46	-0.25	-0.23	-0.15	-0.29	0.13	-0.22	-0.24	1.00								
sc	-0.03	0.01	-0.10	0.08	0.12	0.27	-0.12	-0.11	-0.06	-0.06	0.16	0.17	-0.11	0.82	-0.10	-0.04	0.24	0.58	0.29	0.39	-0.15	0.41	-0.07	0.86	0.16	0.22	-0.05	0.56	0.32	0.75	0.81	0.49	0.67	0.19	0.65	0.67	-0.25	1.00							
sn	0.19	0.02	0.02	-0.02	-0.02	-0.05	0.64	0.35	0.23	0.14	0.01	0.02	0.24	-0.11	0.22	0.04	-0.10	-0.12	-0.02	-0.19	0.09	-0.09	0.07	-0.09	0.39	-0.01	0.06	-0.07	-0.19	-0.09	-0.08	-0.07	-0.10	0.12	-0.09	-0.06	0.36	-0.06	1.00						
sr	-0.01	0.05	-0.03	0.00	-0.06	-0.07	-0.03	-0.01	-0.03	0.00	0.00	-0.06	0.04	0.13	0.00	0.01	0.02	0.28	0.35	0.07	-0.06	0.13	0.01	0.06	0.04	0.02	0.07	0.04	0.06	0.01	0.02	-0.05	-0.05	0.02	0.16	0.01	0.10	0.02	-0.02	1.00					
ta	0.00	-0.01	-0.01	0.03	0.00	0.07	0.01	0.03	0.00	0.03	0.01	0.06	0.02	-0.02	0.01	0.00	-0.01	-0.02	0.00	-0.01	0.00	-0.01	0.00	0.00	0.04	0.07	0.07	-0.01	-0.01	-0.01	-0.01	0.01	0.01	0.04	-0.01	-0.02	0.04	0.00	0.01	-0.02	1.00				
th	-0.03	0.02	-0.14	0.04	0.09	0.25	-0.20	-0.10	-0.09	-0.09	-0.08	0.03	-0.08	0.36	-0.14	-0.06	0.25	0.55	0.28	0.67	0.03	0.24	-0.09	0.47	0.04	0.11	-0.02	0.24	0.65	0.43	0.45	0.37	0.33	0.18	0.45	0.30	-0.28	0.34	-0.14	0.12	0.02	1.00			
ti	-0.05	0.00	-0.08	-0.01	0.06	0.12	-0.07	-0.08	-0.03	-0.04	0.07	0.13	-0.10	0.49	-0.06	-0.02	0.30	0.27	0.24	0.20	-0.06	0.31	-0.10	0.59	0.15	0.13	-0.08	0.63	0.16	0.47	0.53	0.31	0.53	0.40	0.38	0.74	-0.19	0.71	0.00	-0.07	0.00	0.21	1.00		
tl	0.14	0.12	0.52	0.11	0.00	-0.01	0.00	0.12	0.03	0.13	0.25	0.03	0.12	0.06	0.09	-0.01	-0.02	-0.12	-0.12	-0.11	-0.04	0.05	0.22	-0.01	-0.03	0.03	0.20	0.29	-0.11	-0.05	-0.06	-0.08	0.00	-0.05	-0.10	0.30	0.02	0.08	0.03	-0.04	0.01	-0.20	0.11	1.00	
u	0.19	0.09	0.18	0.20	0.05	0.13	-0.05	0.18	0.02	0.05	0.06	-0.01	0.06	0.05	0.03	-0.04	0.07	0.17	0.07	0.28	0.05	0.15	0.06	0.07	0.01	0.10	0.27	0.07	0.30	0.03	0.02	0.05	0.05	0.04	0.06	0.11	-0.02	0.05	-0.02	0.16	-0.01	0.35	-0.02	0.10	1.00	
v	-0.04	0.01	-0.13	0.07	0.12	0.30	-0.10	-0.12	-0.05	-0.07	0.17	0.16	-0.14	0.86	-0.10	-0.04	0.23	0.62	0.31	0.41	-0.15	0.44	-0.10	0.90	0.18	0.20	-0.07	0.54	0.34	0.81	0.88	0.56	0.70	0.19	0.70	0.63	-0.26	0.94	-0.04	0.01	-0.01	0.34	0.70	0.03	0.01	1.00	
w	0.06	0.00	0.01	0.00	0.01	0.00	0.06	0.05	0.04	0.13	0.06	0.00	0.04	-0.03	0.06	0.00	-0.01	-0.03	0.01	-0.05	0.01	-0.01	-0.01	-0.02	0.05	0.00	0.03	-0.01	-0.05	-0.03	-0.03	-0.02	0.00	0.02	-0.02	-0.01	0.07	-0.02	0.13	0.03	0.00	-0.02	0.00	0.02	0.02	-0.02	1.00	
y	0.00	0.04	-0.13	0.17	0.14	0.32	-0.14	-0.11	-0.06	-0.08	0.05	0.11	-0.08	0.67	-0.12	-0.05	0.19	0.72	0.22	0.59	-0.12	0.40	0.00	0.70	0.10	0.35	-0.04	0.32	0.53	0.72	0.71	0.47	0.49	0.07	0.69	0.39	-0.27	0.62	-0.09	0.10	0.00	0.50	0.33	-0.09	0.20	0.65	-0.03	1.00	
zr	-0.02	0.01	-0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.01	-0.02	0.02	0.01	0.03	0.06	0.02	0.05	-0.01	0.00	0.02	-0.03	0.03	0.02	0.42	-0.01	0.01	-0.01	0.02	0.02	0.00	0.02	0.01	0.00	0.01	0.00	0.04	0.00	0.01	0.00	0.01	0.02	0.00	0.01	0.03	0.00	0.01	1.00

Base de datos Geo-Química Histograma Longitudes de compósitos.



Estadísticos base de datos unida, sin filtrar

Variable	Media	Varianza de la muestra	Desviación estándar	Min	Max	Numero de datos validos
Largo	14.8	2.22	2	2	20	15011
Ag	3.02	13.15	0.01	0.01	83.1	15011
Al	0.93	0.45	0.06	0.06	4.7	15011
As	41.25	28573.39	0.1	0.1	7390	15011
Ba	40.86	550.79	0.2	0.2	270	15011
Be	0.25	0.04	0.05	0.05	1.76	15011
Bi	1.21	3.45	0.01	0.01	59.1	15011
Ca	0.26	0.28	0.01	0.01	7.41	15011
Cd	4.18	187.47	0.01	0.01	601	15011
Ce	16.49	73.94	0.02	0.02	51.9	15011
Co	12.45	95.22	0.1	0.1	349	15011
Cr	107.61	10069.45	1	1	1560	15011
Cs	1.14	0.47	0.05	0.05	15.5	15011
Cu	5919.04	7910220.19	23	23	10000	15011
In	0.27	0.2	0.01	0.01	16.1	15011
K	0.23	0.01	0.01	0.01	1.61	15011
Mg	0.42	0.32	0.01	0.01	3.13	15011
Mn	266.53	193393.34	5	5	7830	15011
Mo	89.9	13550.52	0.1	0.1	2890	15011
Na	0.06	0	0.01	0.01	0.5	15011
Ni	22.14	1866.63	0.2	0.2	717	15011
P	439.06	213925.19	10	10	4420	15011
Pb	71.56	65416.77	0.2	0.2	10000	15011
Rb	12.57	43.15	0.1	0.1	109	15011
Re	0.53	0.97	0	0	25.5	15011
S	2.27	2.74	0.01	0.01	10	15011
Sb	2.1	121.59	0.05	0.05	472	15011
Sc	1.64	3.72	0.1	0.1	21.5	15011
Se	3.34	2.79	0.2	0.2	26.5	15011
Sn	0.8	3.55	0.2	0.2	101	15011
Sr	92.36	11151.02	0.2	0.2	1240	15011
Te	0.75	3.9	0.01	0.01	68.3	15011
Th	1.75	1.22	0.2	0.2	37.3	15011
Tl	0.18	0.01	0.02	0.02	2.06	15011
U	0.42	0.1	0.05	0.05	8.35	15011
W	1.04	60.49	0.05	0.05	540	15011
Y	4.68	31.61	0.05	0.05	83.9	15011
Zn	506.29	867708.63	2	2	10000	15011
CuT	0.72	0.37	0.01	0.01	8.04	15011
CuS	0.06	0.03	0	0	5.29	15011
Fe	2.58	2.33	0.11	0.11	36.66	15011
KxAl	0.34	0.03	0.03	0.03	5.67	15011
KxNA/Al	0.02	0	0	0	0.56	15011
Na/Al	0.08	0	0	0	1.67	15011
(Al+K)/(Na+Ca+Mg)	3.88	10.83	0.08	0.08	26.75	15011
(Al+K+Na)/(Ca+Mg)	6.62	48.56	0.08	0.08	93	15011
(Ca+Na)/(K+Al)	0.3	0.28	0.02	0.02	12.37	15011
Al/(Na+Ca+K)	1.95	1.13	0.06	0.06	11.8	15011
3Al/(K+Na)	9.07	26.63	0.5	0.5	45.14	15011
(K+Al+S)/(Fe+S)	0.75	0.03	0.13	0.13	3.66	15011
(CuxAsxSbxS)/Fe	12767444.2	1.67E+17	0.52	0.52	3.40E+10	15011
(CuxAsxSbxS/Fe) x(Al+K+S)	94556052	1.18E+19	0.71	0.71	3.02E+11	15011
K/(Ca+Na)	1.71	2.09	0.02	0.02	21.86	15011
K/Mg	4.61	28.85	0.06	0.06	153	15011
Mg/Al	0.3	0.09	0.01	0.01	4.19	15011
Mn/Al	235.99	135473.36	2.72	2.72	7395.83	15011

Ocurrencia alteraciones sin filtrar.

Alteración	*Frecuencia*	*% acumulado*	Alteracion	*Frecuencia*	*% acumulado*
10	1	0.01 %	51	4558	30.36 %
20	8	0.06 %	52	2977	50.20 %
30	527	3.57 %	50	2411	66.26 %
31	218	5.02 %	41	2070	80.05 %
32	3	5.04 %	40	1885	92.61 %
33	3	5.06 %	30	527	96.12 %
40	1885	17.62 %	31	218	97.57 %
41	2070	31.41 %	61	197	98.88 %
42	7	31.46 %	53	51	99.22 %
43	46	31.76 %	43	46	99.53 %
50	2411	47.82 %	70	29	99.72 %
51	4558	78.19 %	62	13	99.81 %
52	2977	98.02 %	20	8	99.86 %
53	51	98.36 %	42	7	99.91 %
54	7	98.41 %	54	7	99.95 %
61	197	99.72 %	32	3	99.97 %
62	13	99.81 %	33	3	99.99 %
70	29	100.00 %	10	1	100.00 %

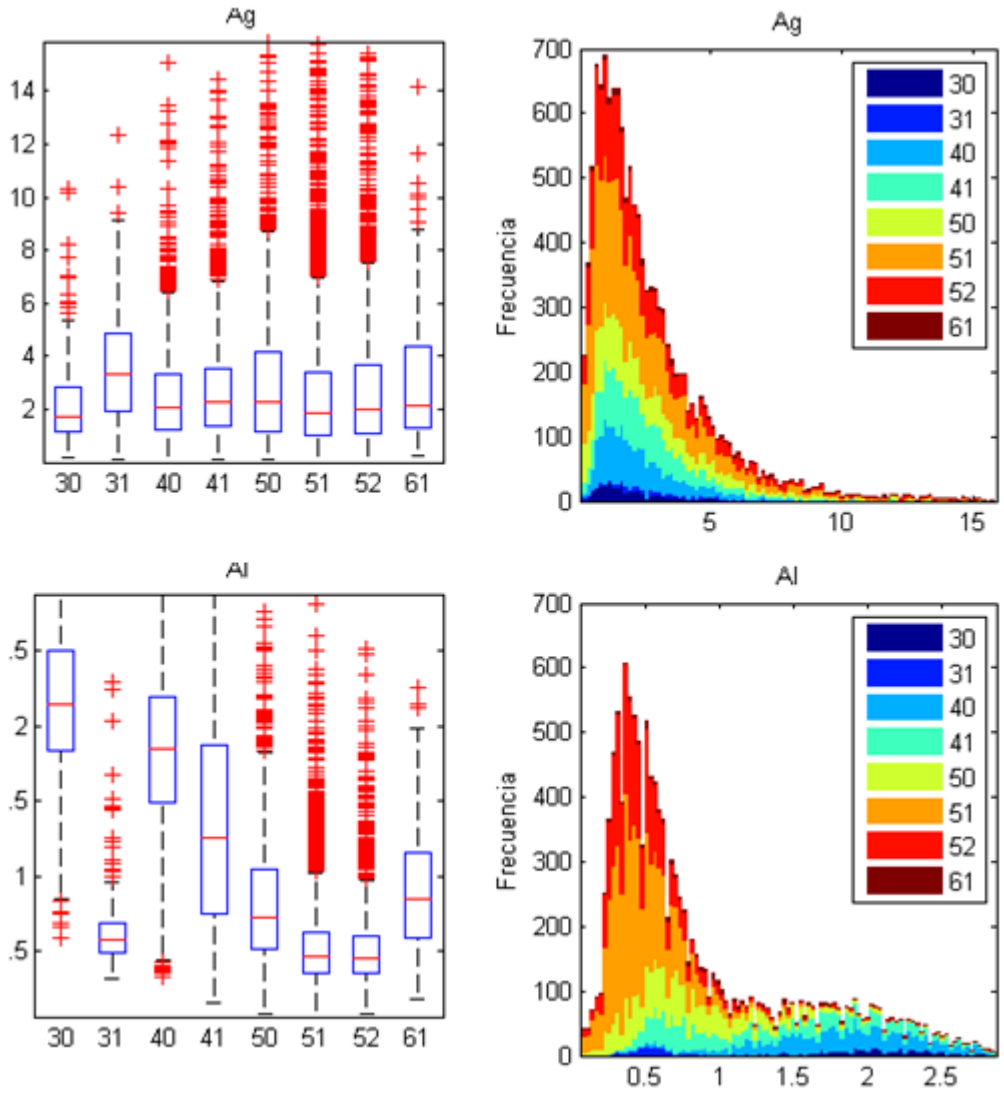
Base de datos con un filtro 70 % de pertenencia al sondaje.

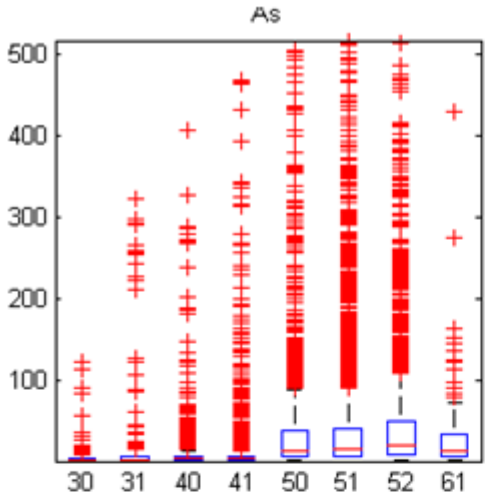
Variable	Media	Varianza de la muestra	Desviación estándar	Min	Max	Numero de datos validos
Largo	14.76	2.29	2	2	18	13165
Ag	2.95	12.19	0.01	0.01	83.1	13165
Al	0.89	0.45	0.06	0.06	4.7	13165
As	42.22	30772.14	0.1	0.1	7390	13165
Ba	40.62	565.48	0.2	0.2	270	13165
Be	0.24	0.04	0.05	0.05	1.76	13165
Bi	1.2	3.59	0.01	0.01	59.1	13165
Ca	0.26	0.29	0.01	0.01	7.41	13165
Cd	3.96	178.57	0.01	0.01	601	13165
Ce	16.15	73.54	0.02	0.02	46.3	13165
Co	12.5	100.69	0.1	0.1	349	13165
Cr	109.2	10406.2	1	1	1560	13165
Cs	1.11	0.45	0.05	0.05	15.5	13165
Cu	5885.93	7947149.66	23	23	10000	13165
In	0.26	0.19	0.01	0.01	9.25	13165
K	0.23	0.01	0.01	0.01	1.53	13165
Mg	0.39	0.31	0.01	0.01	3.13	13165
Mn	248.96	176382.75	5	5	7830	13165
Mo	89.63	13810.56	0.1	0.1	2890	13165
Na	0.06	0	0.01	0.01	0.5	13165
Ni	20.52	1589.58	0.2	0.2	717	13165
P	417.33	211171.77	10	10	4420	13165
Pb	65.35	58756.25	0.2	0.2	10000	13165
Rb	12.29	38.64	0.1	0.1	85.2	13165
Re	0.53	1.01	0	0	25.5	13165
S	2.3	2.79	0.01	0.01	10	13165
Sb	2.1	123.44	0.05	0.05	472	13165
Sc	1.54	3.53	0.1	0.1	21.5	13165
Se	3.36	2.72	0.2	0.2	23.9	13165
Sn	0.82	3.8	0.2	0.2	101	13165
Sr	91.45	10774.03	0.2	0.2	1240	13165
Te	0.76	4.16	0.01	0.01	68.3	13165
Th	1.73	1.26	0.2	0.2	37.3	13165
Tl	0.17	0.01	0.02	0.02	2.06	13165
U	0.41	0.1	0.05	0.05	8.35	13165
W	1.07	68.09	0.05	0.05	540	13165
Y	4.36	29.5	0.05	0.05	74.2	13165
Zn	474.74	808281.53	2	2	10000	13165
CuT	0.71	0.36	0.01	0.01	8.04	13165
CuS	0.06	0.03	0	0	5.29	13165
Fe	2.58	2.4	0.11	0.11	36.66	13165
KxAl	0.35	0.03	0.03	0.03	5.67	13165
KxNA/Al	0.02	0	0	0	0.56	13165
Na/Al	0.09	0	0	0	1.67	13165
(Al+K)/(Na+Ca+Mg)	3.99	10.99	0.08	0.08	26.75	13165
(Al+K+Na)/(Ca+Mg)	6.88	49.72	0.08	0.08	93	13165
(Ca+Na)/(K+Al)	0.3	0.29	0.02	0.02	12.37	13165
Al/(Na+Ca+K)	1.91	1.1	0.06	0.06	11.8	13165
3Al/(K+Na)	8.86	26.56	0.5	0.5	44.25	13165
(K+Al+S)/(Fe+S)	0.75	0.03	0.13	0.13	3.57	13165
CuxAsxSbxS)/Fe	13691918.3	1.90E+17	0.52	0.52	3.40E+10	13165
((CuxAsxSbxS)/Fe)x(Al+K+S)	101767604	1.34E+19	0.71	0.71	3.02E+11	13165
K/(Ca+Na)	1.74	2.16	0.02	0.02	21.86	13165
K/Mg	4.9	29.69	0.07	0.07	153	13165
Mg/Al	0.28	0.09	0.01	0.01	4.19	13165
Mn/Al	224.72	127163.23	2.72	2.72	7395.83	13165

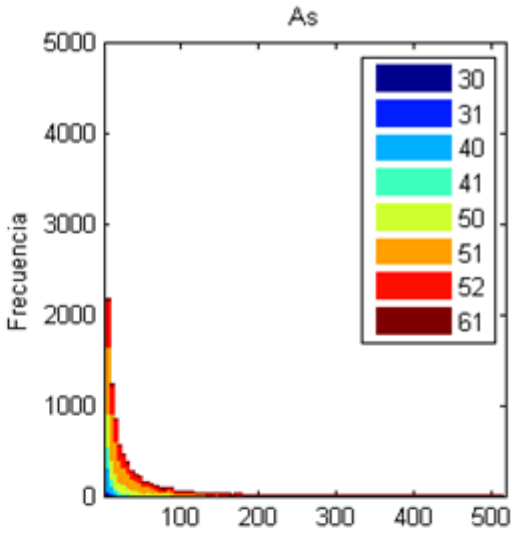
Ocurrencia alteraciones, con 70 % de filtro.

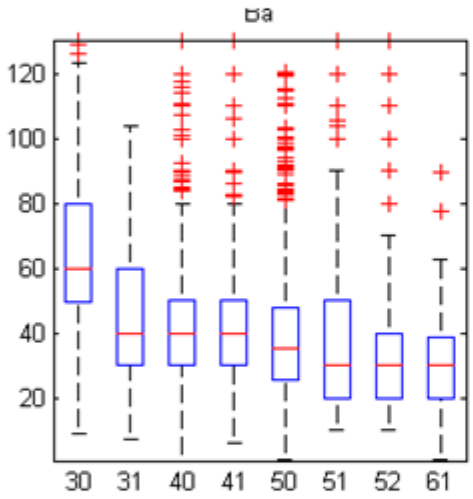
Alteraciones	*Frecuencia*	*% acumulado*	*Alteraciones*	*Frecuencia*	*% acumulado*
20	5	0.04 %	51	4265	32.40 %
30	381	2.93 %	52	2713	53.00 %
31	179	4.29 %	50	2011	68.28 %
32	2	4.31 %	41	1826	82.15 %
33	3	4.33 %	40	1520	93.70 %
40	1520	15.88 %	30	381	96.59 %
41	1826	29.75 %	31	179	97.95 %
42	6	29.79 %	61	143	99.04 %
43	32	30.03 %	53	48	99.40 %
50	2011	45.31 %	43	32	99.64 %
51	4265	77.71 %	70	16	99.76 %
52	2713	98.31 %	62	8	99.83 %
53	48	98.68 %	54	7	99.88 %
54	7	98.73 %	42	6	99.92 %
61	143	99.82 %	20	5	99.96 %
62	8	99.88 %	33	3	99.98 %
70	16	100.00 %	32	2	100.00 %

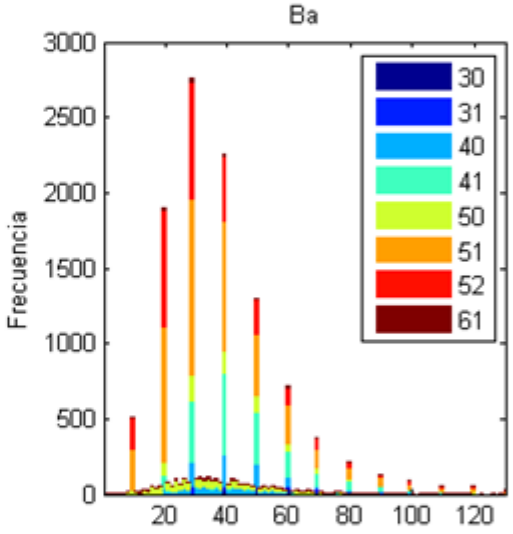
Histograma y Boxplots por alteraciones de mayor ocurrencia

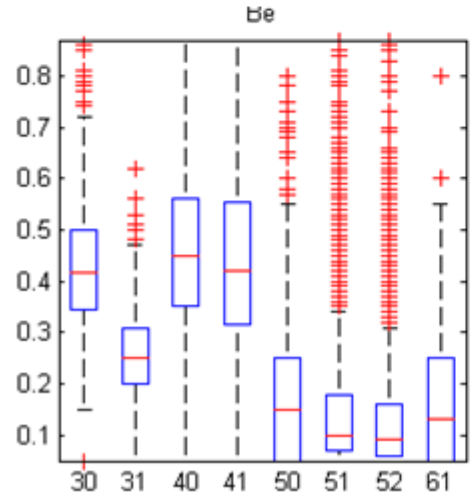


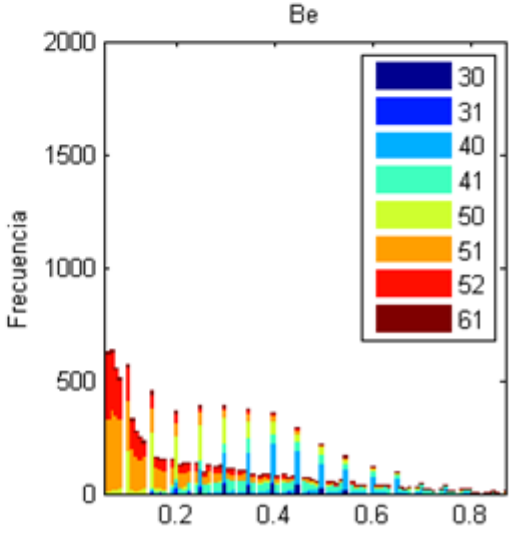

As
500
400
300
200
100
30 31 40 41 50 51 52 61

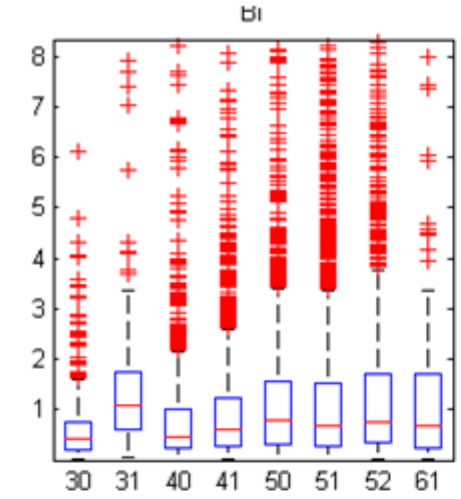

As
5000
4000
3000
2000
1000
0
Frecuencia
100 200 300 400 500
30
31
40
41
50
51
52
61

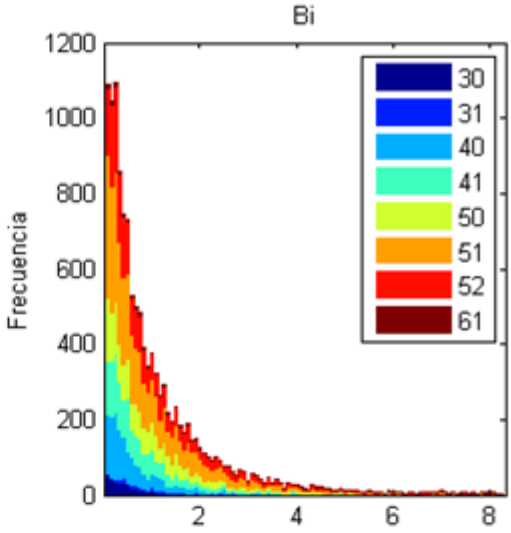

Ba
120
100
80
60
40
20
30 31 40 41 50 51 52 61

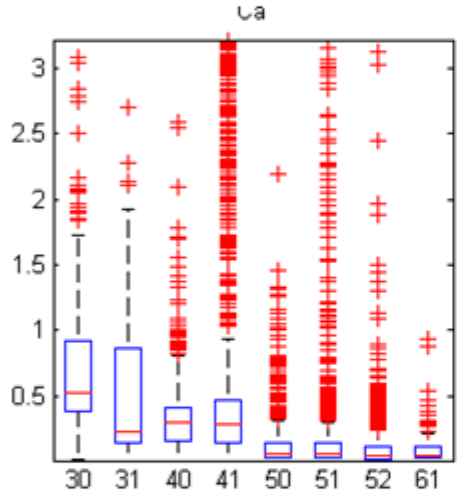

Ba
3000
2500
2000
1500
1000
500
0
Frecuencia
20 40 60 80 100 120
30
31
40
41
50
51
52
61

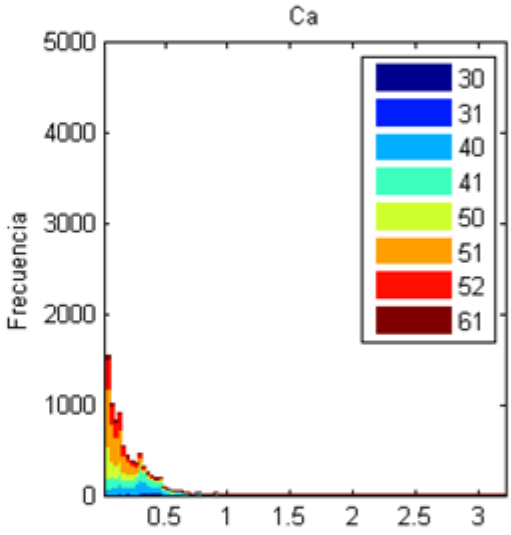

Be
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
30 31 40 41 50 51 52 61

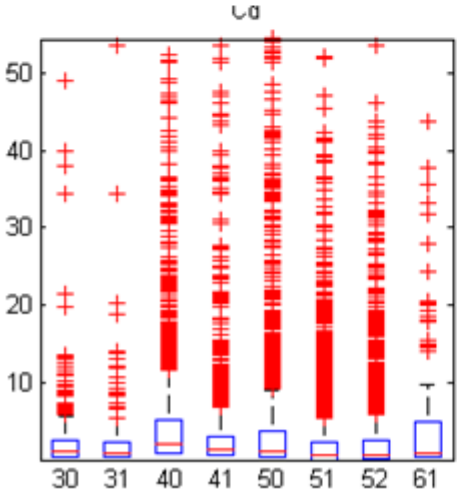

Be
2000
1500
1000
500
0
Frecuencia
0.2 0.4 0.6 0.8
30
31
40
41
50
51
52
61

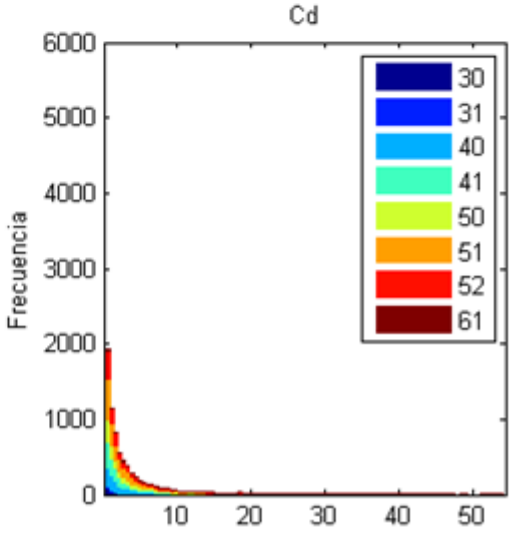

Bi
30
31
40
41
50
51
52
61

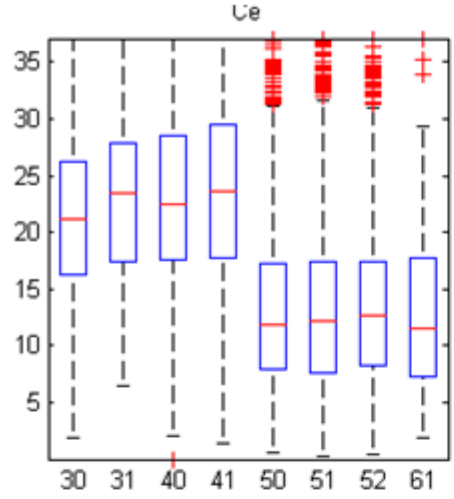

Bi
Frecuencia
30
31
40
41
50
51
52
61

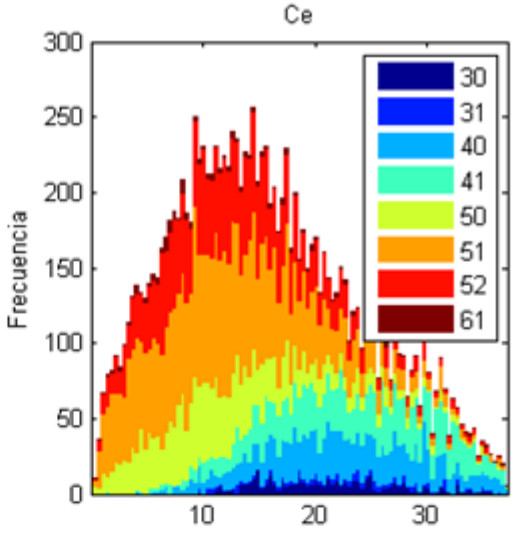

Ca
30
31
40
41
50
51
52
61

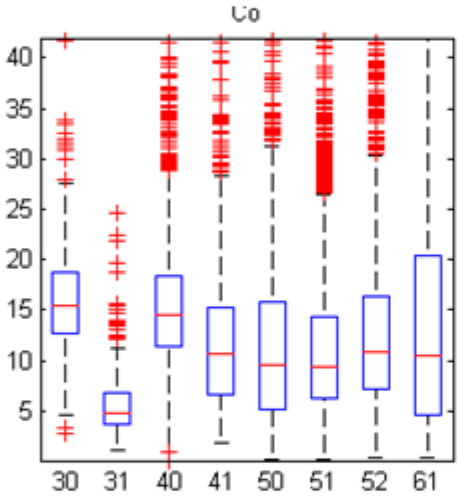

Ca
Frecuencia
30
31
40
41
50
51
52
61

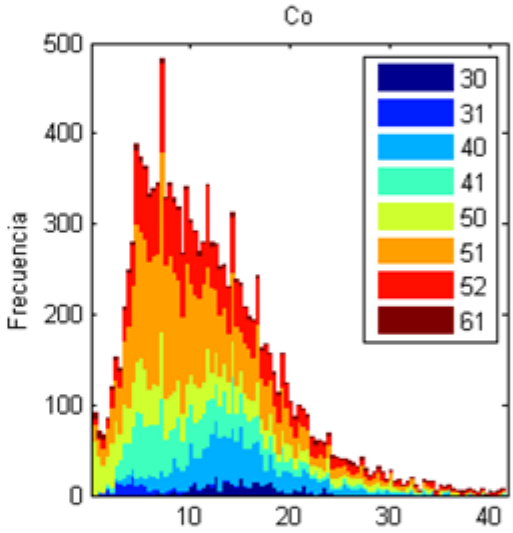

Cd
30
31
40
41
50
51
52
61

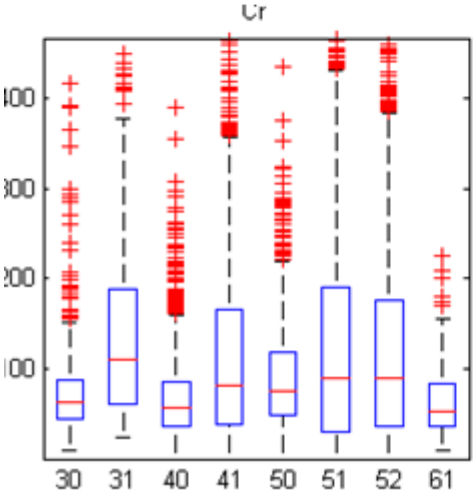

Cd
Frecuencia
30
31
40
41
50
51
52
61

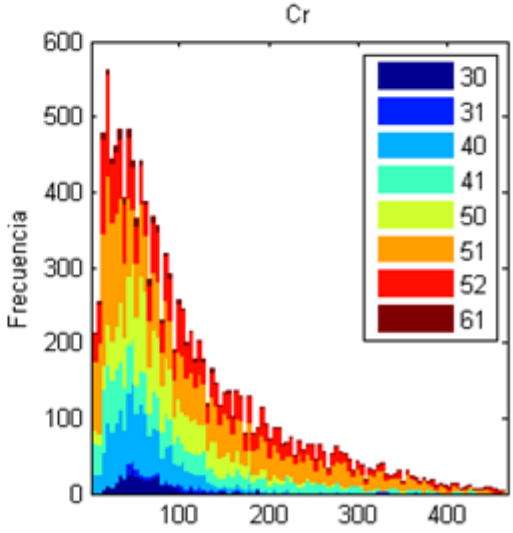

Ce
35
30
25
20
15
10
5
30 31 40 41 50 51 52 61

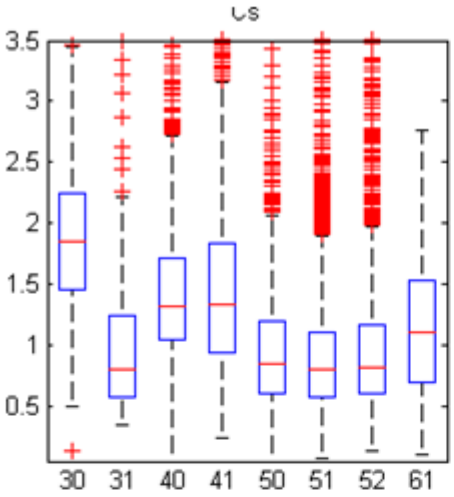

Ce
300
250
200
150
100
50
0
Frecuencia
10 20 30
30
31
40
41
50
51
52
61

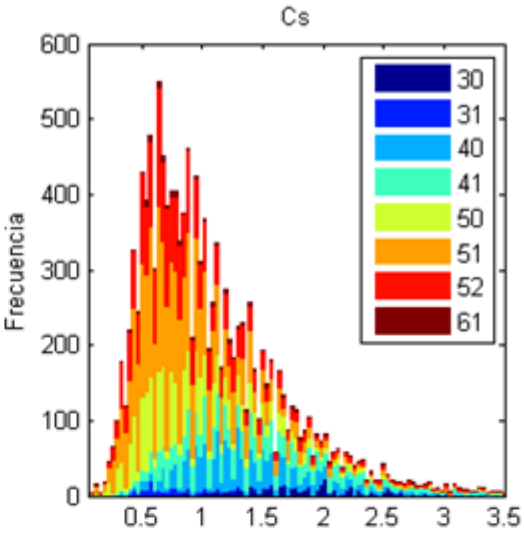

Co
40
35
30
25
20
15
10
5
30 31 40 41 50 51 52 61

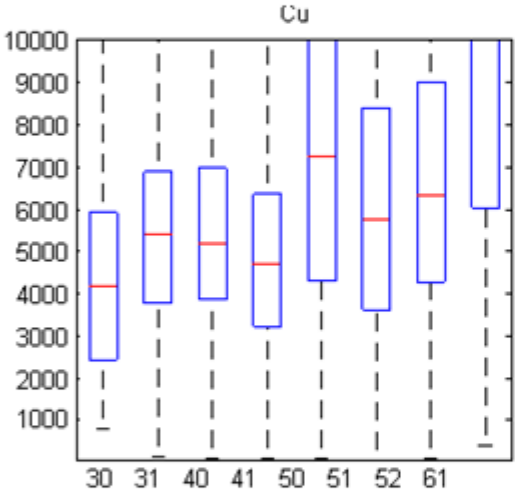

Co
500
400
300
200
100
0
Frecuencia
10 20 30 40
30
31
40
41
50
51
52
61

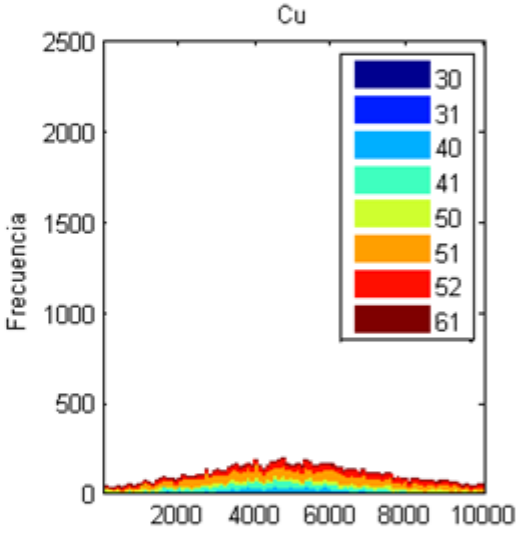

Cr
30 31 40 41 50 51 52 61

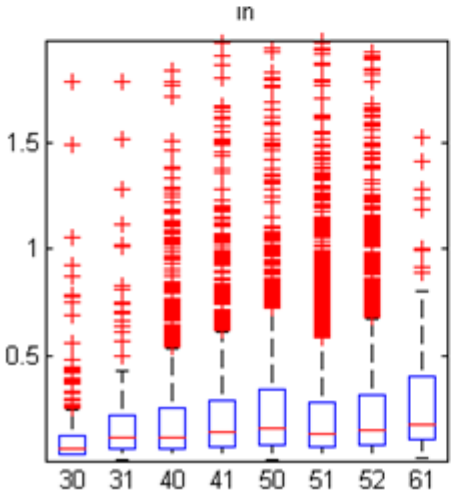

Cr
600
500
400
300
200
100
0
Frecuencia
100 200 300 400
30
31
40
41
50
51
52
61

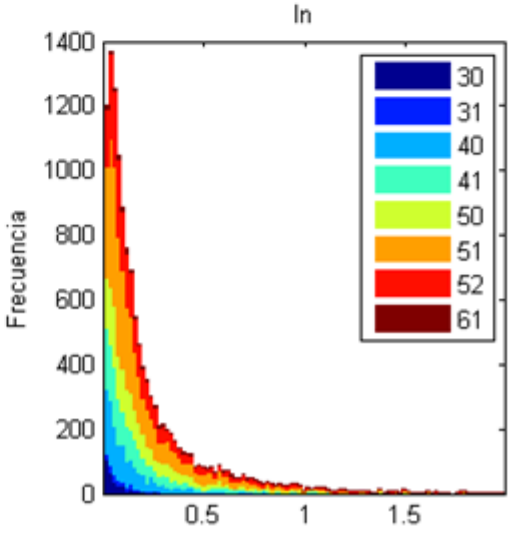

Cs
3.5
3
2.5
2
1.5
1
0.5
30 31 40 41 50 51 52 61

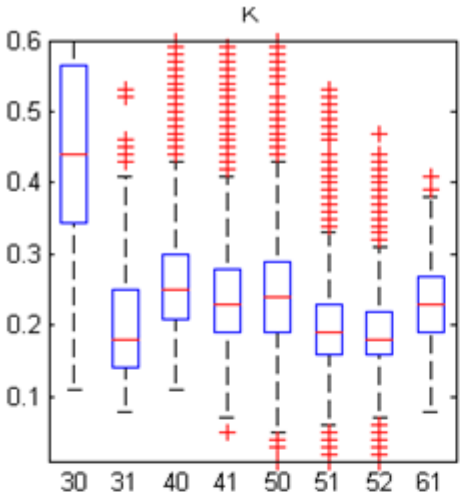

Cs
Frecuencia
600
500
400
300
200
100
0
0.5 1 1.5 2 2.5 3 3.5
30
31
40
41
50
51
52
61

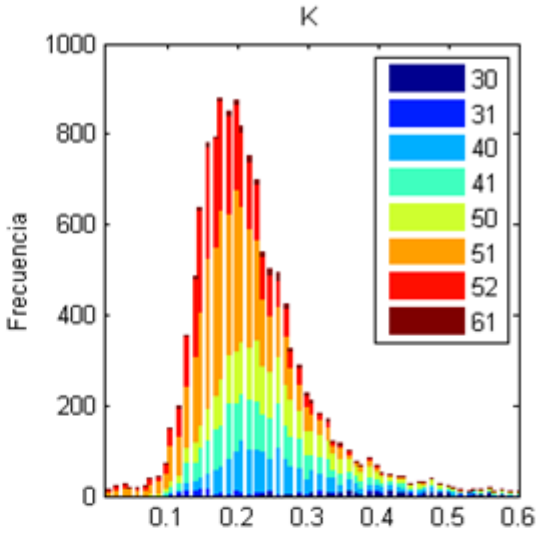

Cu
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
30 31 40 41 50 51 52 61

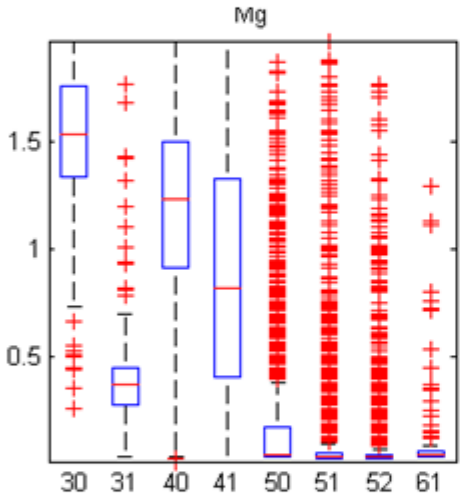

Cu
Frecuencia
2500
2000
1500
1000
500
0
2000 4000 6000 8000 10000
30
31
40
41
50
51
52
61

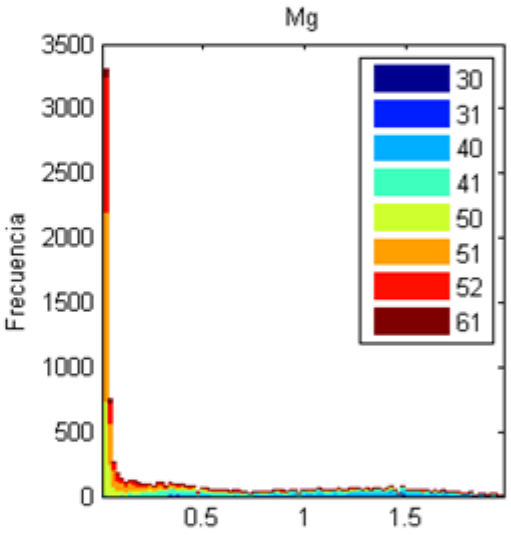

In
1.5
1
0.5
30 31 40 41 50 51 52 61

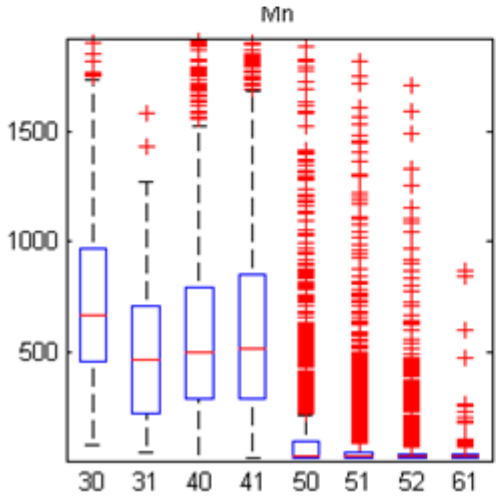

In
Frecuencia
1400
1200
1000
800
600
400
200
0
0.5 1 1.5
30
31
40
41
50
51
52
61

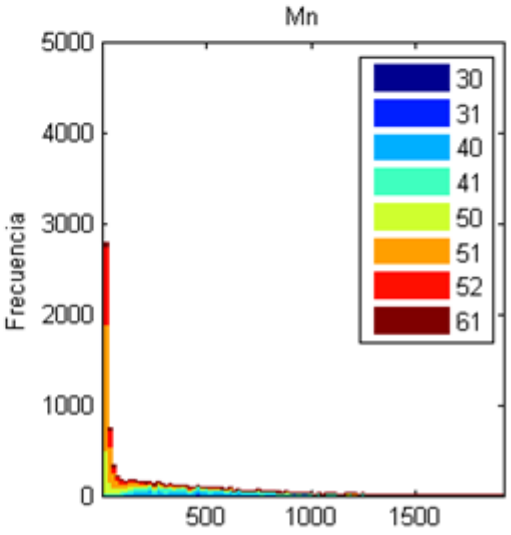

K
0.6
0.5
0.4
0.3
0.2
0.1
30 31 40 41 50 51 52 61

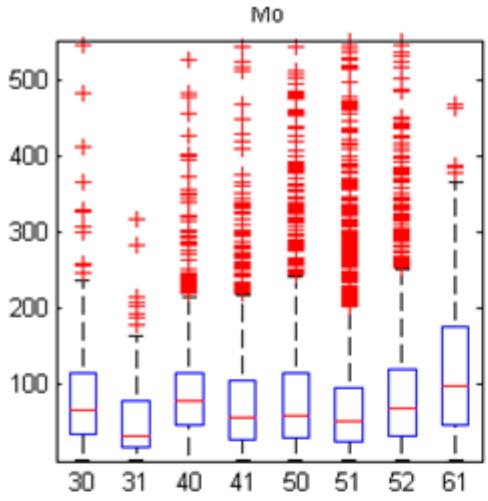

K
Frecuencia
1000
800
600
400
200
0
0.1 0.2 0.3 0.4 0.5 0.6
30
31
40
41
50
51
52
61

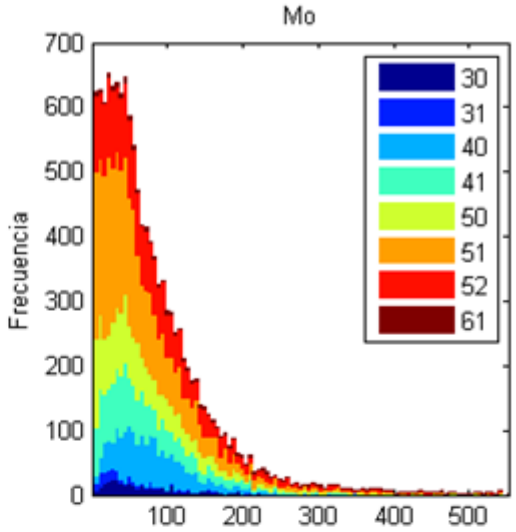

Mg
1.5
1
0.5
30 31 40 41 50 51 52 61

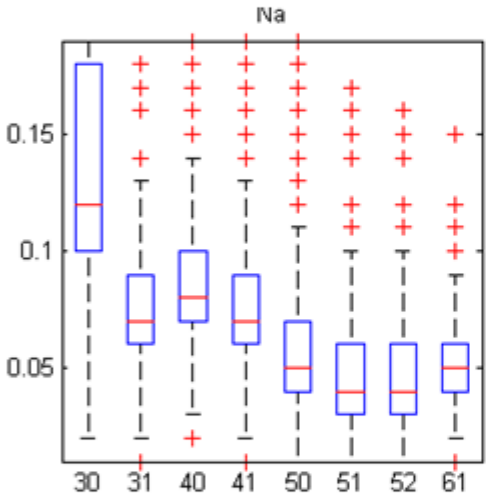

Mg
Frecuencia
3500
3000
2500
2000
1500
1000
500
0
0.5 1 1.5
30
31
40
41
50
51
52
61

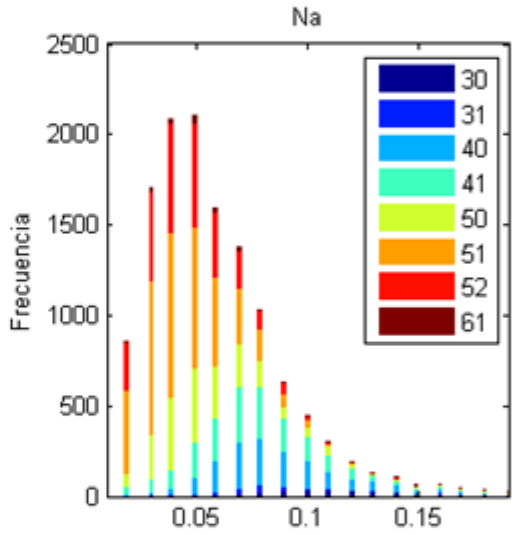

Mn
1500
1000
500
30 31 40 41 50 51 52 61

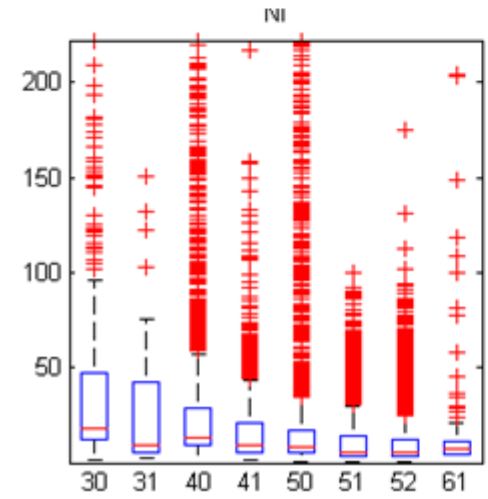

Mn
Frecuencia
5000
4000
3000
2000
1000
0
500 1000 1500
30
31
40
41
50
51
52
61

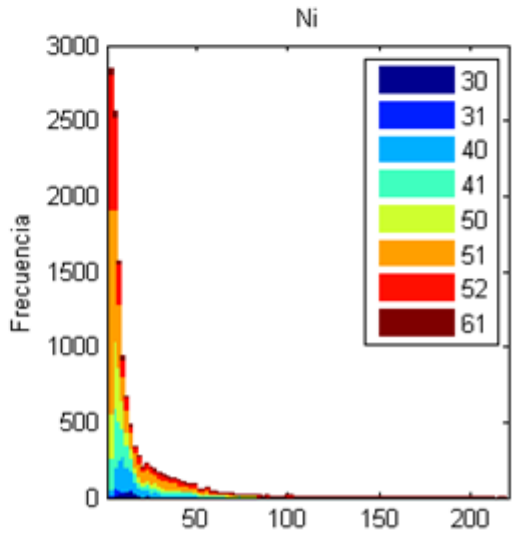

Mo
500
400
300
200
100
30 31 40 41 50 51 52 61

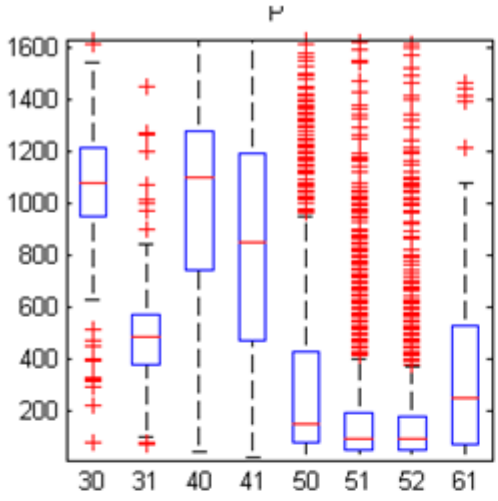

Mo
Frecuencia
700
600
500
400
300
200
100
0
100 200 300 400 500
30
31
40
41
50
51
52
61

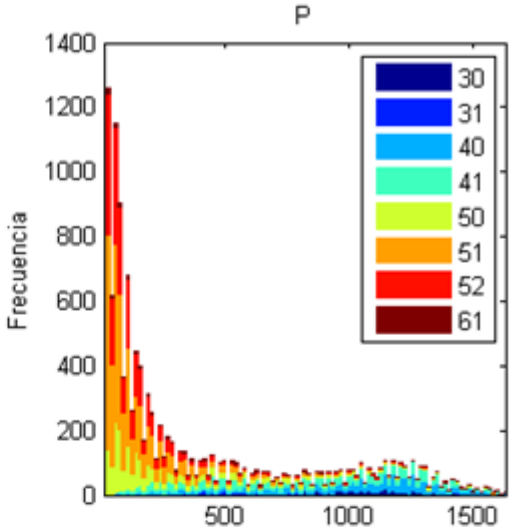

Na
0.15
0.1
0.05
30 31 40 41 50 51 52 61

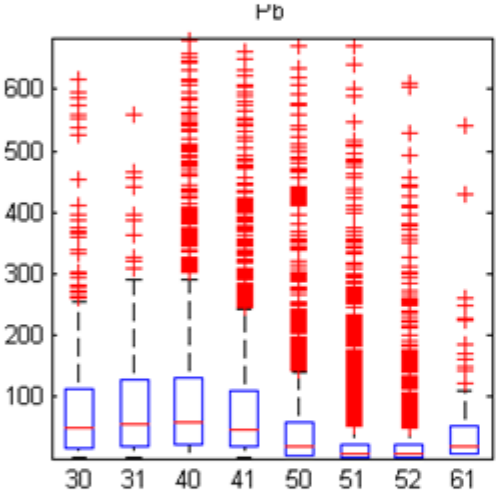

Na
Frecuencia
2500
2000
1500
1000
500
0
0.05 0.1 0.15
30
31
40
41
50
51
52
61

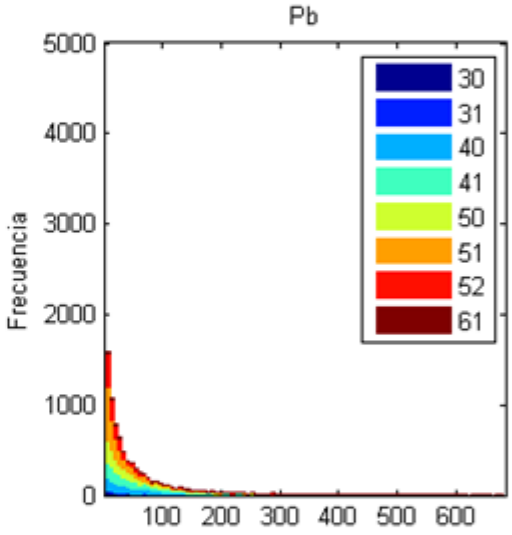

Ni
200
150
100
50
30 31 40 41 50 51 52 61

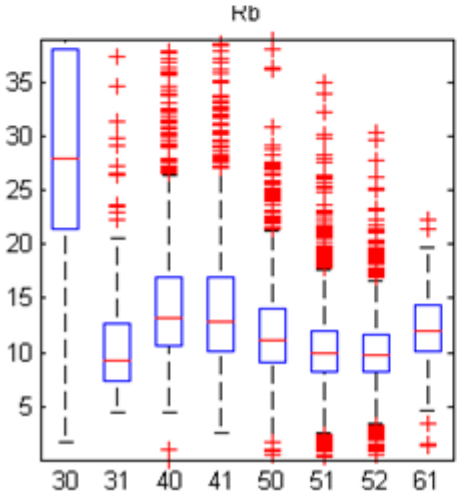

Ni
Frecuencia
3000
2500
2000
1500
1000
500
0
50 100 150 200
30
31
40
41
50
51
52
61

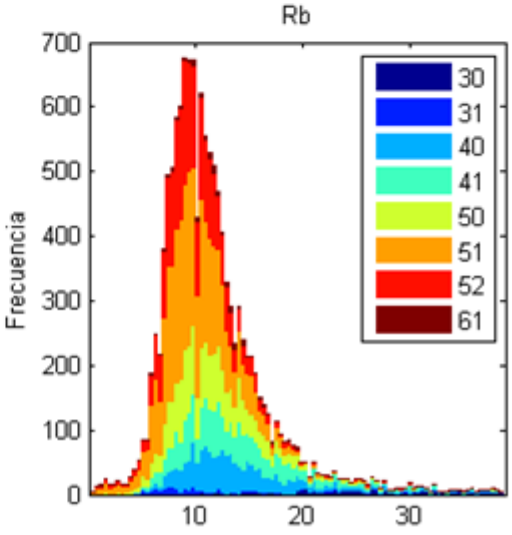

1600
1400
1200
1000
800
600
400
200
30 31 40 41 50 51 52 61

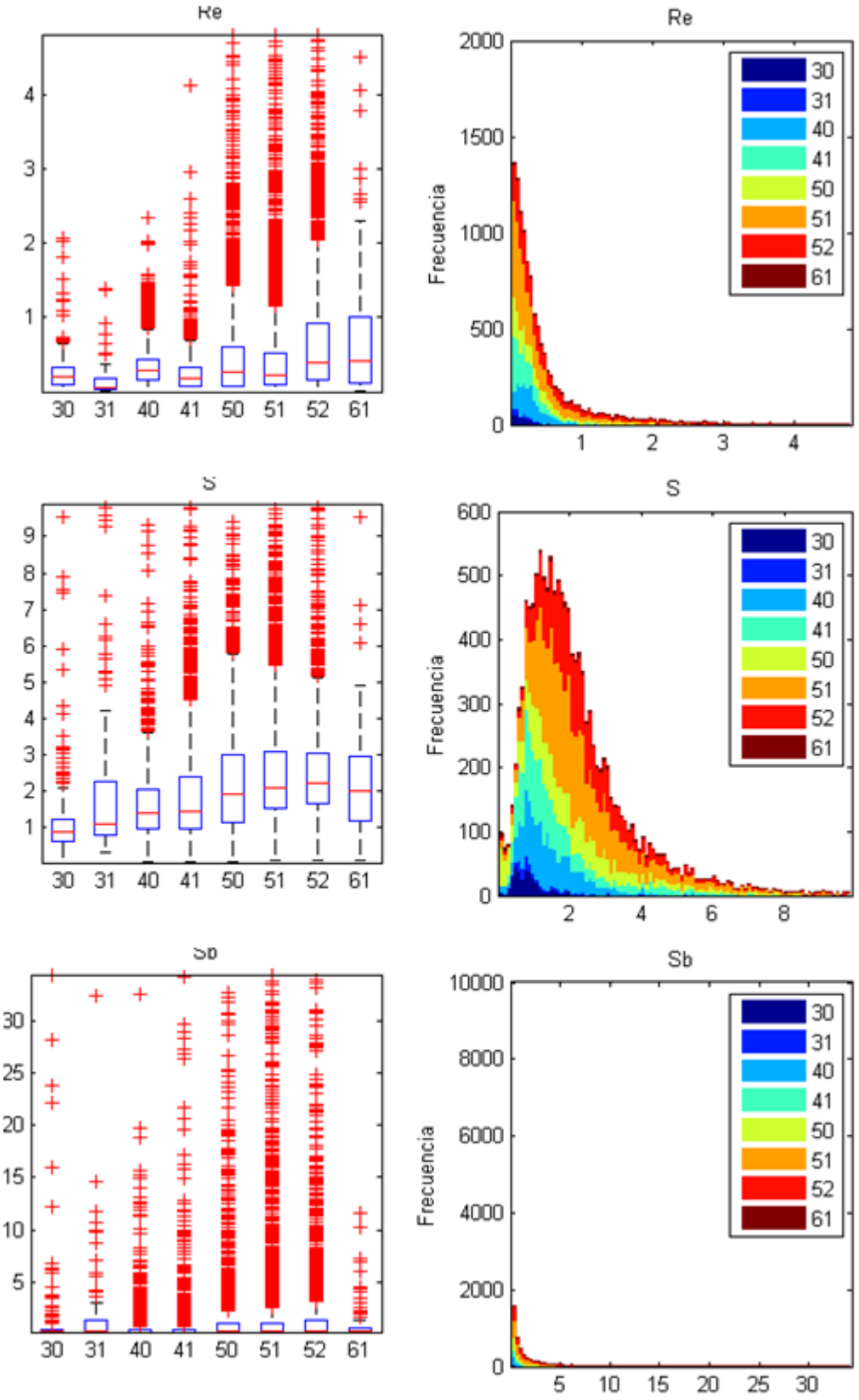

P
Frecuencia
1400
1200
1000
800
600
400
200
0
500 1000 1500
30
31
40
41
50
51
52
61

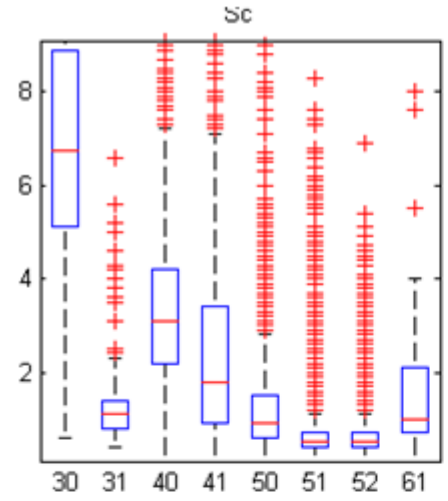

Pb
600
500
400
300
200
100
30 31 40 41 50 51 52 61

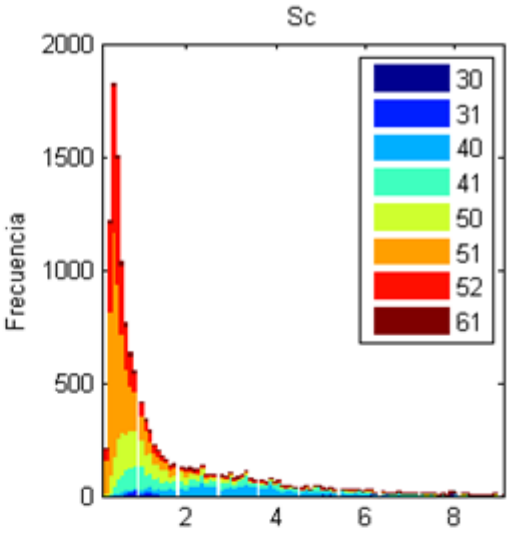

Pb
Frecuencia
5000
4000
3000
2000
1000
0
100 200 300 400 500 600
30
31
40
41
50
51
52
61

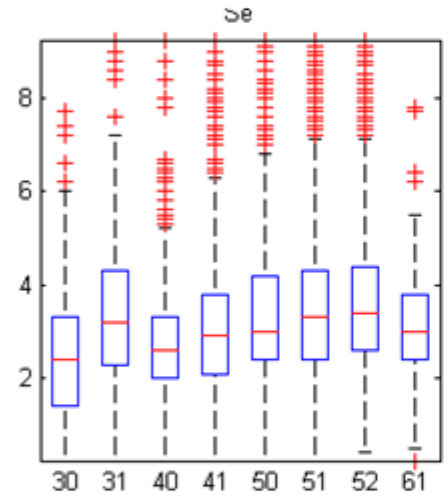

Rb
35
30
25
20
15
10
5
30 31 40 41 50 51 52 61

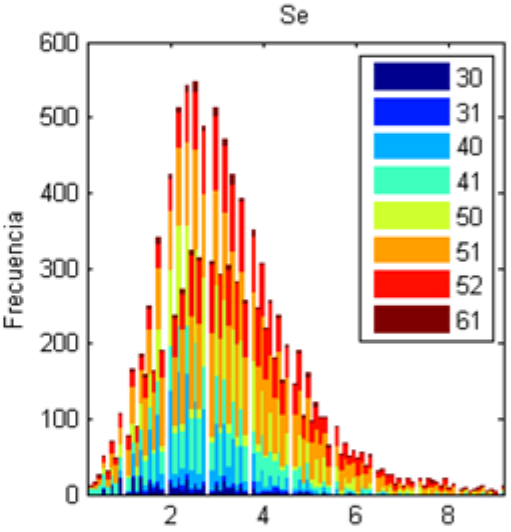

Rb
Frecuencia
700
600
500
400
300
200
100
0
10 20 30
30
31
40
41
50
51
52
61

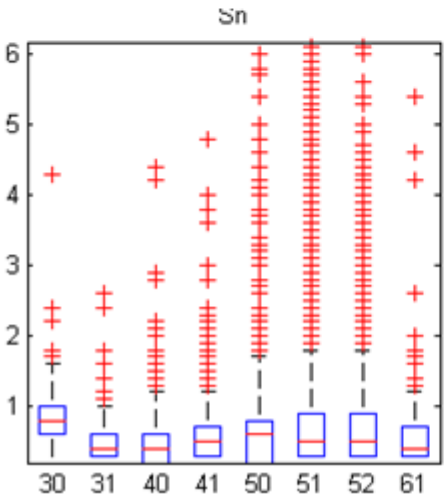

Re
Re
Frecuencia
30
31
40
41
50
51
52
61
S
S
Frecuencia
Sb
Sb
Frecuencia

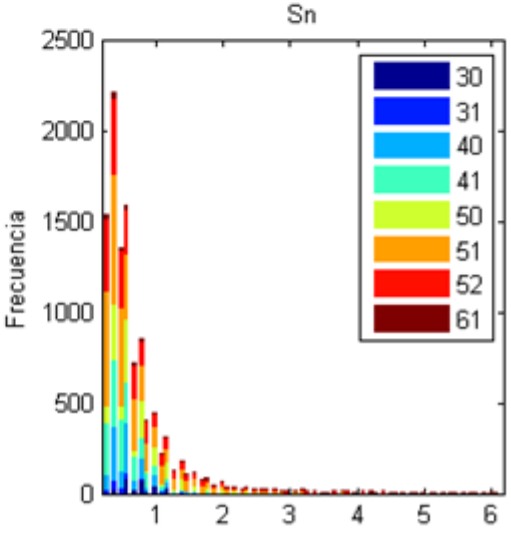

Sc
8
6
4
2
30 31 40 41 50 51 52 61

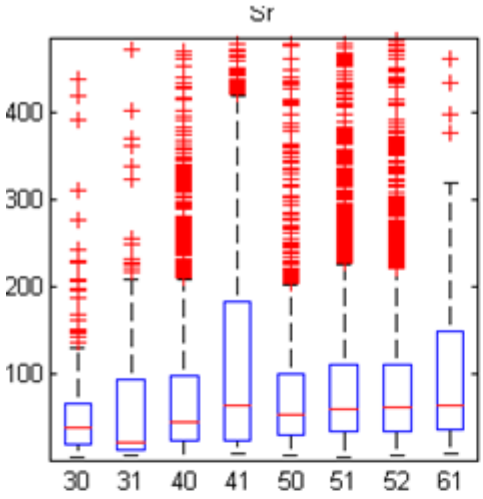

Sc
Frecuencia
2000
1500
1000
500
0
2 4 6 8
30
31
40
41
50
51
52
61

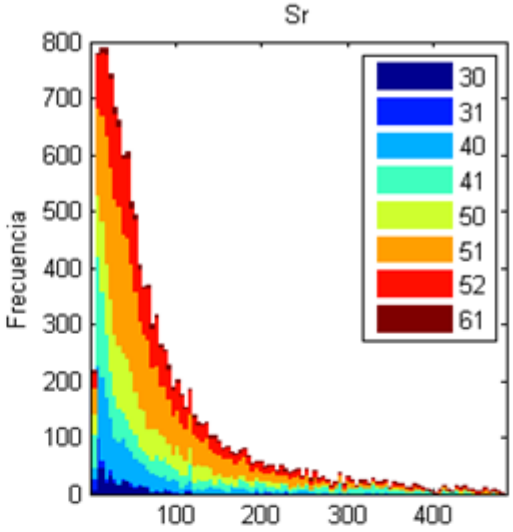

Se
8
6
4
2
30 31 40 41 50 51 52 61

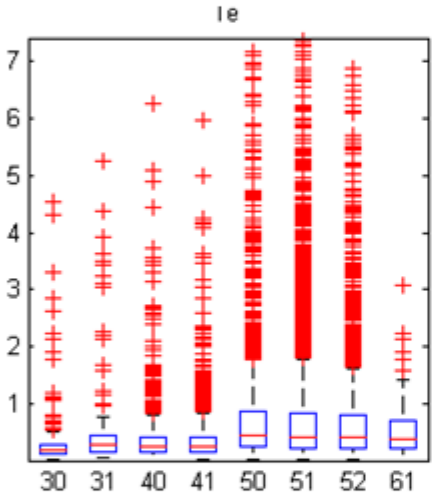

Se
Frecuencia
600
500
400
300
200
100
0
2 4 6 8
30
31
40
41
50
51
52
61

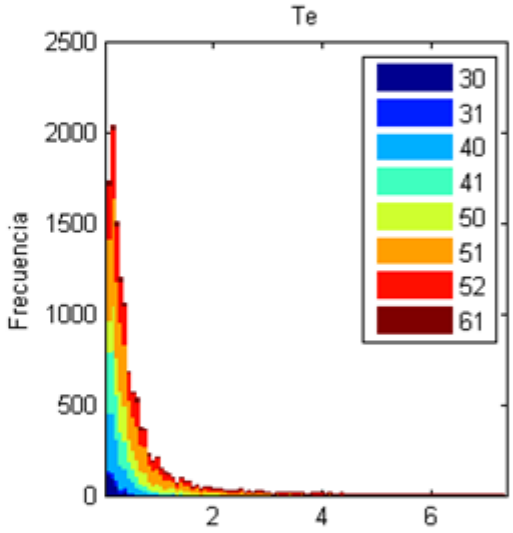

Sn
6
5
4
3
2
1
30 31 40 41 50 51 52 61

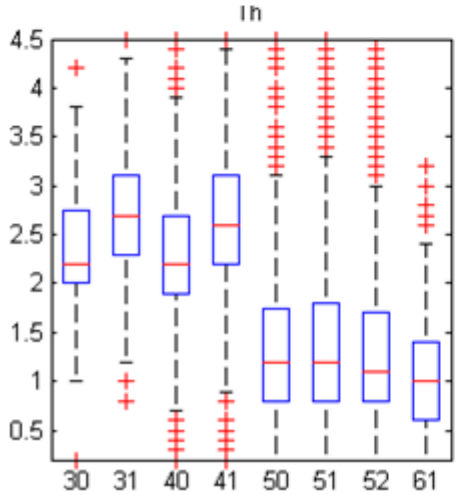

Sn
Frecuencia
2500
2000
1500
1000
500
0
1 2 3 4 5 6
30
31
40
41
50
51
52
61

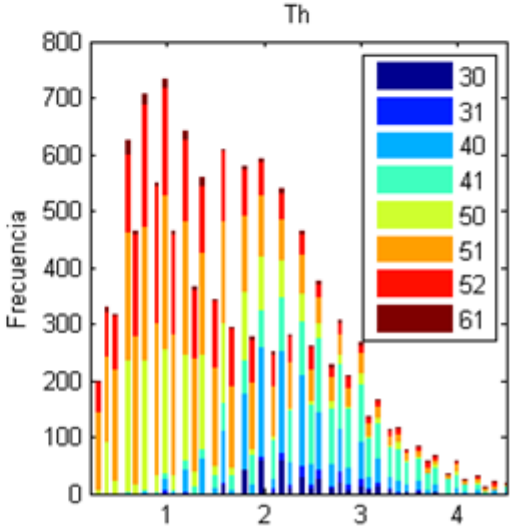

Sr
400
300
200
100
30
31
40
41
50
51
52
61

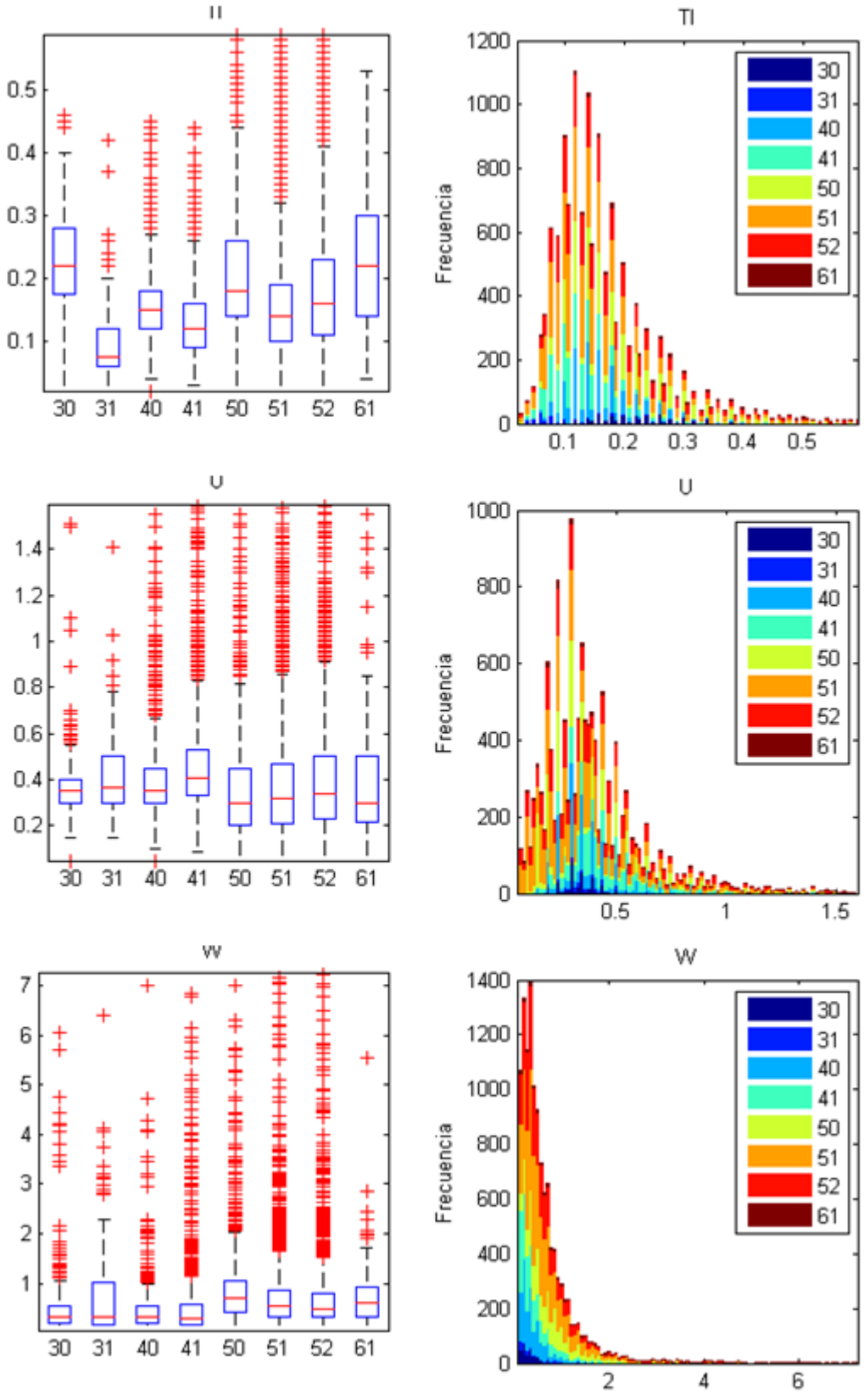

Sr
Frecuencia
800
700
600
500
400
300
200
100
0
100
200
300
400
30
31
40
41
50
51
52
61

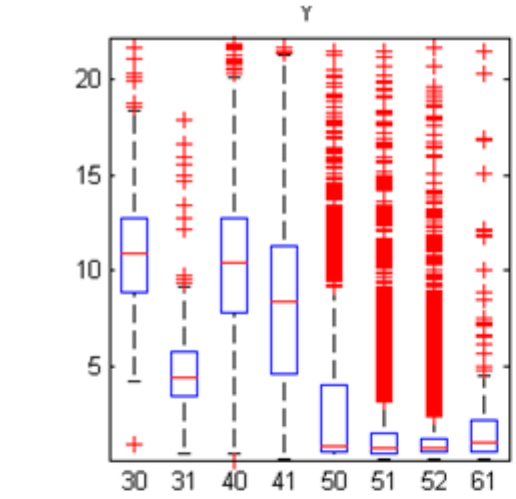

Ie
7
6
5
4
3
2
1
30
31
40
41
50
51
52
61

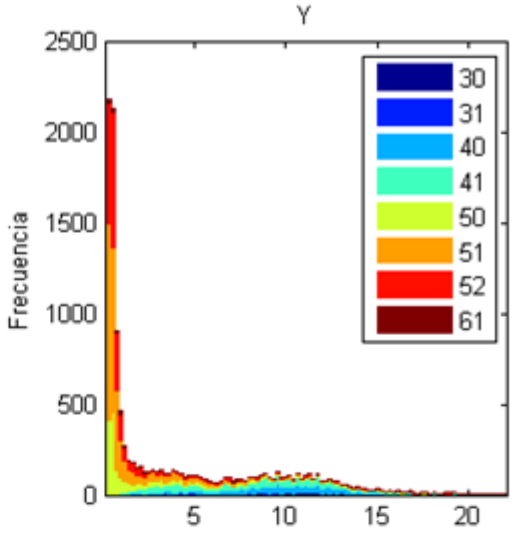

Te
Frecuencia
2500
2000
1500
1000
500
0
2
4
6
30
31
40
41
50
51
52
61

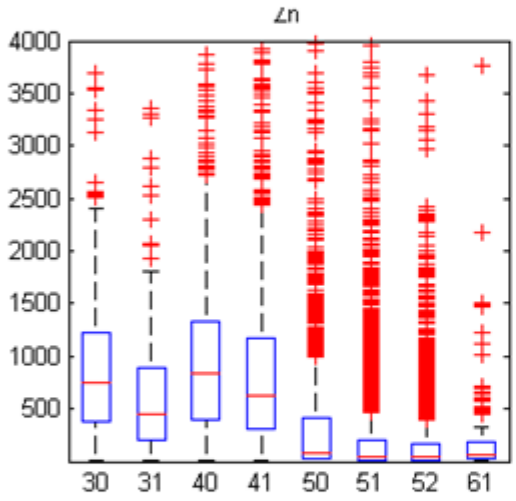

Ih
4.5
4
3.5
3
2.5
2
1.5
1
0.5
30
31
40
41
50
51
52
61

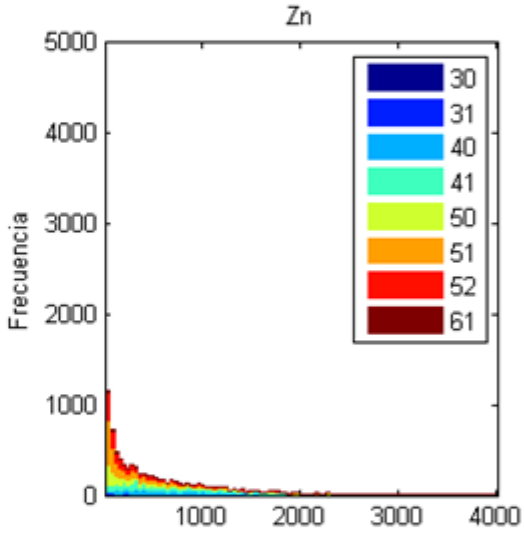

Th
Frecuencia
800
700
600
500
400
300
200
100
0
1
2
3
4
30
31
40
41
50
51
52
61

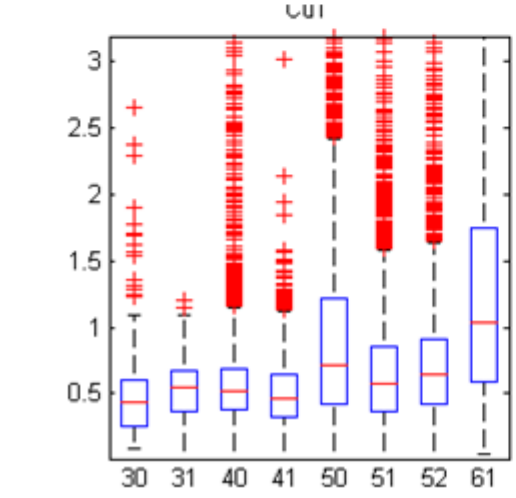

TI
Frecuencia
30
31
40
41
50
51
52
61
U
W

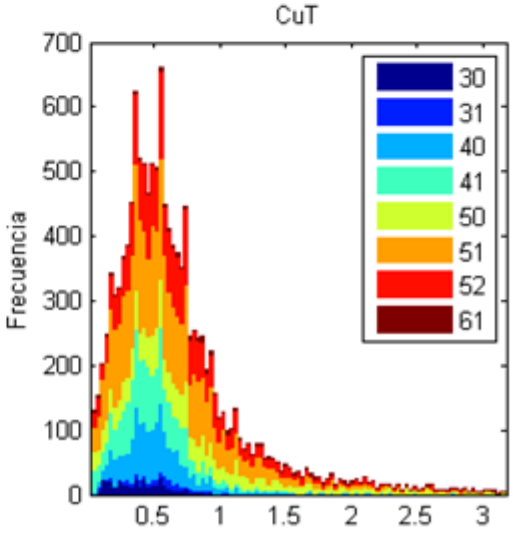

Y
20
15
10
5
30 31 40 41 50 51 52 61

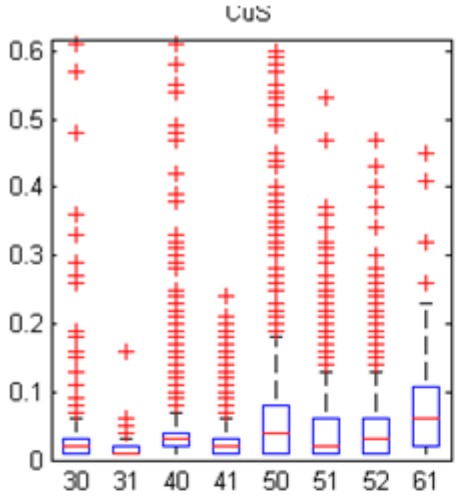

Y
2500
2000
1500
1000
500
0
Frecuencia
5 10 15 20
30
31
40
41
50
51
52
61

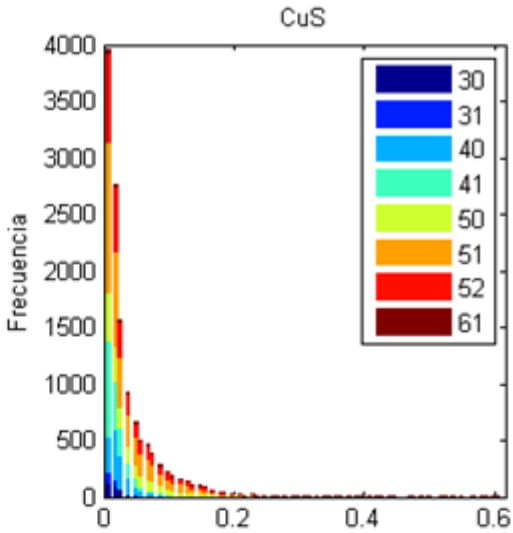

Zn
4000
3500
3000
2500
2000
1500
1000
500
30 31 40 41 50 51 52 61

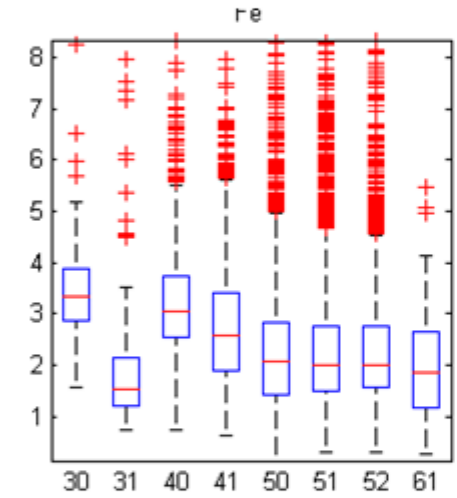

Zn
5000
4000
3000
2000
1000
0
Frecuencia
1000 2000 3000 4000
30
31
40
41
50
51
52
61

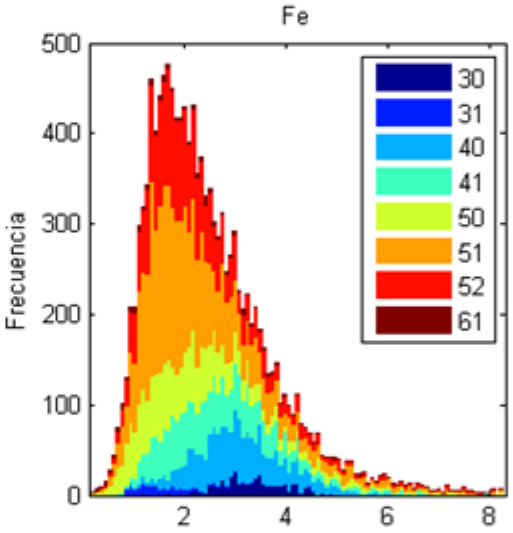

CuT
3
2.5
2
1.5
1
0.5
30 31 40 41 50 51 52 61

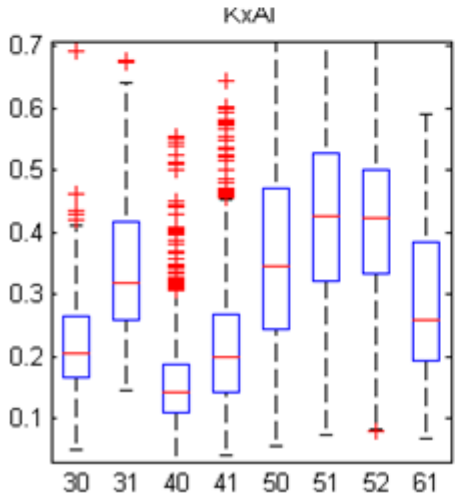

CuT
700
600
500
400
300
200
100
0
Frecuencia
0.5 1 1.5 2 2.5 3
30
31
40
41
50
51
52
61

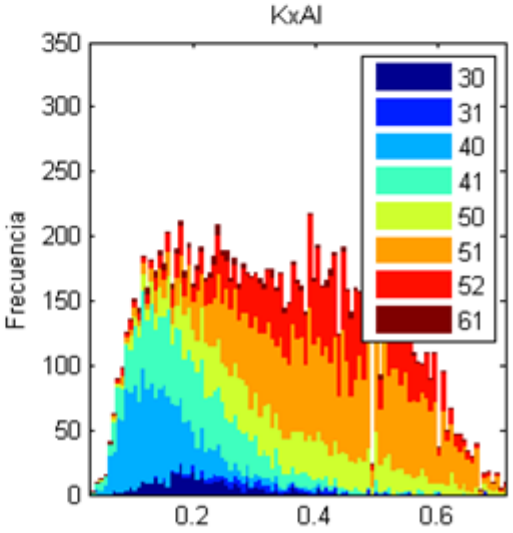

CuS
0.6
0.5
0.4
0.3
0.2
0.1
0
30 31 40 41 50 51 52 61

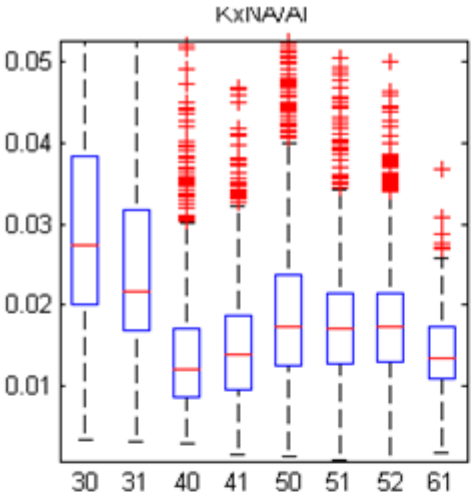

CuS
Frecuencia
4000
3500
3000
2500
2000
1500
1000
500
0
0 0.2 0.4 0.6
30
31
40
41
50
51
52
61

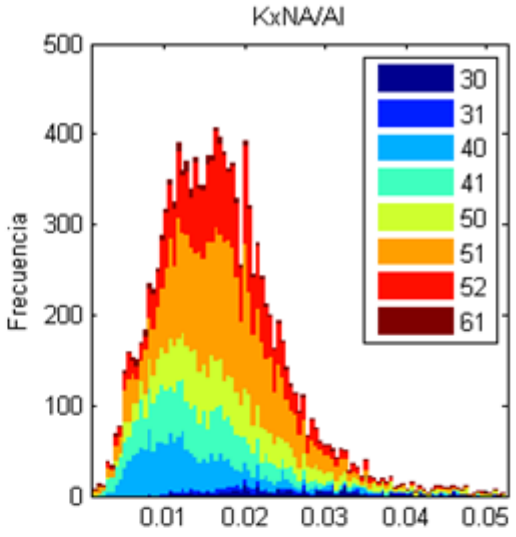

Fe
8
7
6
5
4
3
2
1
30 31 40 41 50 51 52 61

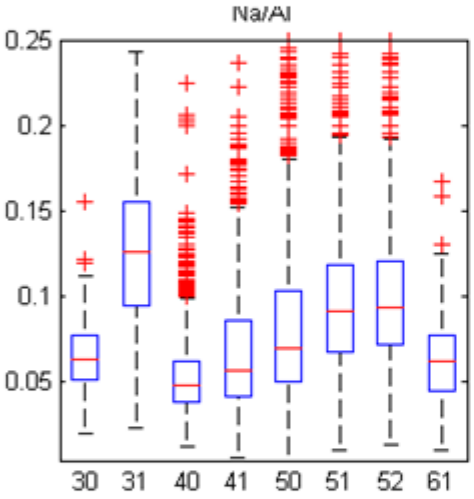

Fe
Frecuencia
500
400
300
200
100
0
2 4 6 8
30
31
40
41
50
51
52
61

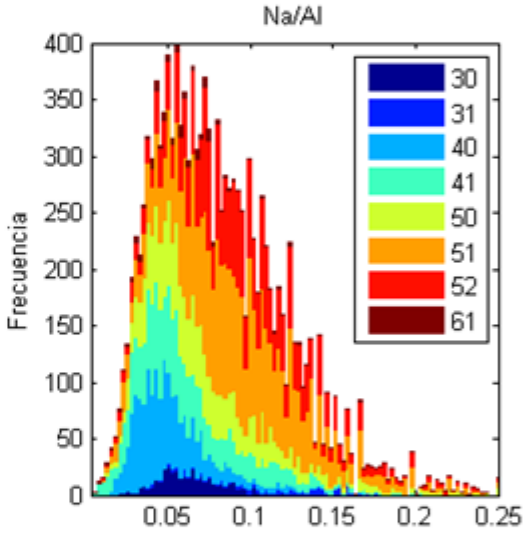

KxAl
0.7
0.6
0.5
0.4
0.3
0.2
0.1
30 31 40 41 50 51 52 61

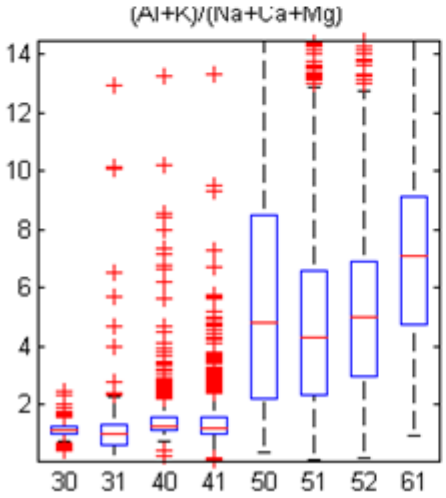

KxAl
Frecuencia
350
300
250
200
150
100
50
0
0.2 0.4 0.6
30
31
40
41
50
51
52
61

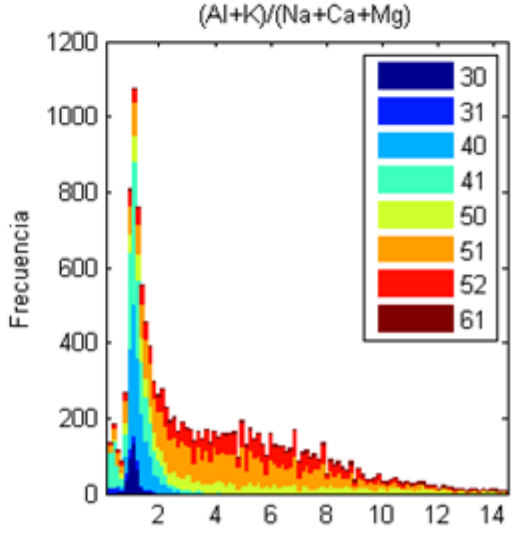

KxNA/Al
0.05
0.04
0.03
0.02
0.01
30 31 40 41 50 51 52 61

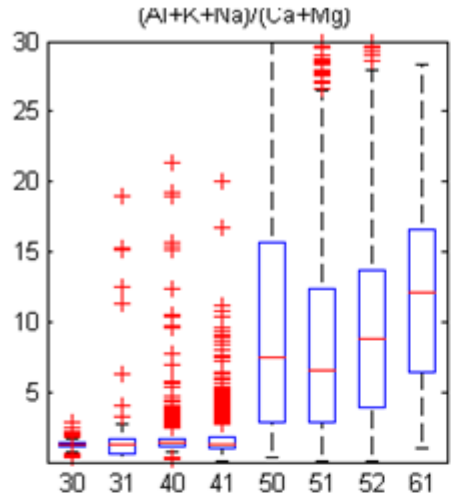

KxNA/Al
Frecuencia
500
400
300
200
100
0
0.01 0.02 0.03 0.04 0.05
30
31
40
41
50
51
52
61

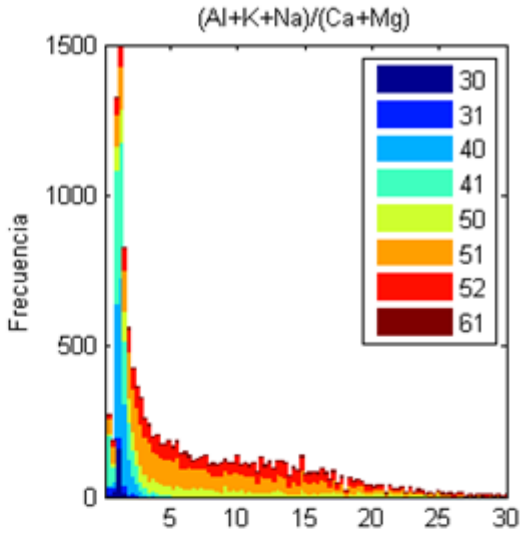

Na/Al
0.25
0.2
0.15
0.1
0.05
30 31 40 41 50 51 52 61

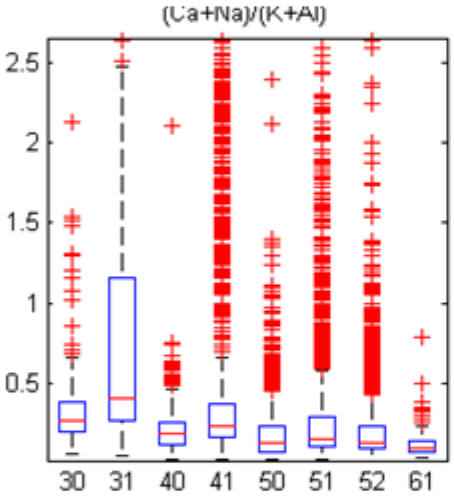

Na/Al
Frecuencia
400
350
300
250
200
150
100
50
0
0.05 0.1 0.15 0.2 0.25
30
31
40
41
50
51
52
61

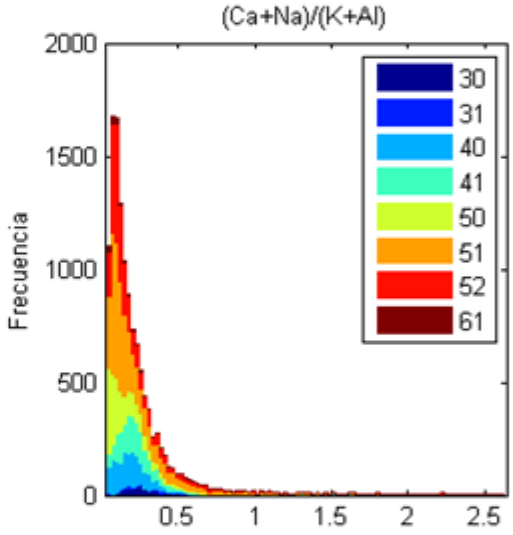

(Al+K)/(Na+Ca+Mg)
14
12
10
8
6
4
2
30 31 40 41 50 51 52 61

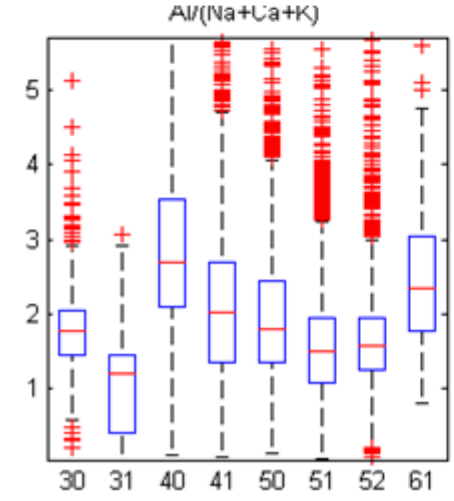

(Al+K)/(Na+Ca+Mg)
Frecuencia
1200
1000
800
600
400
200
0
2 4 6 8 10 12 14
30
31
40
41
50
51
52
61

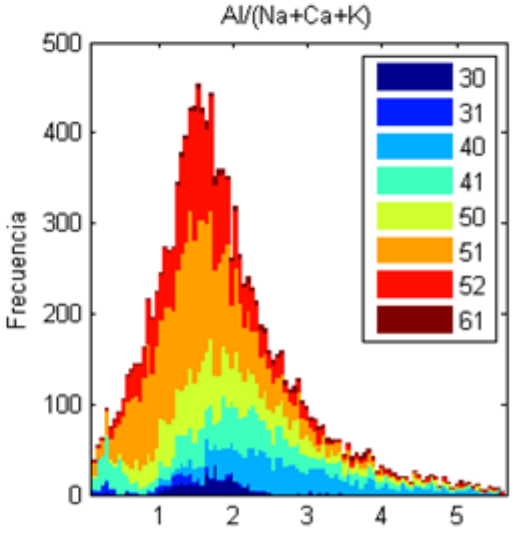

(Al+K+Na)/(Ca+Mg)
30
25
20
15
10
5
30 31 40 41 50 51 52 61

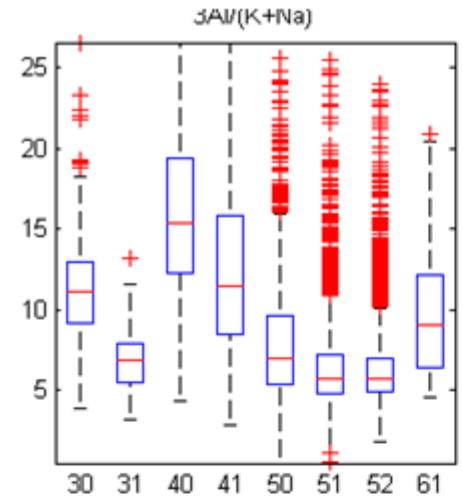

(Al+K+Na)/(Ca+Mg)
Frecuencia
1500
1000
500
0
5 10 15 20 25 30
30
31
40
41
50
51
52
61

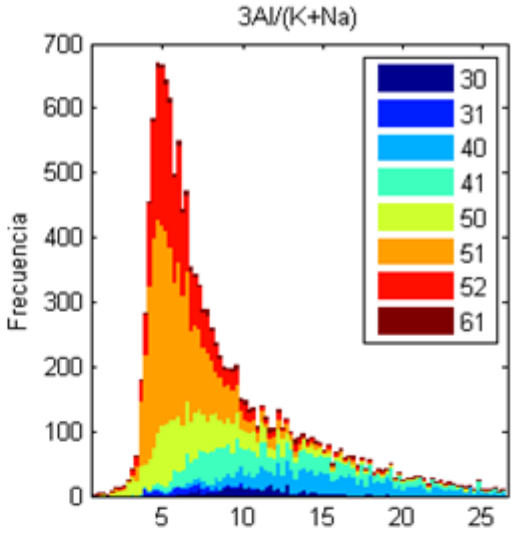

(Ca+Na)/(K+Al)
2.5
2
1.5
1
0.5
30 31 40 41 50 51 52 61

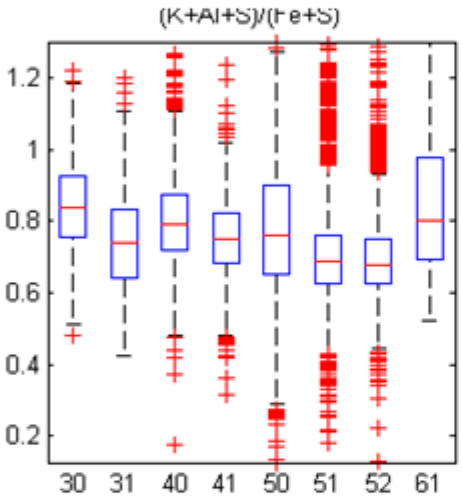

(Ca+Na)/(K+Al)
Frecuencia
2000
1500
1000
500
0
0.5 1 1.5 2 2.5
30
31
40
41
50
51
52
61

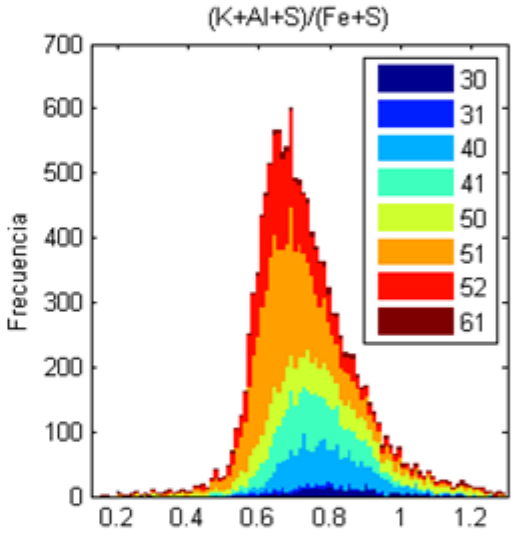

Al/(Na+Ca+K)
5
4
3
2
1
30 31 40 41 50 51 52 61

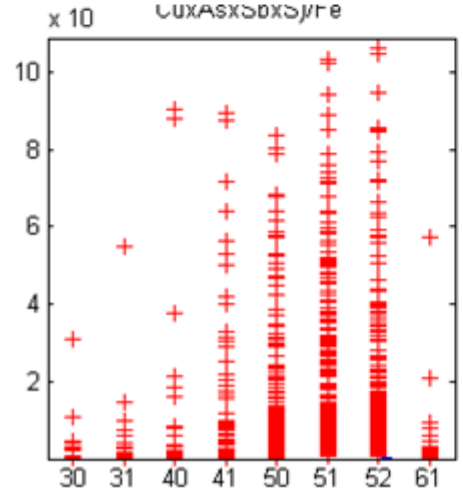

Al/(Na+Ca+K)
Frecuencia
500
400
300
200
100
0
1 2 3 4 5
30
31
40
41
50
51
52
61

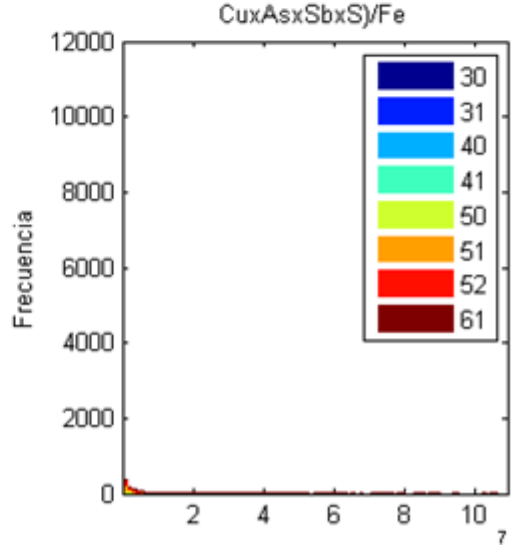

3Al/(K+Na)
25
20
15
10
5
30 31 40 41 50 51 52 61

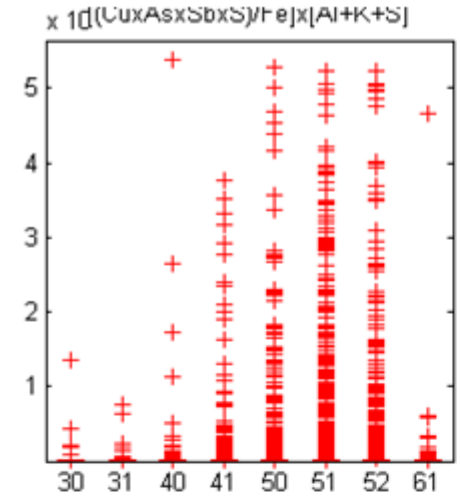

3Al/(K+Na)
Frecuencia
700
600
500
400
300
200
100
0
5 10 15 20 25
30
31
40
41
50
51
52
61

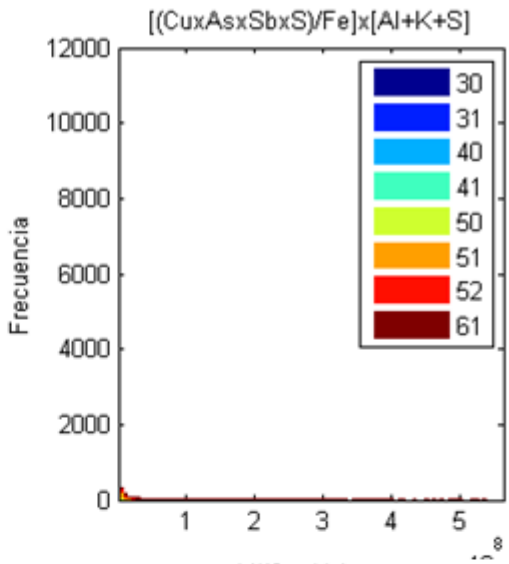

(K+Al+S)/(Fe+S)
1.2
1
0.8
0.6
0.4
0.2
30 31 40 41 50 51 52 61

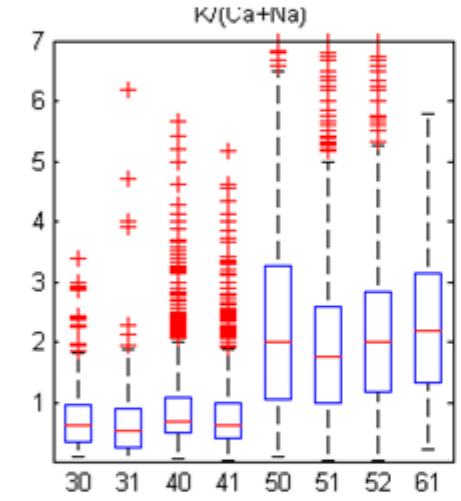

(K+Al+S)/(Fe+S)
Frecuencia
700
600
500
400
300
200
100
0
0.2 0.4 0.6 0.8 1 1.2
30
31
40
41
50
51
52
61

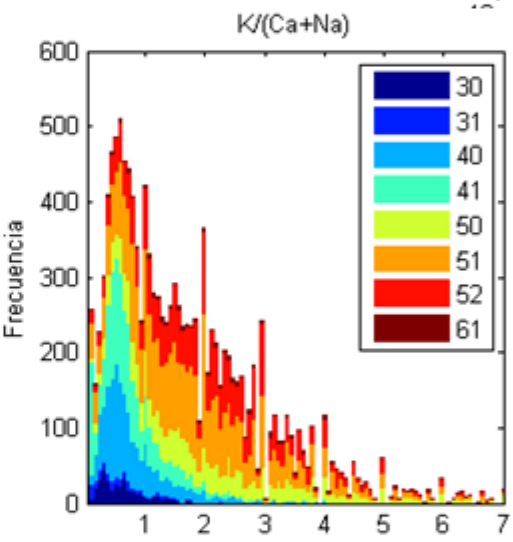

x 10
10
8
6
4
2
30 31 40 41 50 51 52 61

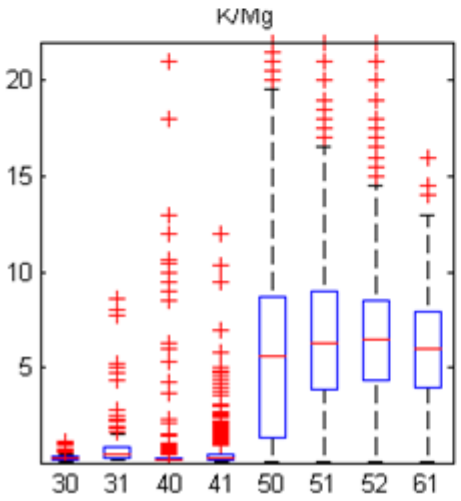

CuxAsxSbxS)/Fe
Frecuencia
12000
10000
8000
6000
4000
2000
0
2 4 6 8 10
7
30
31
40
41
50
51
52
61

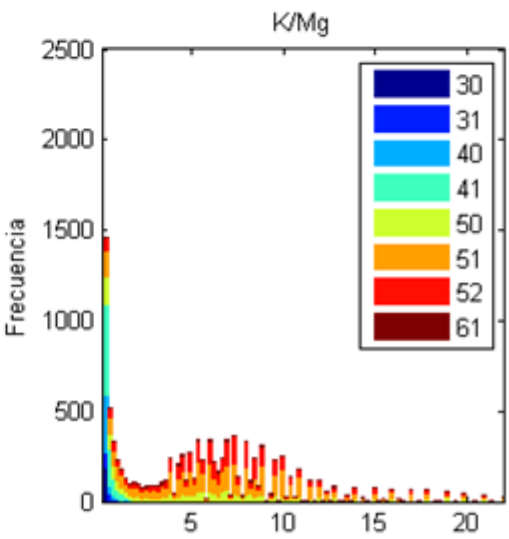

x 10
[(CuxAsxSbxS)/Fe]x[Al+K+S]
30
31
40
41
50
51
52
61

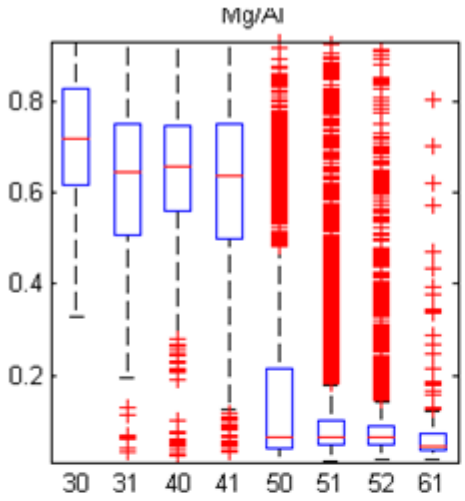

[(CuxAsxSbxS)/Fe]x[Al+K+S]
Frecuencia
30
31
40
41
50
51
52
61

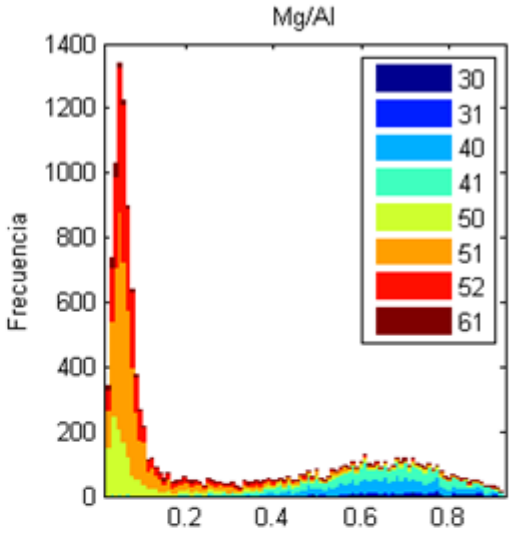

K/(Ca+Na)
30
31
40
41
50
51
52
61

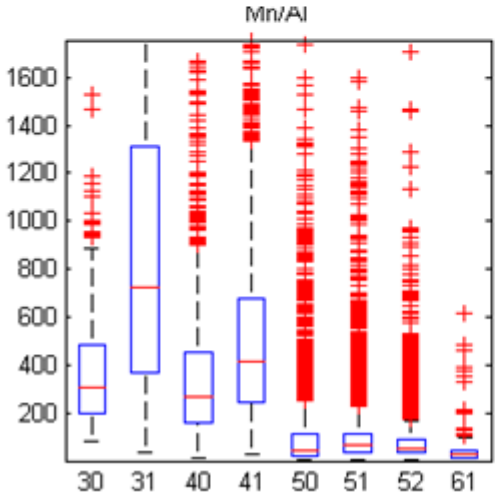

K/(Ca+Na)
Frecuencia
30
31
40
41
50
51
52
61

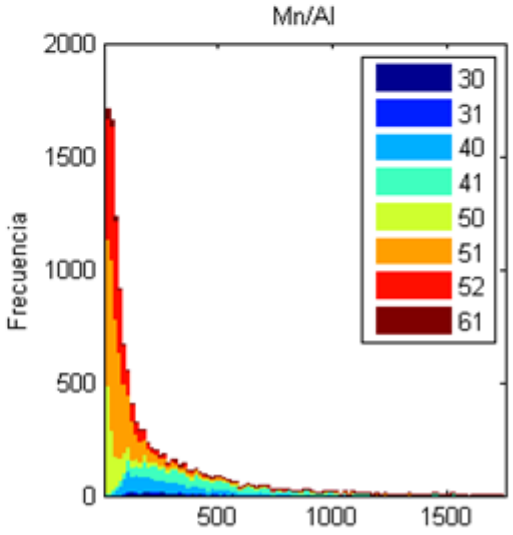

K/Mg
30
31
40
41
50
51
52
61


K/Mg
Frecuencia
30
31
40
41
50
51
52
61


Mg/Al
0.8
0.6
0.4
0.2
30 31 40 41 50 51 52 61


Mg/Al
Frecuencia
1400
1200
1000
800
600
400
200
0
0.2 0.4 0.6 0.8
30
31
40
41
50
51
52
61


Mn/Al
1600
1400
1200
1000
800
600
400
200
30 31 40 41 50 51 52 61


Mn/Al
Frecuencia
2000
1500
1000
500
0
500 1000 1500
30
31
40
41
50
51
52
61

Apéndice C

Anexos: métodos de análisis

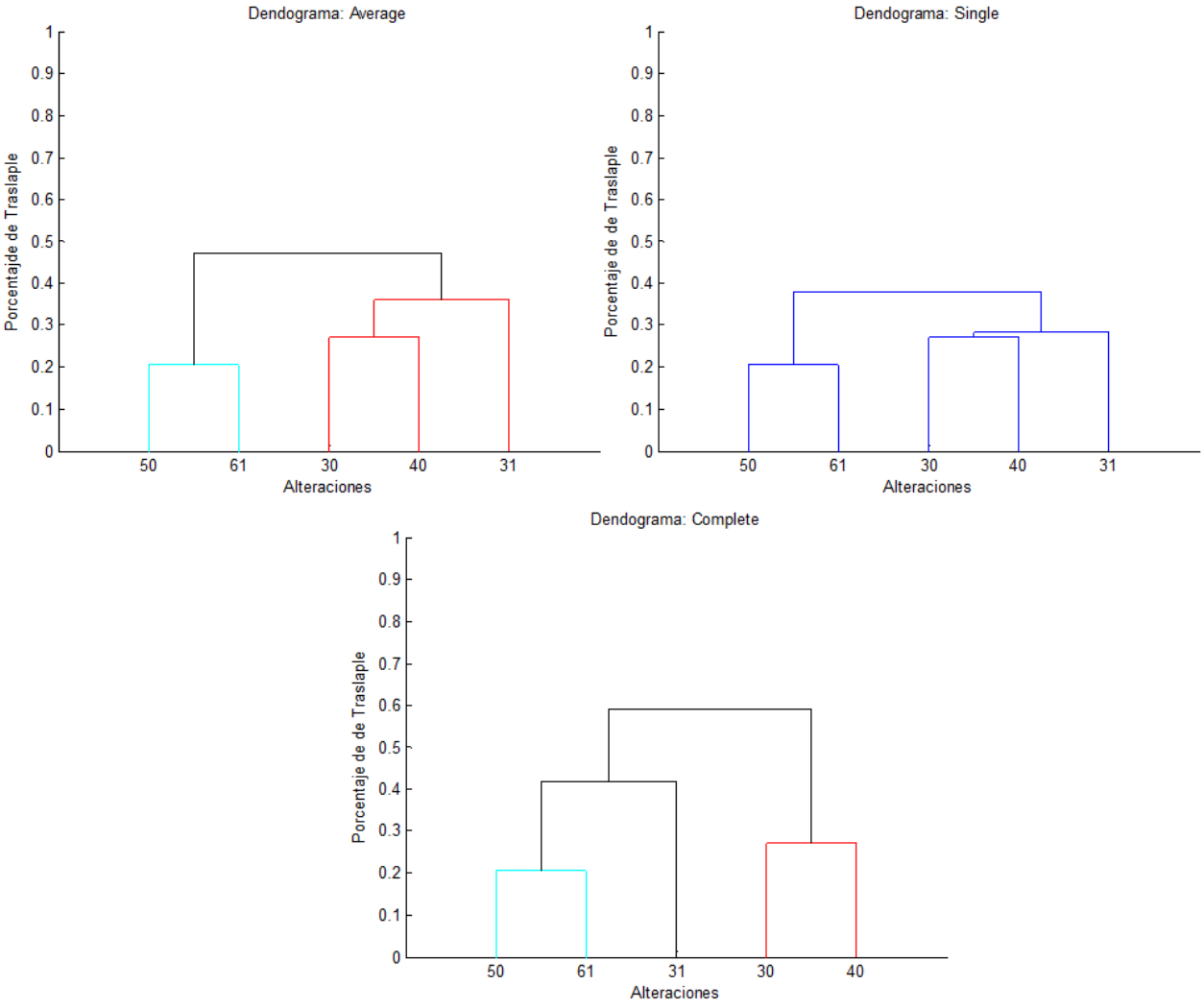
La sección «C» de los anexos incluye la información procesada para la creación de los modelos, en este se estudian las relaciones entre las variables utilizando los estadísticos desarrollados (traslape de población y dendrograma de categorías) y la distribución de las variables según las categorías unidas.

Traslapes de familias para cluster de categorías

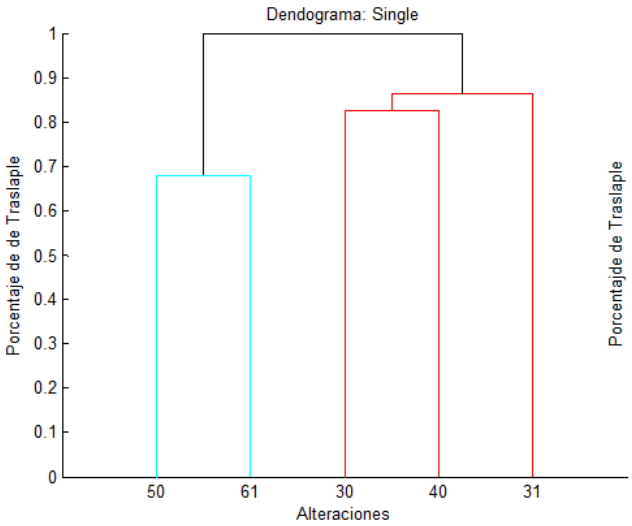
	Familia 30				Familia 31				Familia 40			
Elelementos	31	40	50	61	30	40	50	61	30	31	50	61
Ag	0.84	0.96	1.00	0.98	0.47	0.65	0.73	0.87	0.75	0.90	1.00	1.00
Al	0.00	0.54	0.00	0.00	0.00	0.24	0.73	0.82	0.41	0.06	0.00	0.46
As	1.00	0.94	0.42	0.64	1.00	1.00	0.83	0.89	0.53	1.00	0.70	0.83
Ba	0.52	0.34	0.15	0.15	0.70	1.00	0.50	0.50	0.60	1.00	0.67	0.67
Be	0.29	1.00	0.00	0.00	0.37	0.42	0.39	0.63	1.00	0.24	0.00	0.03
Bi	0.70	0.97	0.94	0.99	0.33	0.58	0.93	0.90	0.65	0.82	0.98	1.00
Ca	1.00	0.28	0.00	0.00	1.00	0.34	0.04	0.03	0.52	0.94	0.18	0.14
Cd	1.00	0.96	0.96	1.00	1.00	0.97	0.67	1.00	0.62	0.91	0.59	1.00
Ce	1.00	0.98	0.40	0.50	1.00	0.94	0.40	0.49	0.85	0.92	0.33	0.41
Co	0.00	0.82	0.71	1.00	0.00	0.43	0.60	1.00	0.66	0.20	0.91	1.00
Cr	0.84	1.00	1.00	0.92	0.23	0.44	0.70	0.21	1.00	0.85	1.00	1.00
Cs	0.24	0.67	0.10	0.44	0.29	0.68	0.84	0.90	0.68	0.56	0.42	0.76
Cu	0.76	0.83	0.79	0.48	0.84	1.00	1.00	0.69	0.82	1.00	0.97	0.66
In	0.91	0.89	0.84	0.65	0.39	1.00	0.97	0.89	0.39	1.00	0.98	0.90
K	0.08	0.12	0.00	0.00	0.14	0.79	1.00	1.00	0.25	0.92	0.50	0.75
Mg	0.00	0.55	0.00	0.00	0.00	0.37	0.00	0.00	0.29	0.09	0.00	0.00
Mn	0.65	0.85	0.00	0.00	0.73	0.92	0.00	0.00	0.83	0.81	0.00	0.00
Mo	0.64	0.87	0.91	0.95	0.86	0.89	0.96	0.80	0.97	0.74	1.00	0.92
Na	0.14	0.17	0.00	0.00	0.29	1.00	0.18	0.35	0.33	1.00	0.17	0.33
Ni	0.55	0.38	0.20	0.13	0.90	1.00	0.43	0.32	0.88	1.00	0.58	0.43
P	0.00	1.00	0.00	0.00	0.00	0.72	0.00	1.00	1.00	0.25	0.00	0.41
Pb	1.00	0.99	0.17	0.50	1.00	1.00	0.14	0.41	0.97	1.00	0.16	0.47
Rb	0.00	0.07	0.00	0.00	0.00	0.71	1.00	0.76	0.16	0.63	0.37	0.64
Re	0.50	1.00	0.99	1.00	0.77	0.79	0.75	0.97	1.00	0.42	0.98	1.00
S	0.85	0.75	0.32	0.84	0.30	1.00	1.00	1.00	0.41	1.00	0.76	1.00
Sb	1.00	0.86	1.00	1.00	1.00	1.00	1.00	0.45	0.99	1.00	1.00	1.00
Sc	0.00	0.09	0.00	0.00	0.00	0.80	0.20	1.00	0.14	0.21	0.00	0.45
Se	0.72	0.77	0.65	0.69	0.65	0.76	0.94	1.00	0.87	0.94	0.87	0.91
Sn	0.50	0.50	1.00	0.75	0.80	1.00	1.00	1.00	0.67	1.00	0.83	1.00
Sr	1.00	0.98	0.85	0.83	1.00	0.97	1.00	0.91	0.32	0.80	1.00	1.00
Te	0.86	0.96	0.79	0.72	0.53	0.91	0.96	0.91	0.61	0.93	0.89	0.85
Th	0.67	1.00	0.25	0.33	0.62	0.82	0.00	0.00	1.00	0.73	0.21	0.27
Tl	0.00	0.26	0.71	1.00	0.00	0.64	0.64	0.36	0.34	0.66	1.00	0.72
U	1.00	1.00	1.00	1.00	1.00	1.00	0.91	1.00	1.00	1.00	1.00	1.00
W	1.00	1.00	0.82	0.91	1.00	1.00	1.00	1.00	1.00	1.00	0.81	0.90
Y	0.00	0.95	0.00	0.00	0.00	0.78	0.05	0.58	0.55	0.37	0.00	0.17
Zn	1.00	1.00	0.12	0.16	1.00	1.00	0.18	0.21	1.00	1.00	0.12	0.16
CuT	0.78	0.82	0.81	0.54	0.82	0.98	1.00	0.75	0.85	0.95	0.98	0.71
CuS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Fe	0.14	0.89	0.49	0.35	0.13	0.57	0.86	1.00	0.64	0.47	0.72	0.62
KxAl	0.41	0.84	0.17	0.88	0.27	0.16	0.85	0.84	0.80	0.24	0.01	0.68
KxNA/Al	0.73	0.18	0.27	0.13	0.92	0.30	0.42	0.24	0.36	0.48	0.75	1.00
Na/Al	0.37	1.00	0.72	1.00	0.17	0.21	0.69	0.18	1.00	0.33	0.58	1.00
(Al+K)/ (Na+Ca+Mg)	1.00	0.94	0.00	0.00	1.00	0.48	0.00	0.00	0.40	0.66	0.00	0.00
(Ca+Na)/(K+Al)	0.89	0.68	0.69	0.01	0.16	0.10	0.10	0.00	0.80	0.68	1.00	0.21
Al/ (Na+Ca+K)	0.45	0.94	0.98	0.66	0.31	0.27	0.57	0.08	0.39	0.16	0.39	1.00
3Al/(K+Na)	0.08	0.97	0.02	0.97	0.13	0.08	0.90	0.74	0.52	0.03	0.00	0.50
(K+Al+S)/ (Fe+S)	0.58	0.61	0.32	1.00	0.60	0.79	0.73	0.82	0.77	0.97	0.64	1.00
CuxAsxSbxS)/ Fe	0.97	0.97	0.64	0.80	0.03	1.00	0.99	1.00	0.28	1.00	0.91	0.95
K/(Ca+Na)	1.00	0.98	0.49	0.25	1.00	0.82	0.43	0.23	0.85	0.88	0.58	0.37
K/Mg	0.52	1.00	0.00	0.00	0.11	0.24	0.00	0.00	1.00	0.60	0.00	0.00
Mg/Al	0.71	0.75	0.00	0.00	0.56	0.85	0.00	0.00	0.67	0.96	0.00	0.00
Mn/Al	0.81	1.00	0.00	0.00	0.27	0.40	0.00	0.00	1.00	0.85	0.00	0.00

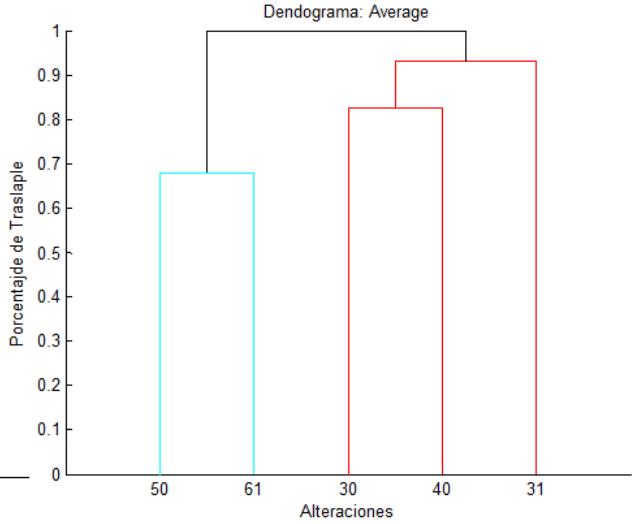
	Familia 50				Familia 61			
Elelementos	30	31	40	61	30	31	40	50
Ag	1.00	0.85	1.00	0.95	0.58	0.92	1.00	0.85
Al	0.00	0.68	0.00	0.51	0.00	0.35	0.76	0.23
As	0.03	0.23	0.10	0.66	0.07	0.37	0.17	0.97
Ba	0.27	0.67	0.67	1.00	0.26	0.64	0.64	1.00
Be	0.00	0.40	0.00	1.00	0.00	0.52	0.04	1.00
Bi	0.43	0.89	0.66	0.97	0.46	0.87	1.00	0.98
Ca	0.00	0.38	0.56	1.00	0.00	0.29	0.50	1.00
Cd	0.95	0.96	0.92	0.96	1.00	1.00	1.00	0.41
Ce	0.38	0.43	0.37	0.96	0.45	0.50	0.44	0.91
Co	0.58	0.29	0.92	1.00	1.00	1.00	1.00	1.00
Cr	1.00	0.88	1.00	1.00	0.92	0.76	1.00	1.00
Cs	0.15	0.98	0.60	0.86	0.47	0.78	0.80	0.64
Cu	0.56	1.00	0.69	1.00	0.44	0.57	0.60	1.00
In	0.33	0.89	0.89	0.92	0.20	0.65	0.64	0.73
K	0.00	1.00	0.73	0.82	0.00	1.00	0.92	0.69
Mg	0.00	0.00	0.00	0.86	0.00	0.00	0.00	0.60
Mn	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Mo	0.92	0.72	1.00	0.88	0.59	0.37	0.51	0.54
Na	0.00	0.33	0.33	1.00	0.00	0.50	0.50	1.00
Ni	0.71	0.91	0.89	1.00	0.64	0.92	0.89	1.00
P	0.00	0.00	0.00	0.96	0.00	1.00	0.52	0.30
Pb	0.76	0.79	0.72	0.93	0.93	0.94	0.91	0.39
Rb	0.00	1.00	0.67	0.74	0.00	0.98	0.94	0.60
Re	0.34	0.17	0.41	1.00	1.00	0.15	1.00	1.00
S	0.14	1.00	0.63	0.81	0.36	1.00	1.00	0.77
Sb	1.00	1.00	1.00	0.41	1.00	0.97	1.00	0.97
Sc	0.00	0.33	0.00	0.67	0.00	1.00	0.81	0.19
Se	0.59	0.93	0.69	0.87	0.69	1.00	0.80	0.96
Sn	1.00	1.00	0.50	0.70	0.75	1.00	1.00	0.88
Sr	0.45	1.00	1.00	0.99	0.29	0.79	1.00	0.65
Te	0.20	0.39	0.36	1.00	0.26	0.53	0.48	1.00
Th	0.20	0.00	0.20	1.00	0.25	0.00	0.25	1.00
Tl	0.59	0.41	1.00	0.82	1.00	0.15	0.29	0.54
U	1.00	0.78	1.00	1.00	1.00	1.00	1.00	1.00
W	0.54	1.00	0.54	1.00	0.52	1.00	0.52	1.00
Y	0.00	0.07	0.00	0.99	0.00	0.49	0.31	0.55
Zn	0.40	0.77	0.45	0.98	0.48	0.81	0.52	0.88
CuT	0.53	1.00	0.63	0.83	0.17	0.22	0.22	0.39
CuS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Fe	0.37	0.74	0.75	0.89	0.25	1.00	0.61	0.84
KxAl	0.10	0.74	0.01	0.60	0.50	0.74	0.41	0.60
KxNA/Al	0.58	0.71	0.80	0.69	0.31	0.46	1.00	0.79
Na/Al	0.39	0.81	0.43	0.40	1.00	0.34	1.00	0.64
(Al+K)/ (Na+Ca+Mg)	0.00	0.00	0.00	0.72	0.00	0.00	0.00	0.65
(Ca+Na)/(K+Al)	0.75	0.63	1.00	0.26	0.04	0.00	0.59	0.81
Al/ (Na+Ca+K)	0.72	0.61	0.68	0.47	0.32	0.06	1.00	0.32
3Al/(K+Na)	0.03	0.96	0.00	0.68	0.72	0.34	0.70	0.29
(K+Al+S)/ (Fe+S)	0.44	0.97	0.70	0.73	1.00	0.46	1.00	0.31
CuxAsxSbxS)/ Fe	0.00	0.20	0.02	0.52	0.01	0.38	0.04	1.00
K/(Ca+Na)	0.18	0.20	0.25	0.91	0.09	0.10	0.15	0.83
K/Mg	0.00	0.00	0.00	0.82	0.00	0.00	0.00	0.92
Mg/Al	0.00	0.00	0.00	0.71	0.00	0.00	0.00	0.90
Mn/Al	0.00	0.00	0.01	0.33	0.00	0.00	0.00	0.74

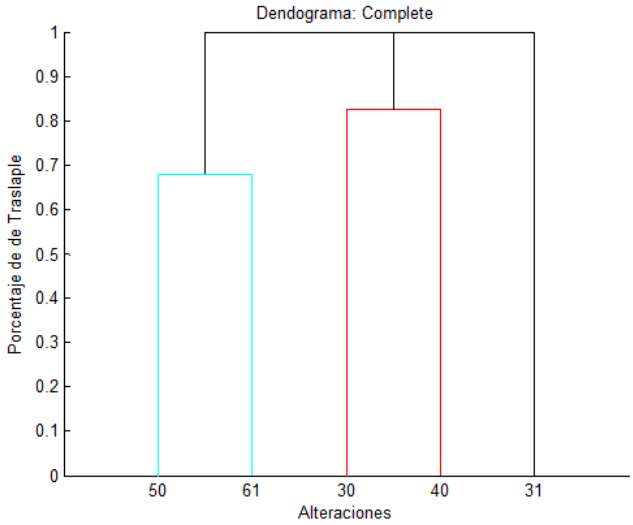
Todas las variables



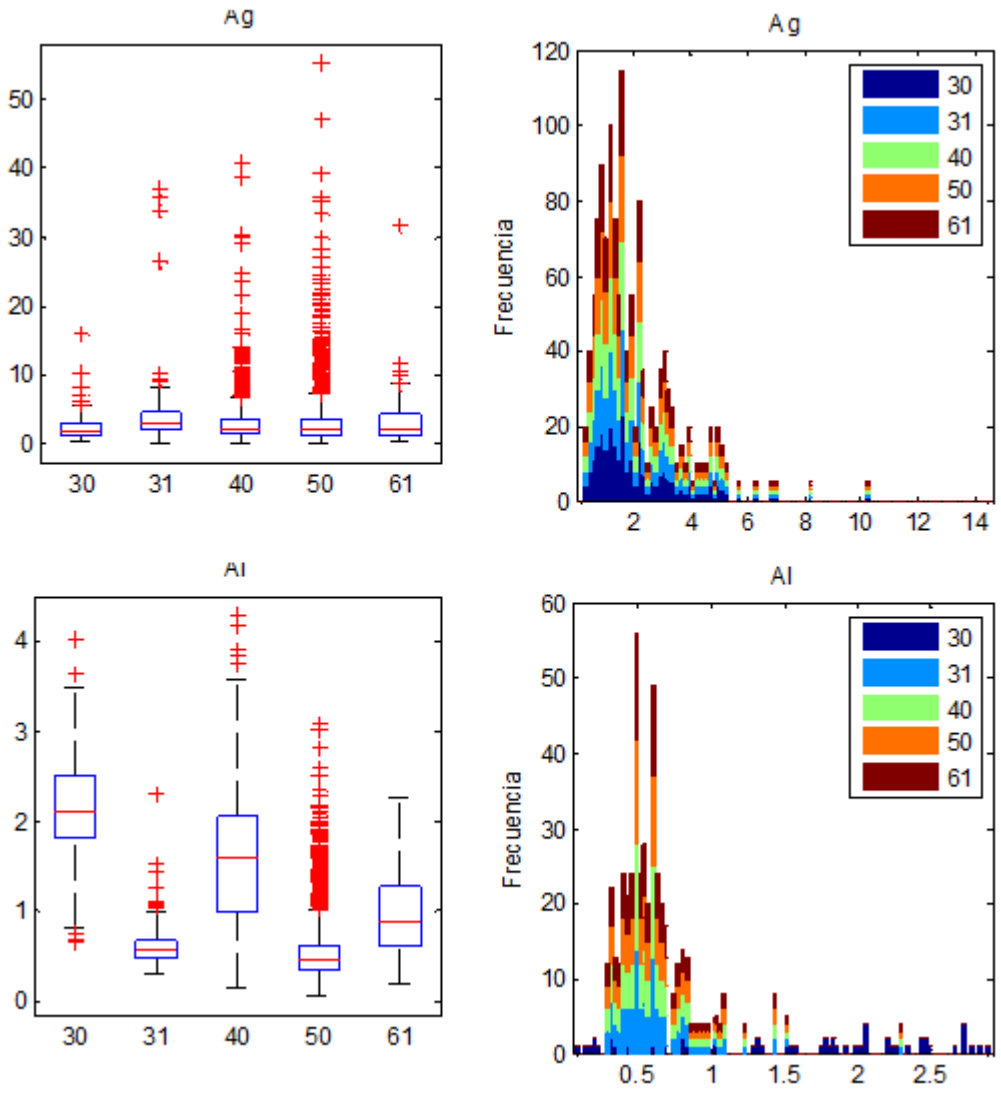
5 variables menos traslapadas

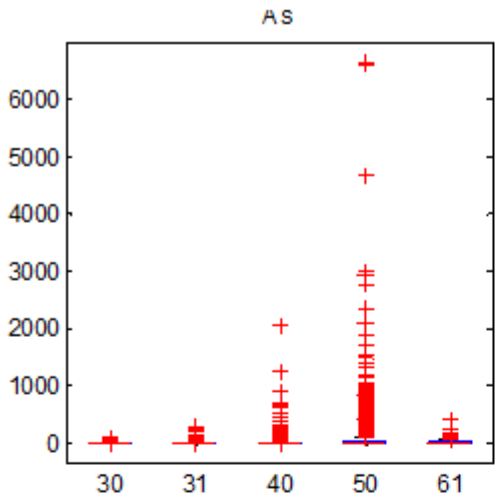


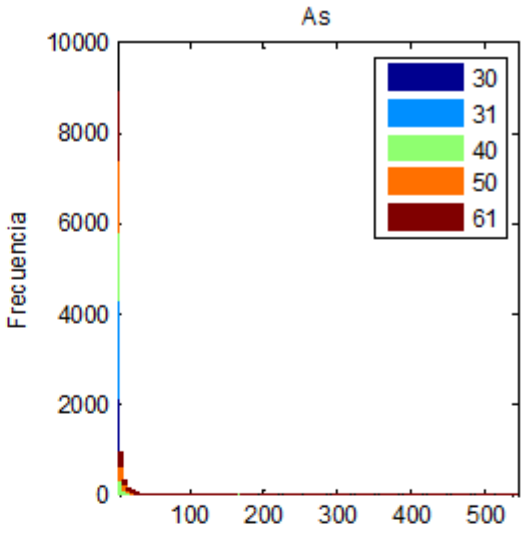


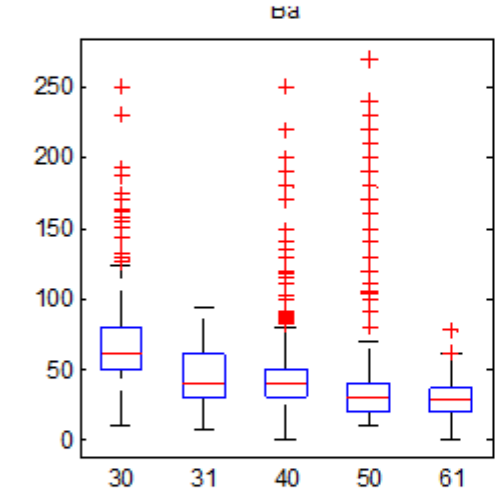


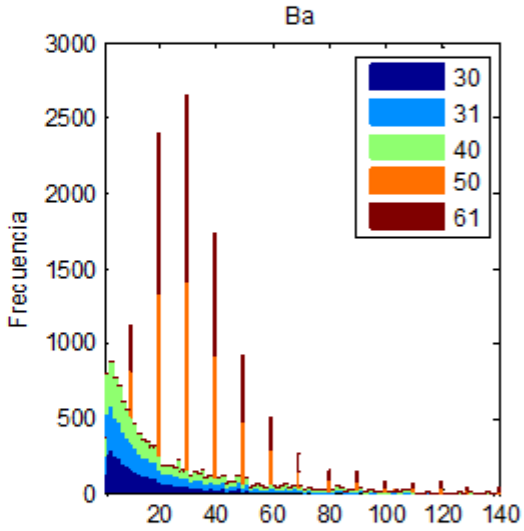
Histograma y Boxplots de elementos por alteraciones unidas

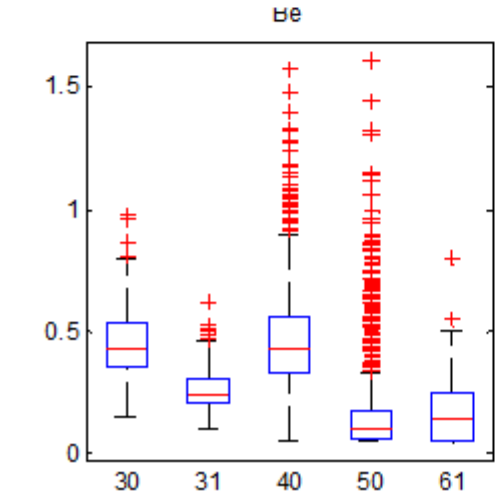


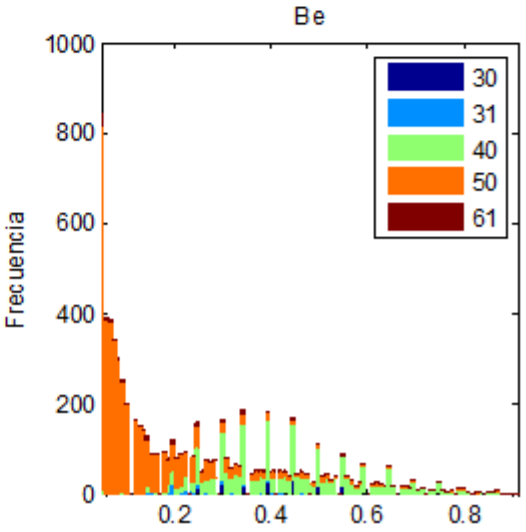

As
6000
5000
4000
3000
2000
1000
0
30
31
40
50
61

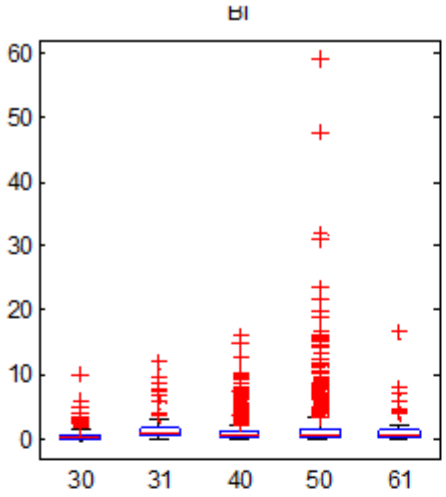

As
Frecuencia
10000
8000
6000
4000
2000
0
100
200
300
400
500
30
31
40
50
61

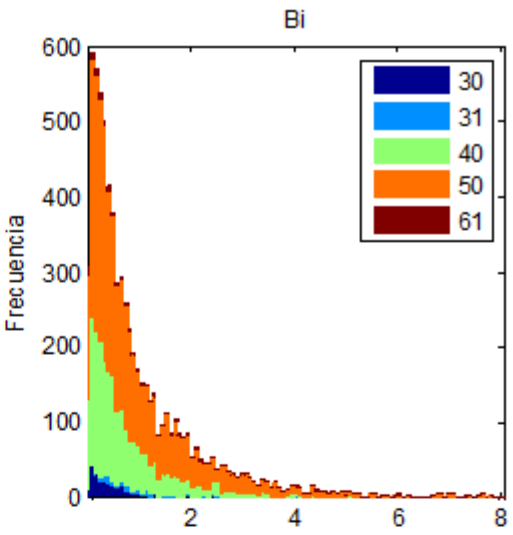

Ba
250
200
150
100
50
0
30
31
40
50
61

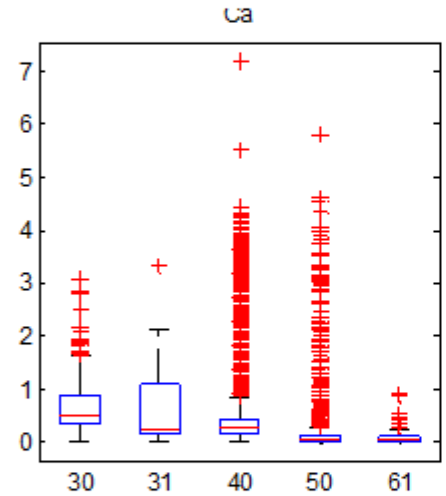

Ba
Frecuencia
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
30
31
40
50
61

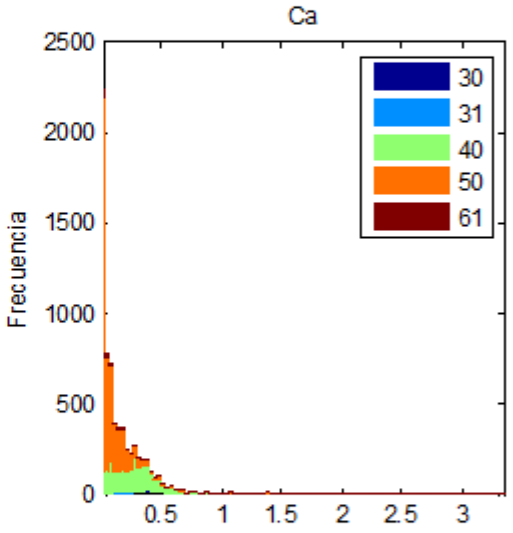

Be
1.5
1
0.5
0
30
31
40
50
61

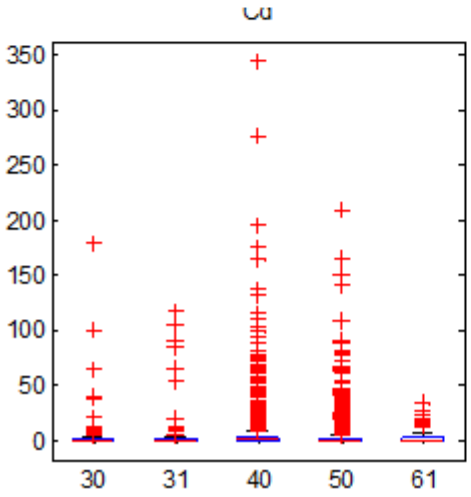

Be
Frecuencia
1000
800
600
400
200
0
0.2
0.4
0.6
0.8
30
31
40
50
61

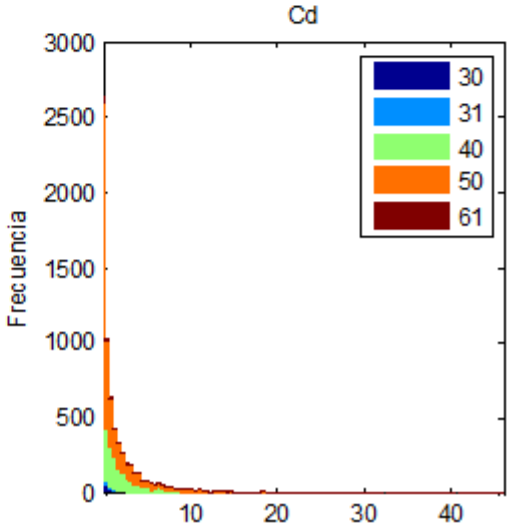

Bi
60
50
40
30
20
10
0
30
31
40
50
61

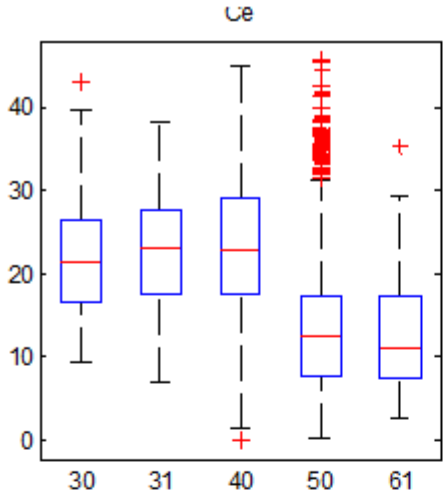

Bi
600
500
400
300
200
100
0
Frecuencia
2
4
6
8
30
31
40
50
61

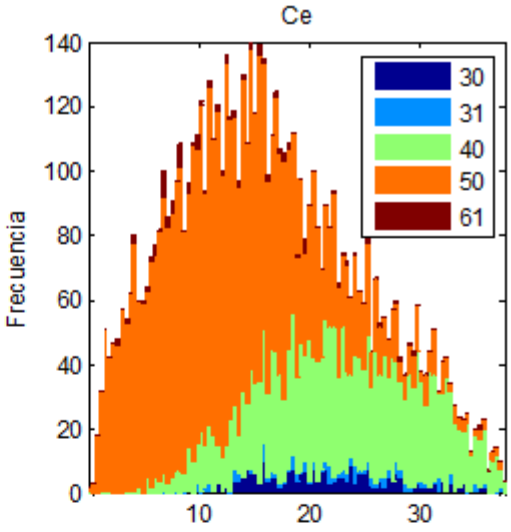

Ca
7
6
5
4
3
2
1
0
30
31
40
50
61

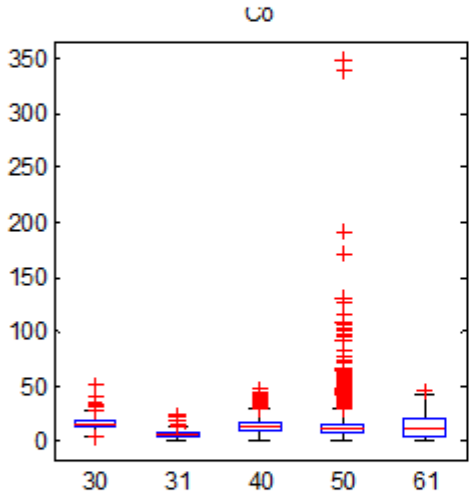

Ca
2500
2000
1500
1000
500
0
Frecuencia
0.5
1
1.5
2
2.5
3
30
31
40
50
61

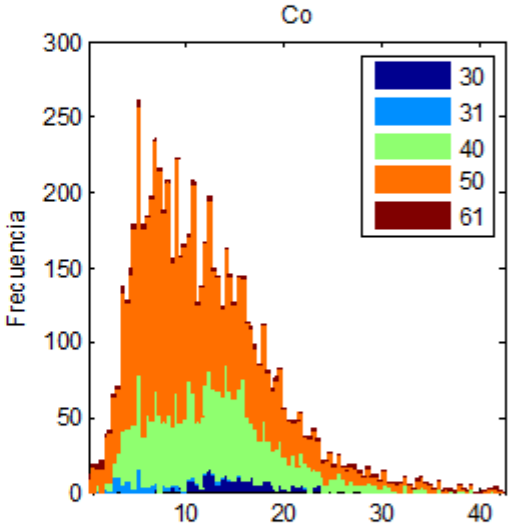

Cd
350
300
250
200
150
100
50
0
30
31
40
50
61

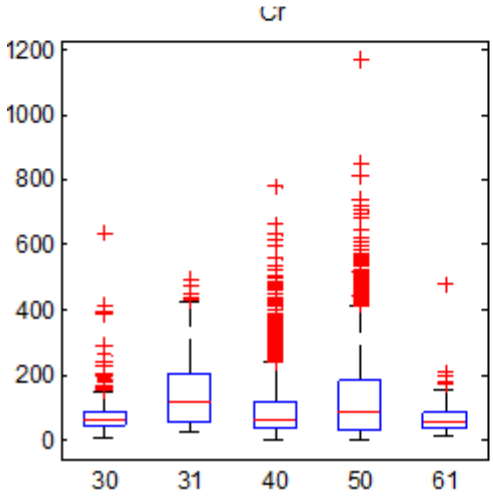

Cd
3000
2500
2000
1500
1000
500
0
Frecuencia
10
20
30
40
30
31
40
50
61

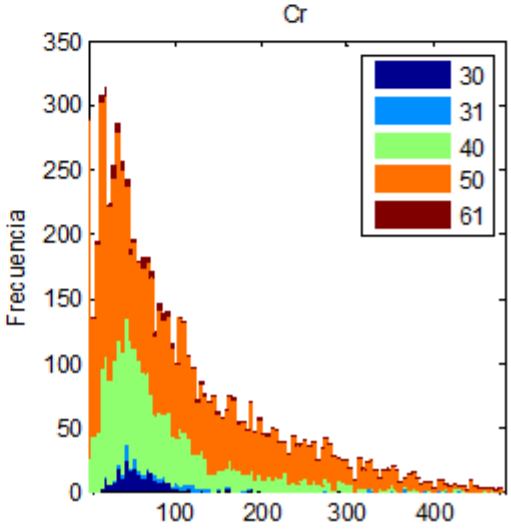

Ce
0
10
20
30
40
30
31
40
50
61

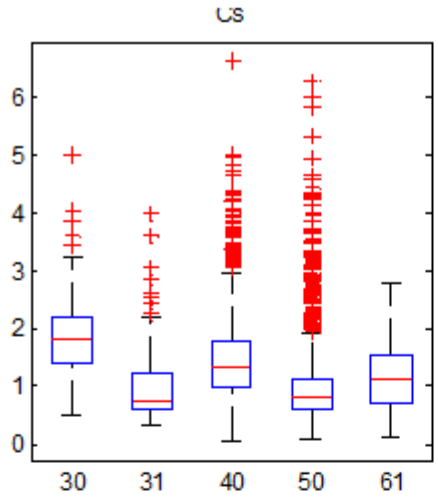

Ce
Frecuencia
0
20
40
60
80
100
120
140
10
20
30
30
31
40
50
61

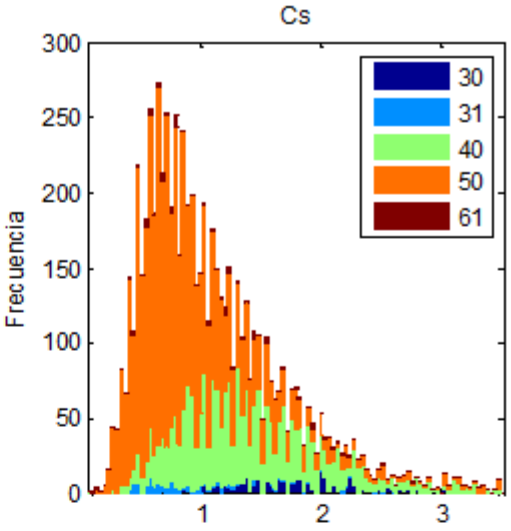

Co
0
50
100
150
200
250
300
350
30
31
40
50
61

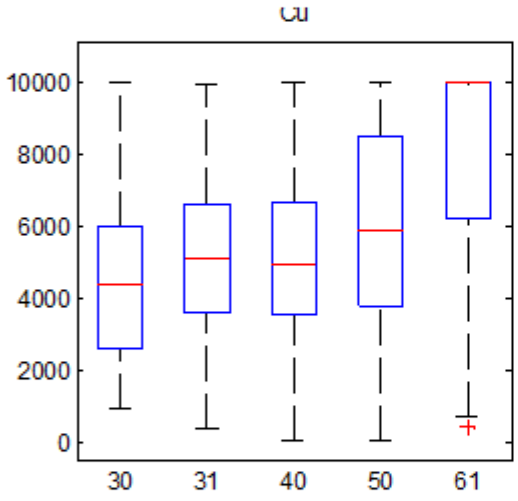

Co
Frecuencia
0
50
100
150
200
250
300
10
20
30
40
30
31
40
50
61

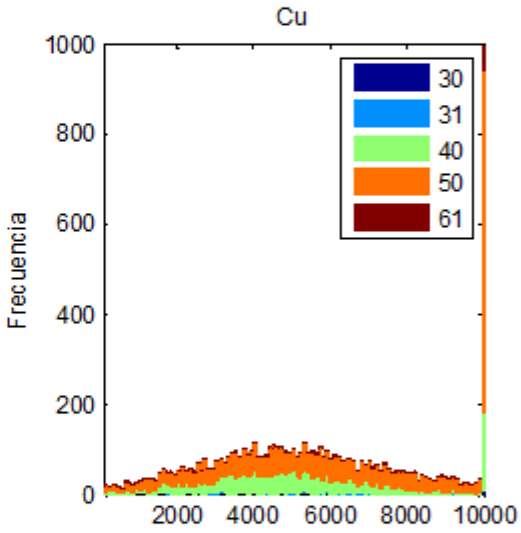

Cr
0
200
400
600
800
1000
1200
30
31
40
50
61

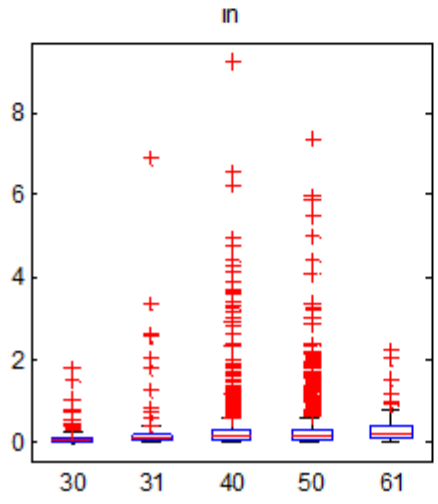

Cr
Frecuencia
0
50
100
150
200
250
300
350
100
200
300
400
30
31
40
50
61

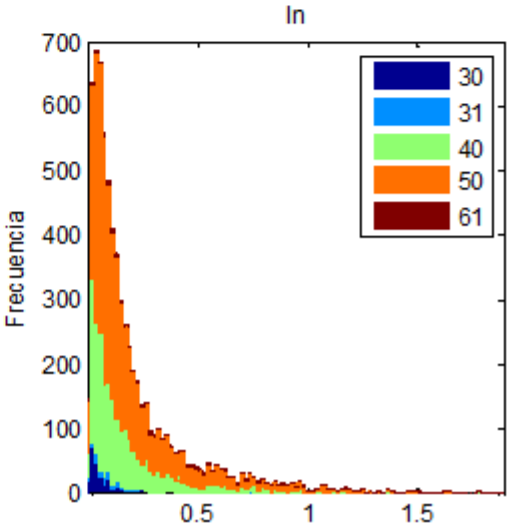

Cs
0
1
2
3
4
5
6
30
31
40
50
61

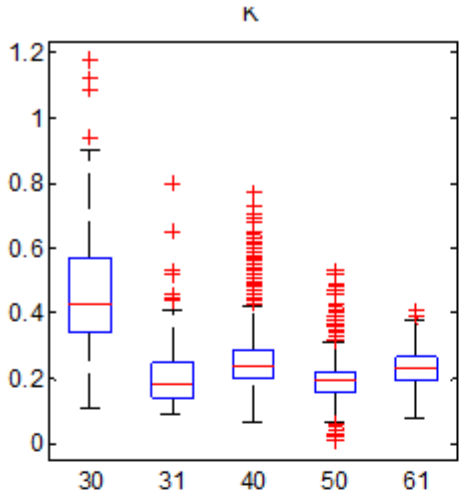

Cs
Frecuencia
0
50
100
150
200
250
300
1
2
3
30
31
40
50
61

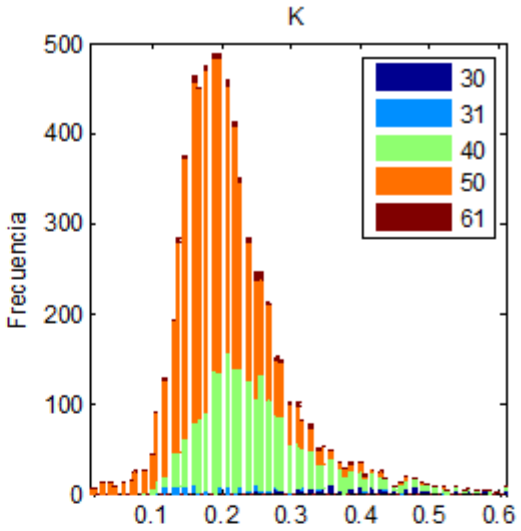

Cu
0
2000
4000
6000
8000
10000
30
31
40
50
61

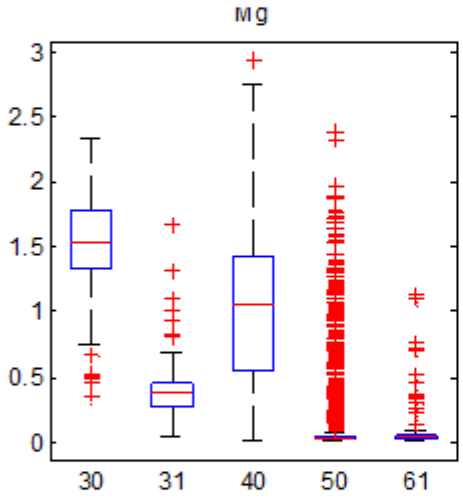

Cu
Frecuencia
0
200
400
600
800
1000
2000
4000
6000
8000
10000
30
31
40
50
61

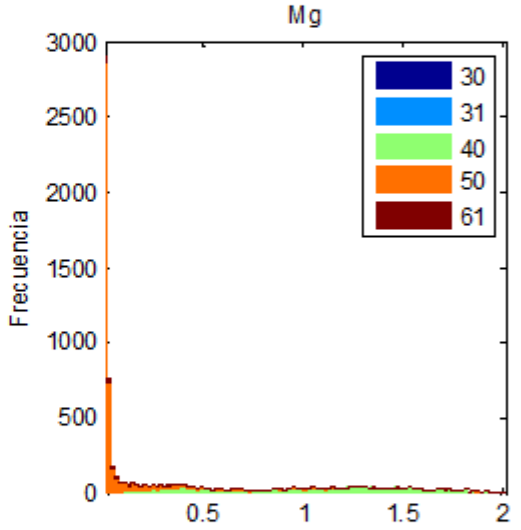

In
0
2
4
6
8
30
31
40
50
61

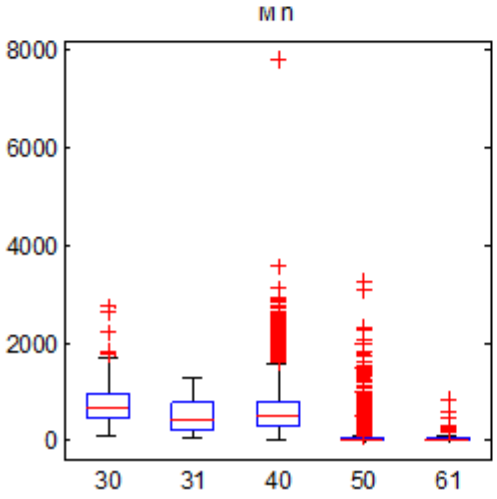

In
Frecuencia
0
100
200
300
400
500
600
700
0.5
1
1.5
30
31
40
50
61

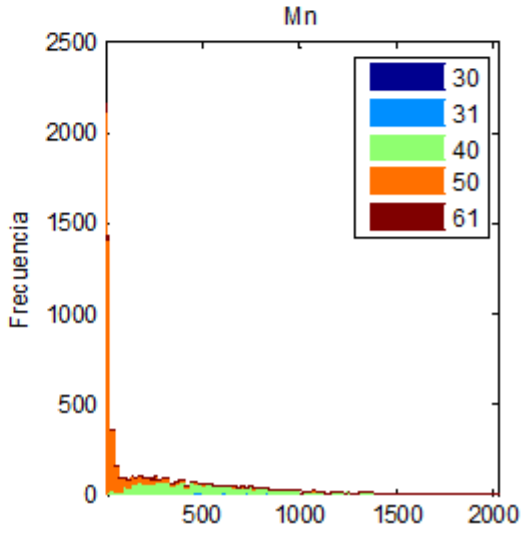

K
1.2
1
0.8
0.6
0.4
0.2
0
30
31
40
50
61

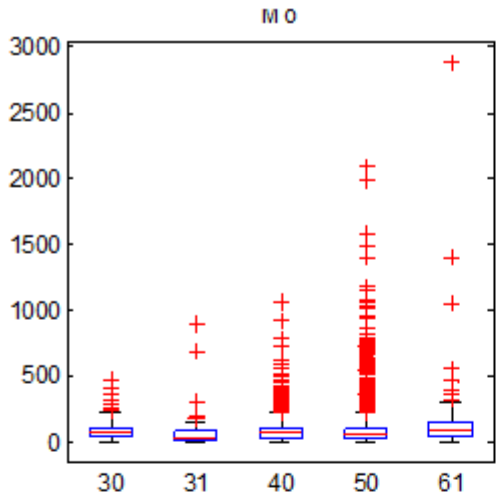

K
500
400
300
200
100
0
Frecuencia
0.1
0.2
0.3
0.4
0.5
0.6
30
31
40
50
61

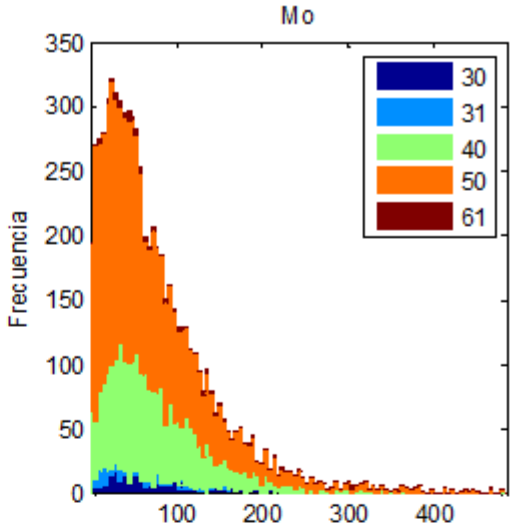

Mg
3
2.5
2
1.5
1
0.5
0
30
31
40
50
61

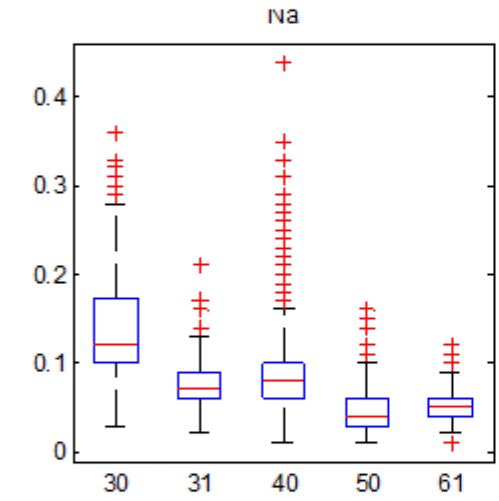

Mg
3000
2500
2000
1500
1000
500
0
Frecuencia
0.5
1
1.5
2
30
31
40
50
61

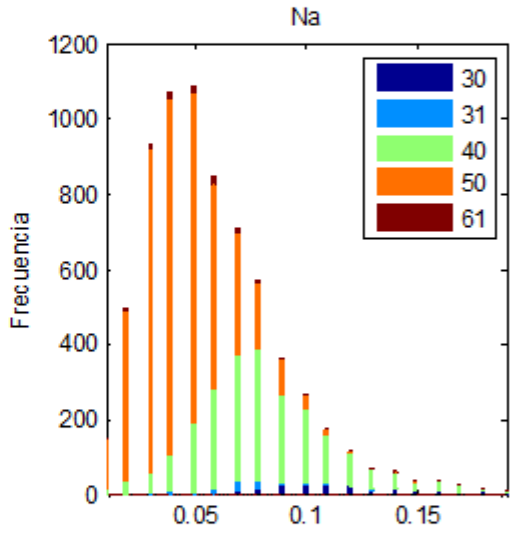

Mn
8000
6000
4000
2000
0
30
31
40
50
61

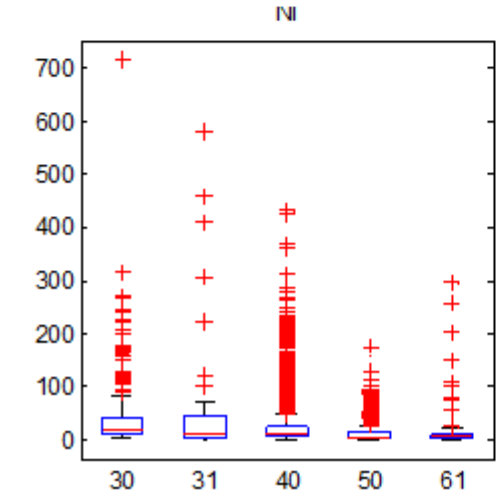

Mn
2500
2000
1500
1000
500
0
Frecuencia
500
1000
1500
2000
30
31
40
50
61

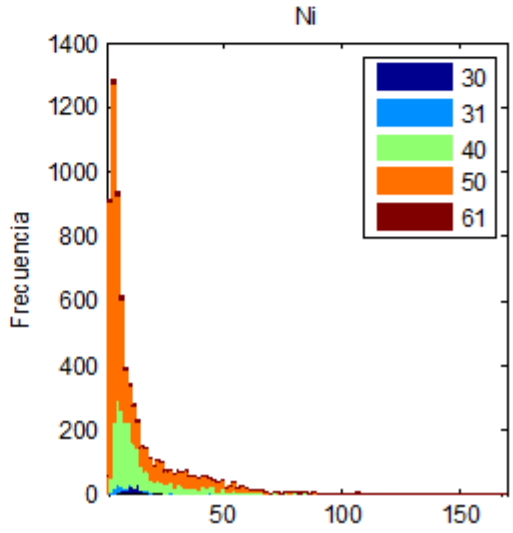

Mo
3000
2500
2000
1500
1000
500
0
30
31
40
50
61

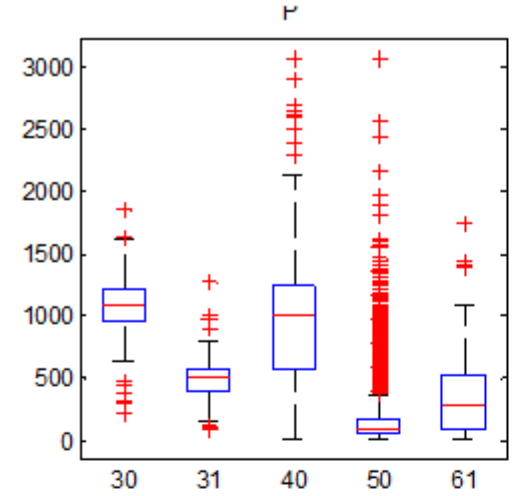

Mo
Frecuencia
350
300
250
200
150
100
50
0
100
200
300
400
30
31
40
50
61

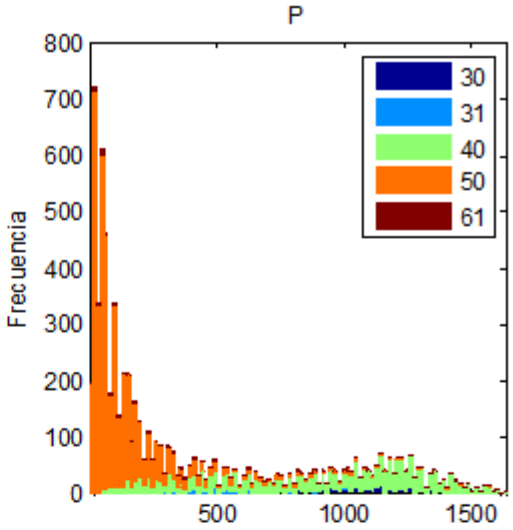

Na
0.4
0.3
0.2
0.1
0
30
31
40
50
61

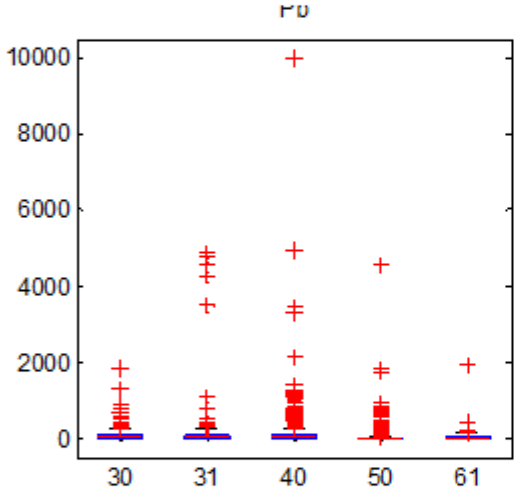

Na
Frecuencia
1200
1000
800
600
400
200
0
0.05
0.1
0.15
30
31
40
50
61

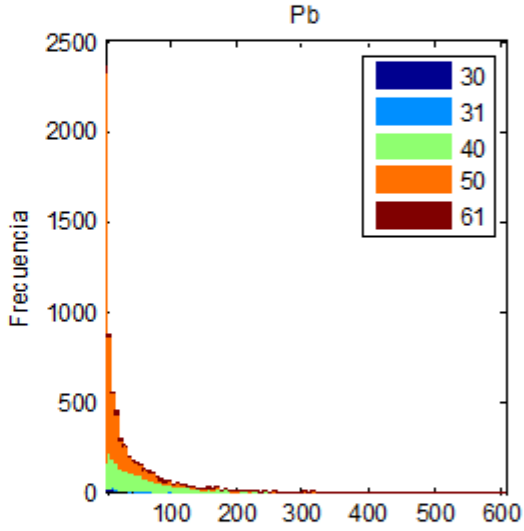

Ni
700
600
500
400
300
200
100
0
30
31
40
50
61

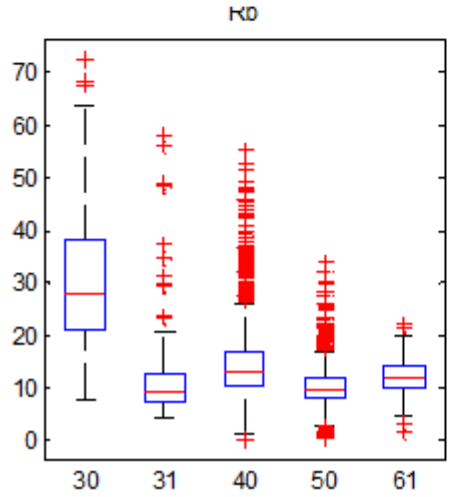

Ni
Frecuencia
1400
1200
1000
800
600
400
200
0
50
100
150
30
31
40
50
61

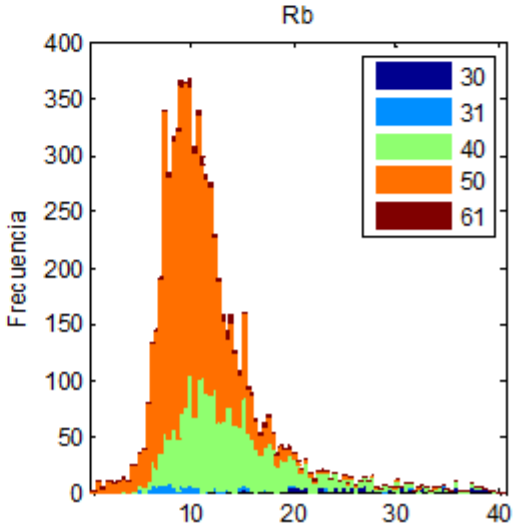

3000
2500
2000
1500
1000
500
0
30
31
40
50
61

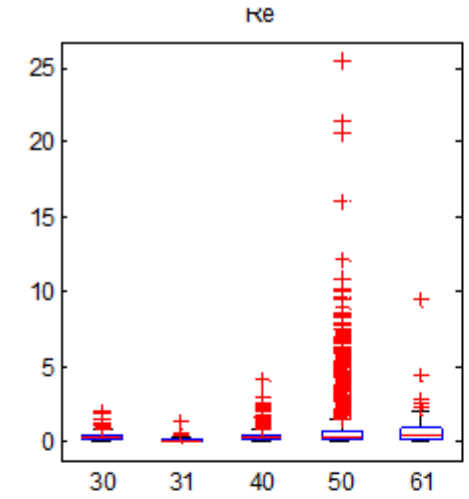

P
800
700
600
500
400
300
200
100
0
Frecuencia
500
1000
1500
30
31
40
50
61

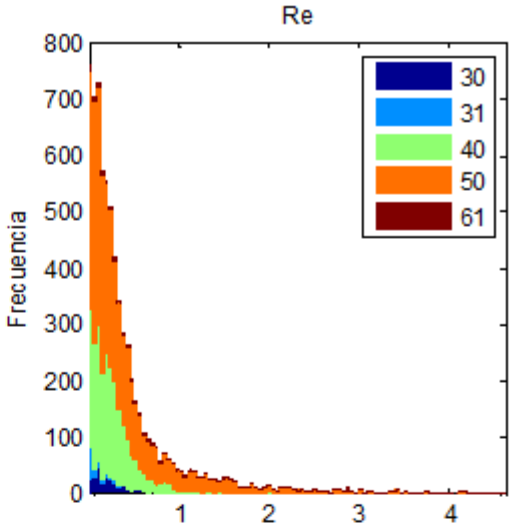

10000
8000
6000
4000
2000
0
30
31
40
50
61

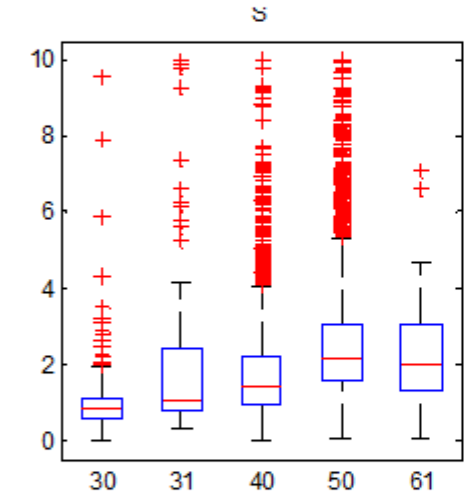

Pb
2500
2000
1500
1000
500
0
Frecuencia
100
200
300
400
500
600
30
31
40
50
61

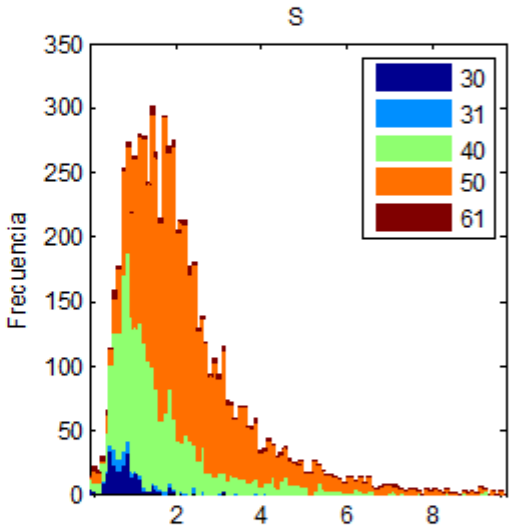

70
60
50
40
30
20
10
0
30
31
40
50
61

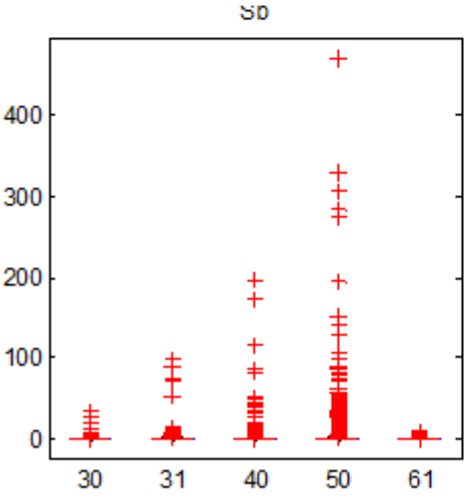

Rb
400
350
300
250
200
150
100
50
0
Frecuencia
10
20
30
40
30
31
40
50
61

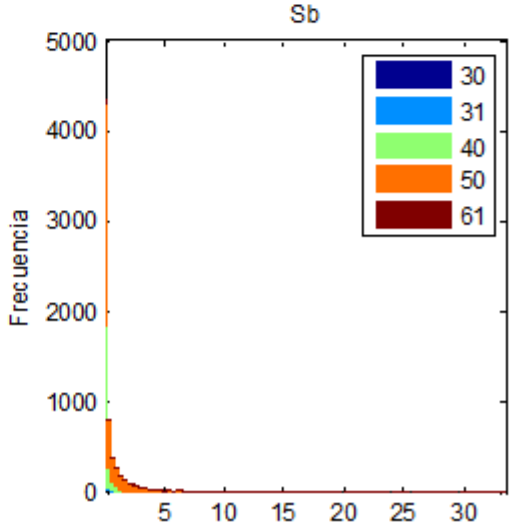

Re
25
20
15
10
5
0
30
31
40
50
61

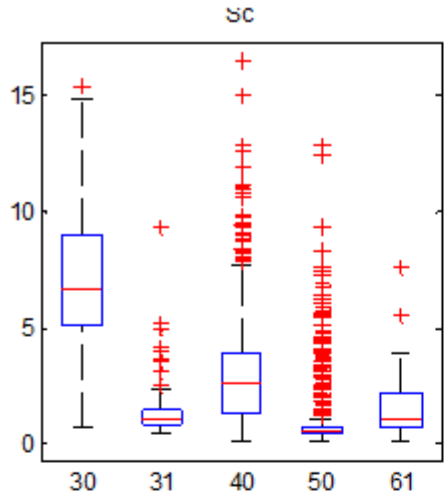

Re
800
700
600
500
400
300
200
100
0
Frecuencia
1
2
3
4
30
31
40
50
61

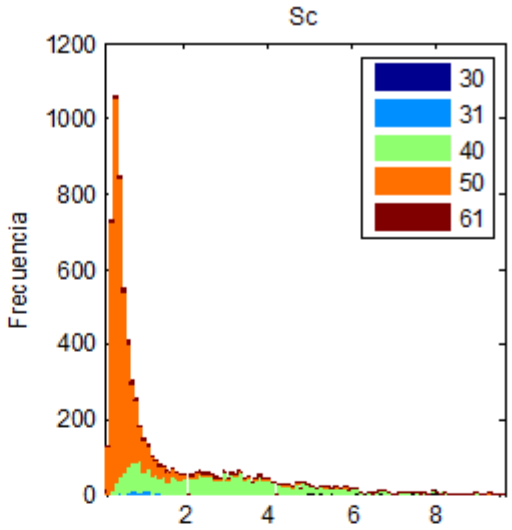

S
10
8
6
4
2
0
30
31
40
50
61

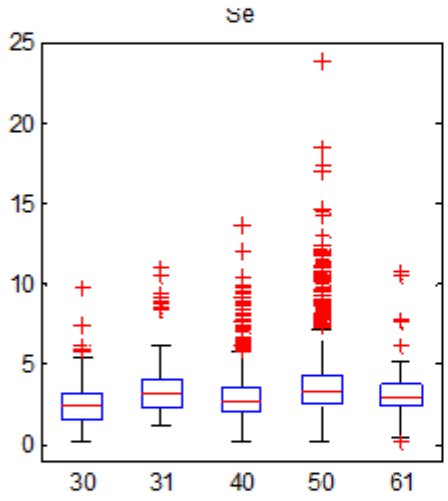

S
350
300
250
200
150
100
50
0
Frecuencia
2
4
6
8
30
31
40
50
61

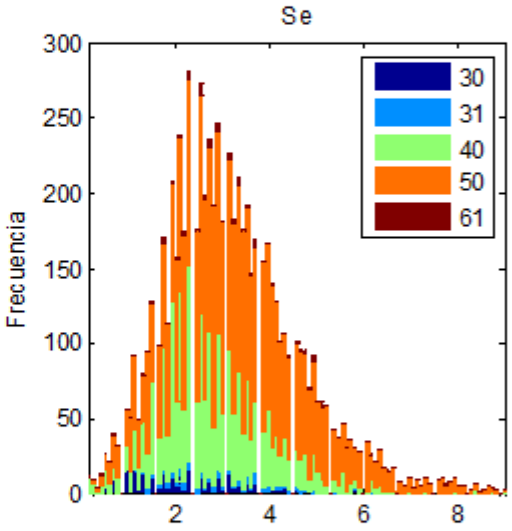

Sb
400
300
200
100
0
30
31
40
50
61

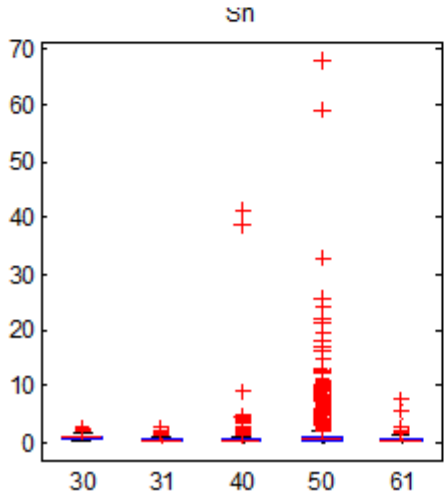

Sb
5000
4000
3000
2000
1000
0
Frecuencia
5
10
15
20
25
30
30
31
40
50
61

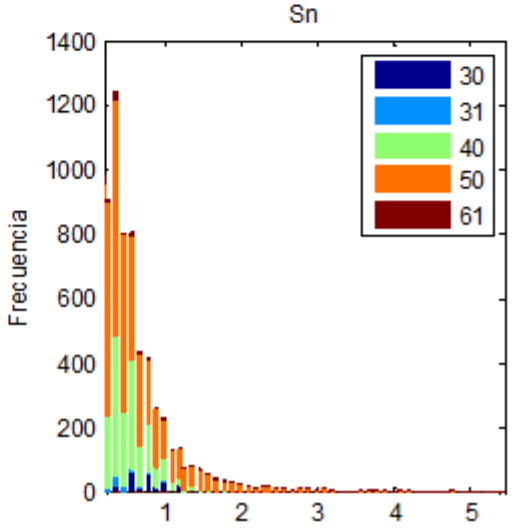

Sc
15
10
5
0
30
31
40
50
61

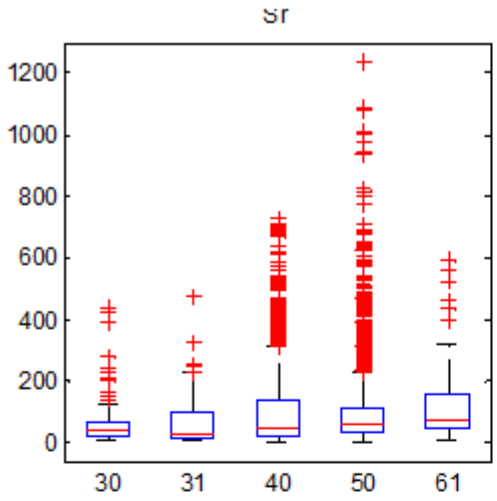

Sc
Frecuencia
1200
1000
800
600
400
200
0
2
4
6
8
30
31
40
50
61

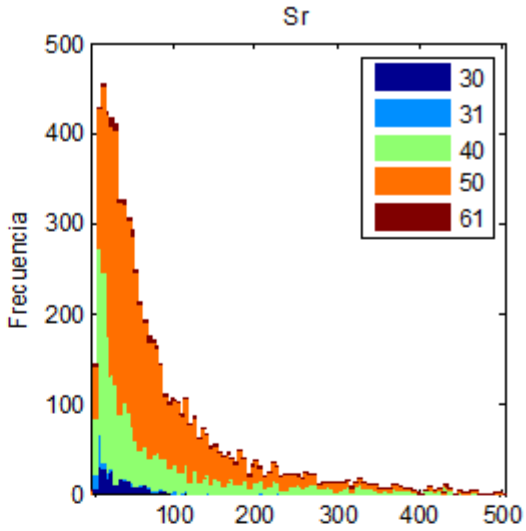

Se
25
20
15
10
5
0
30
31
40
50
61

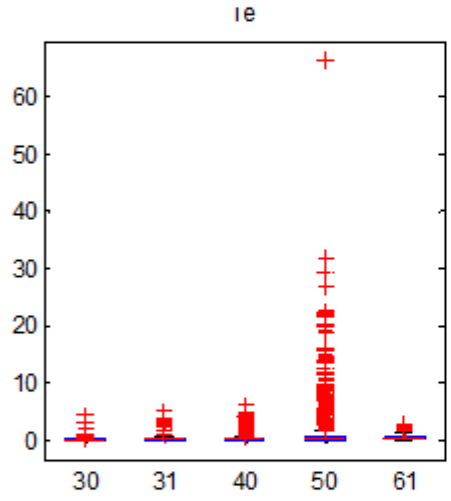

Se
Frecuencia
300
250
200
150
100
50
0
2
4
6
8
30
31
40
50
61

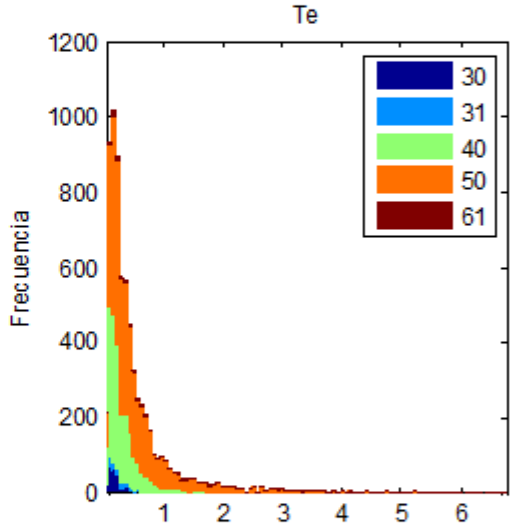

Sn
70
60
50
40
30
20
10
0
30
31
40
50
61

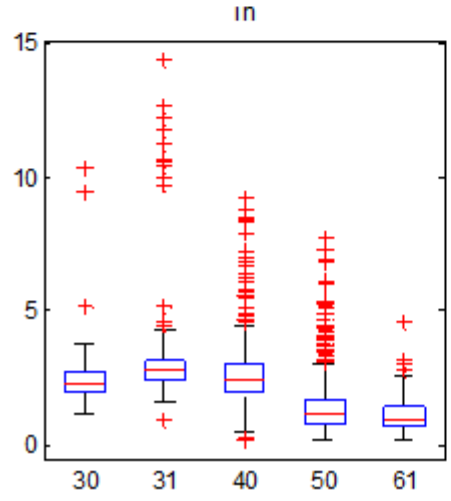

Sn
Frecuencia
1400
1200
1000
800
600
400
200
0
1
2
3
4
5
30
31
40
50
61

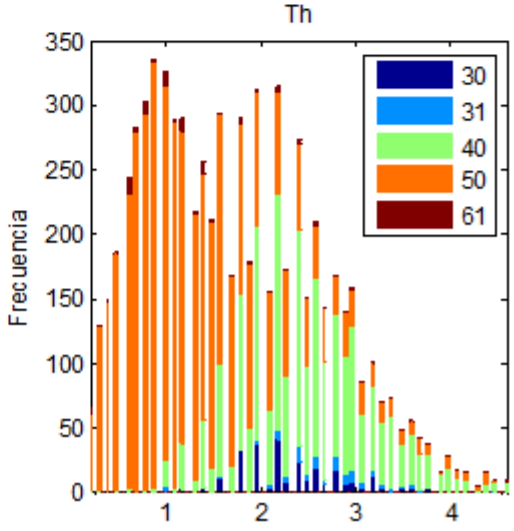

Sr
1200
1000
800
600
400
200
0
30
31
40
50
61

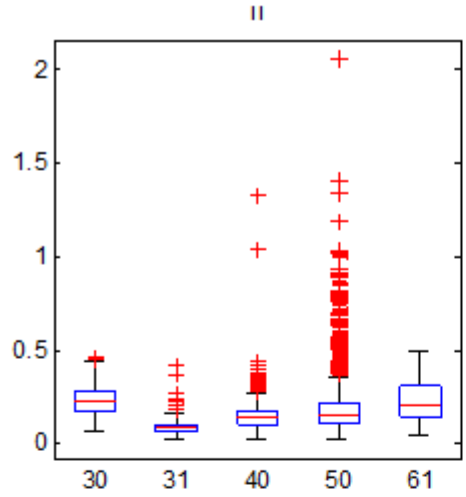

Sr
Frecuencia
500
400
300
200
100
0
100
200
300
400
500
30
31
40
50
61

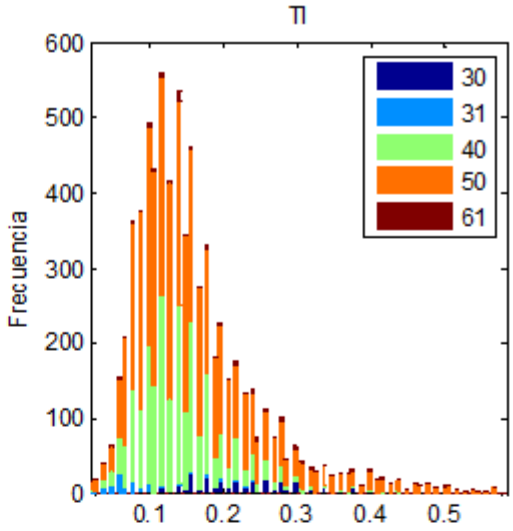

Te
60
50
40
30
20
10
0
30
31
40
50
61

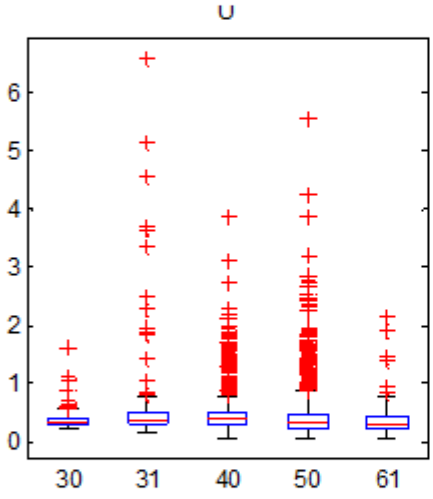

Te
Frecuencia
1200
1000
800
600
400
200
0
1
2
3
4
5
6
30
31
40
50
61

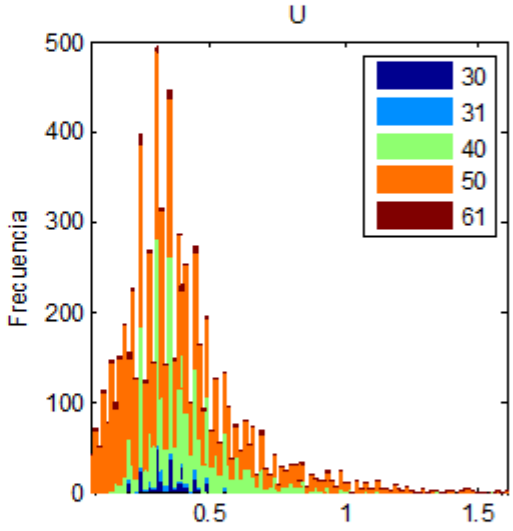

In
15
10
5
0
30
31
40
50
61

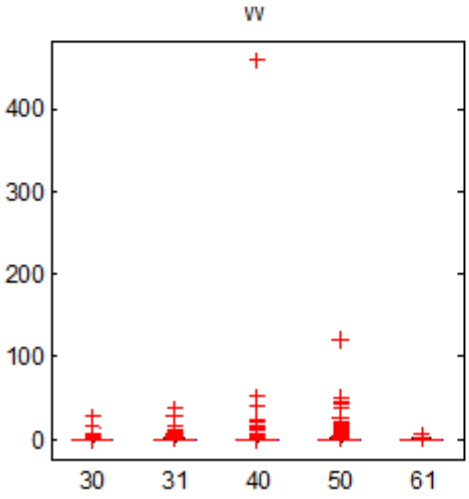

Th
Frecuencia
350
300
250
200
150
100
50
0
1
2
3
4
30
31
40
50
61

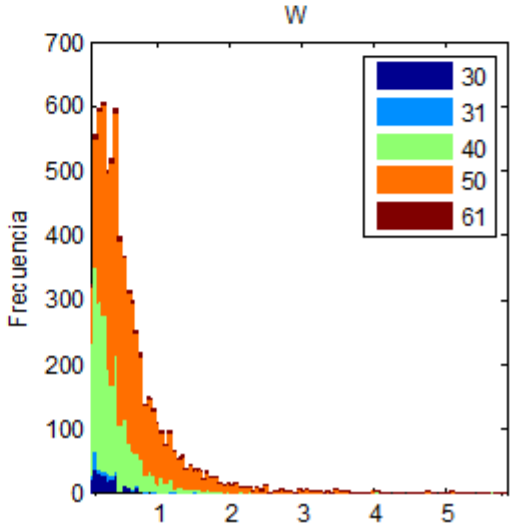

II
2
1.5
1
0.5
0
30
31
40
50
61

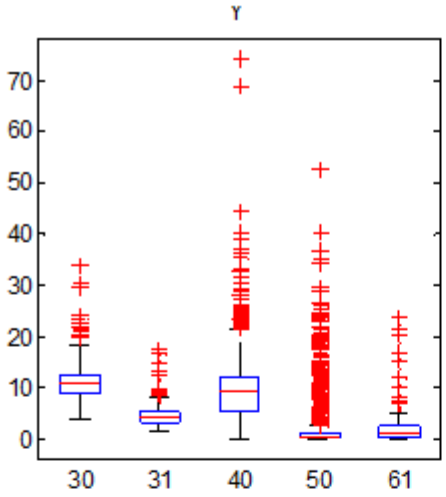

II
600
500
400
300
200
100
0
Frecuencia
0.1
0.2
0.3
0.4
0.5
30
31
40
50
61

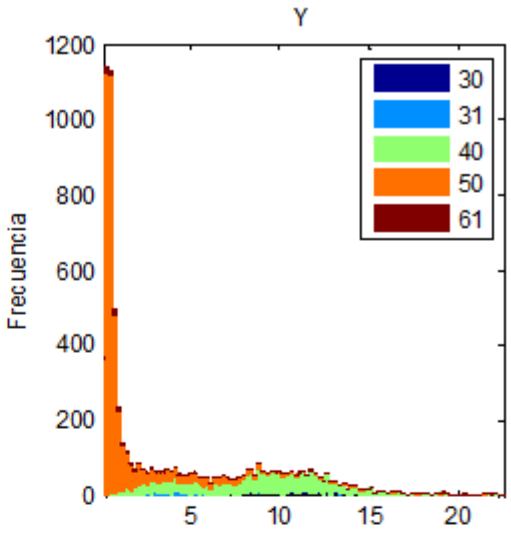

U
6
5
4
3
2
1
0
30
31
40
50
61

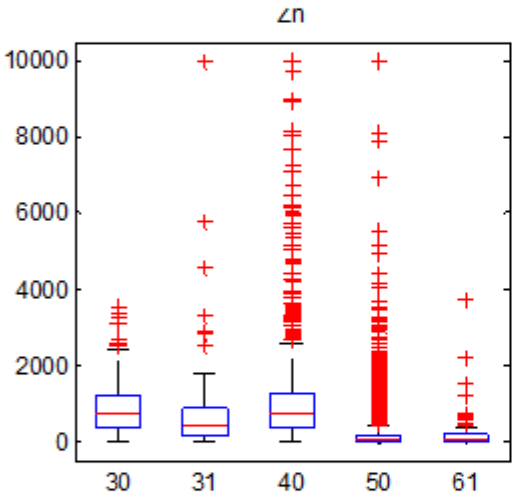

U
500
400
300
200
100
0
Frecuencia
0.5
1
1.5
30
31
40
50
61

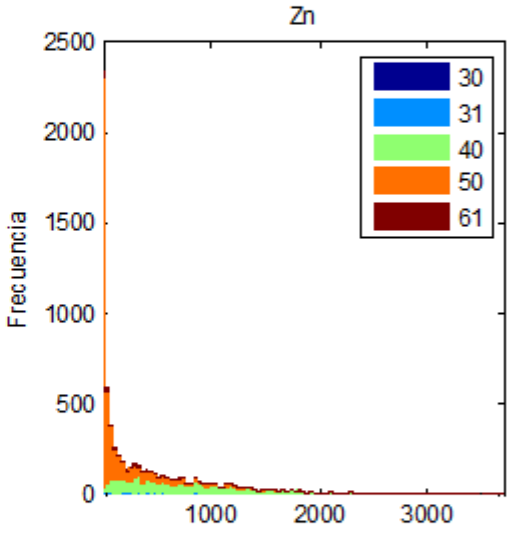

W
400
300
200
100
0
30
31
40
50
61

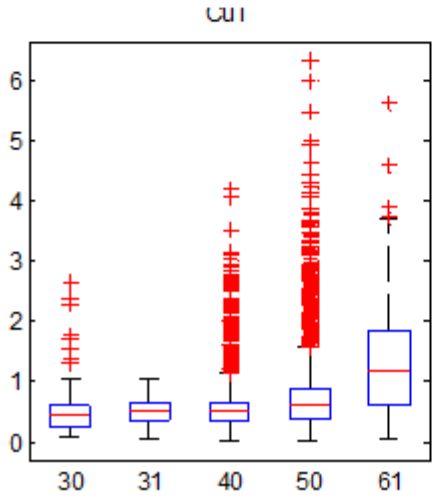

W
700
600
500
400
300
200
100
0
Frecuencia
1
2
3
4
5
30
31
40
50
61

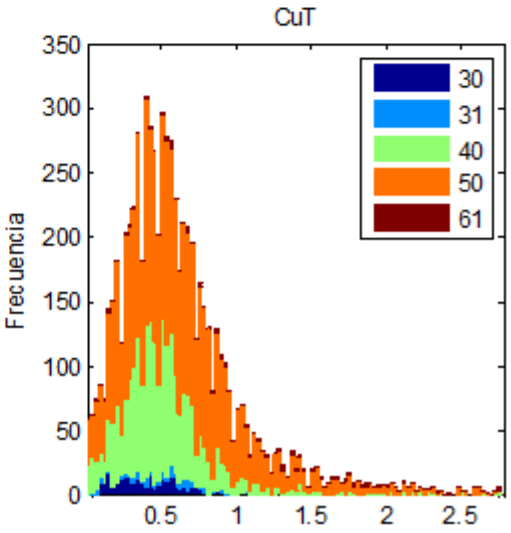

Y
70
60
50
40
30
20
10
0
30
31
40
50
61

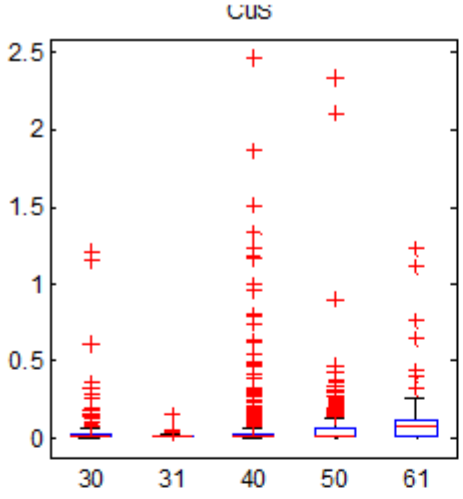

Y
1200
1000
800
600
400
200
0
Frecuencia
5
10
15
20
30
31
40
50
61

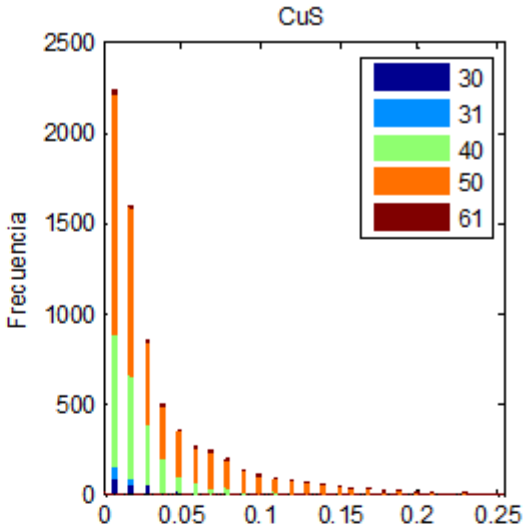

Zn
10000
8000
6000
4000
2000
0
30
31
40
50
61

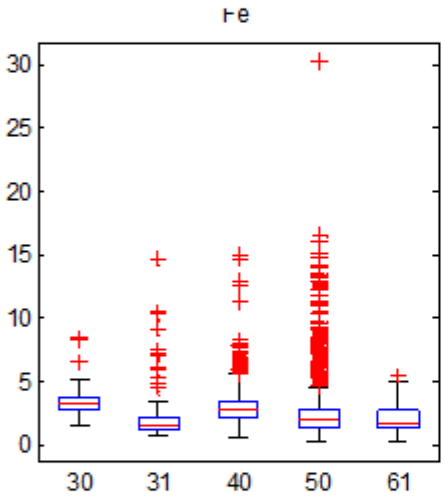

Zn
2500
2000
1500
1000
500
0
Frecuencia
1000
2000
3000
30
31
40
50
61

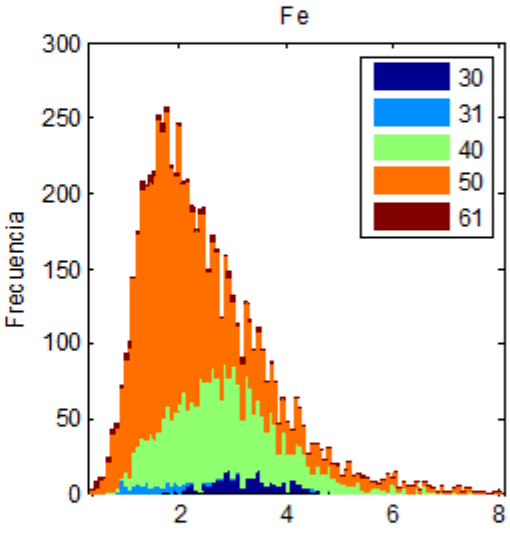

CuT
6
5
4
3
2
1
0
30
31
40
50
61

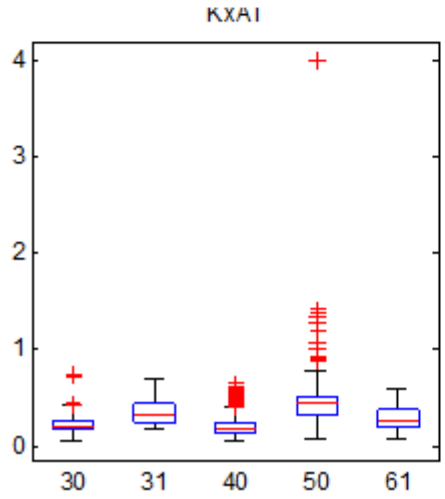

CuT
350
300
250
200
150
100
50
0
Frecuencia
0.5
1
1.5
2
2.5
30
31
40
50
61

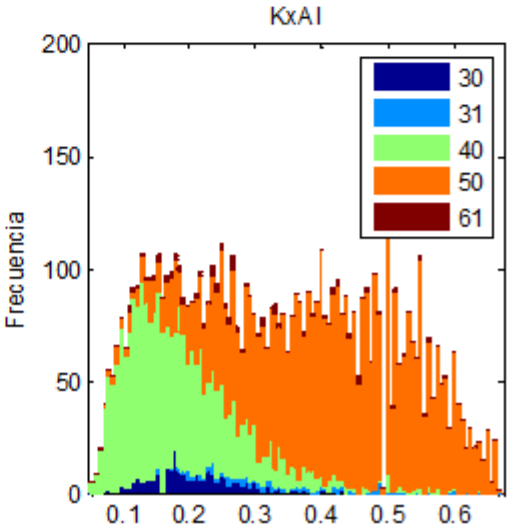

CuS
2.5
2
1.5
1
0.5
0
30
31
40
50
61

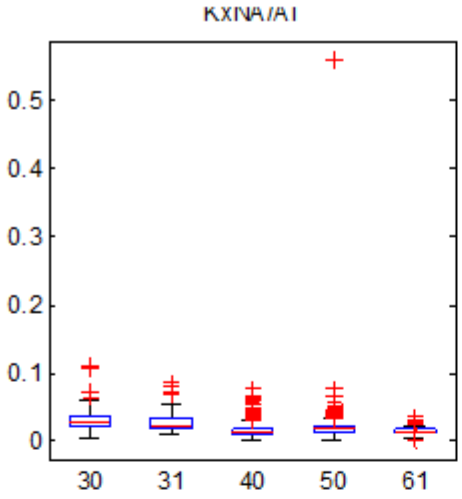

CuS
2500
2000
1500
1000
500
0
Frecuencia
0
0.05
0.1
0.15
0.2
0.25
30
31
40
50
61

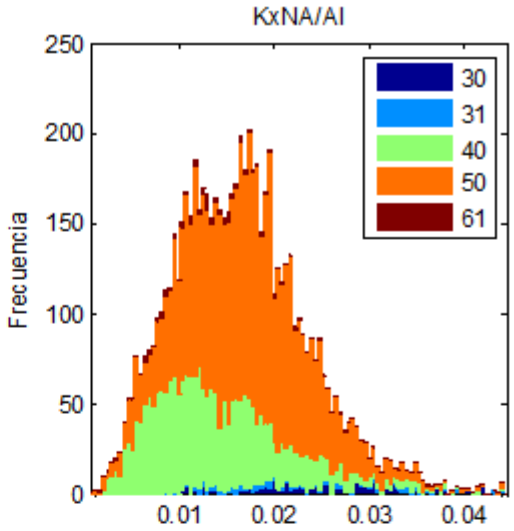

Fe
30
25
20
15
10
5
0
30
31
40
50
61

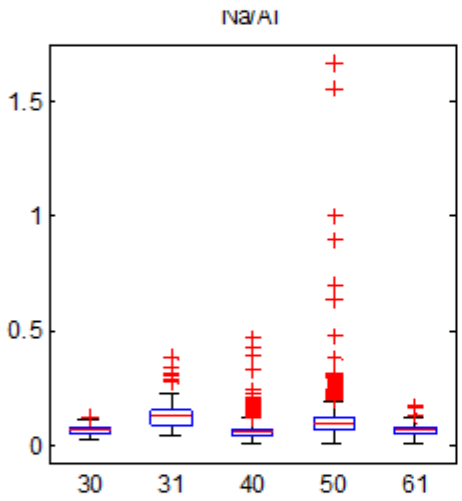

Fe
300
250
200
150
100
50
0
Frecuencia
2
4
6
8
30
31
40
50
61

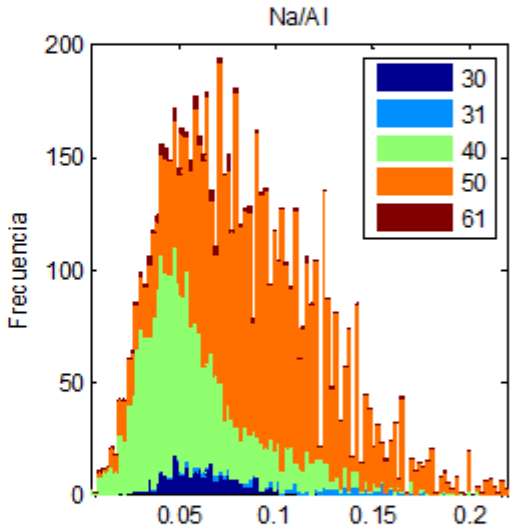

KxAl
4
3
2
1
0
30
31
40
50
61

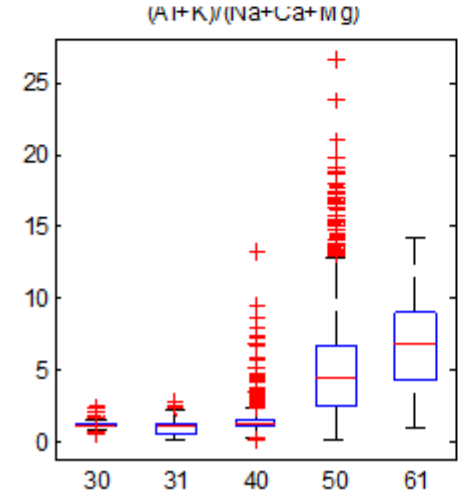

KxAl
200
150
100
50
0
Frecuencia
0.1
0.2
0.3
0.4
0.5
0.6
30
31
40
50
61

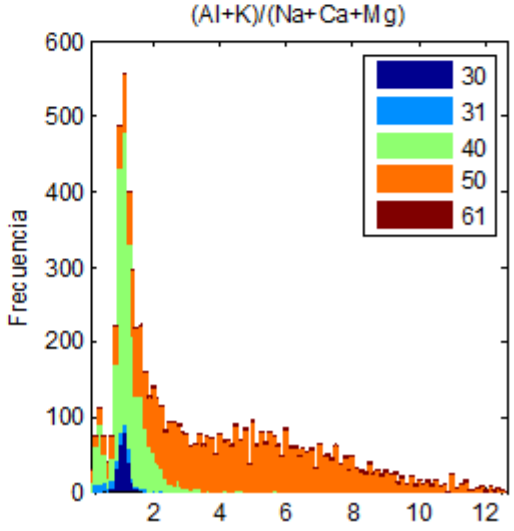

KxNA/AI
0.5
0.4
0.3
0.2
0.1
0
30
31
40
50
61

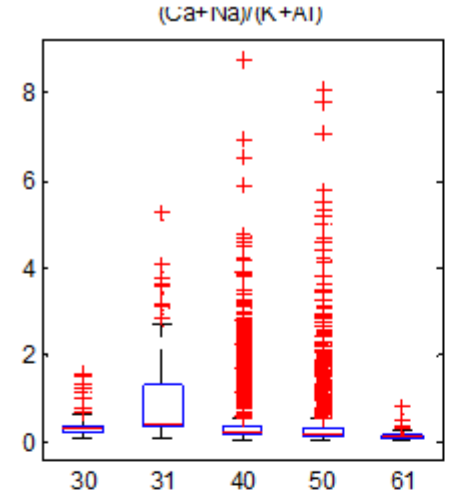

KxNA/AI
Frecuencia
250
200
150
100
50
0
0.01
0.02
0.03
0.04
30
31
40
50
61

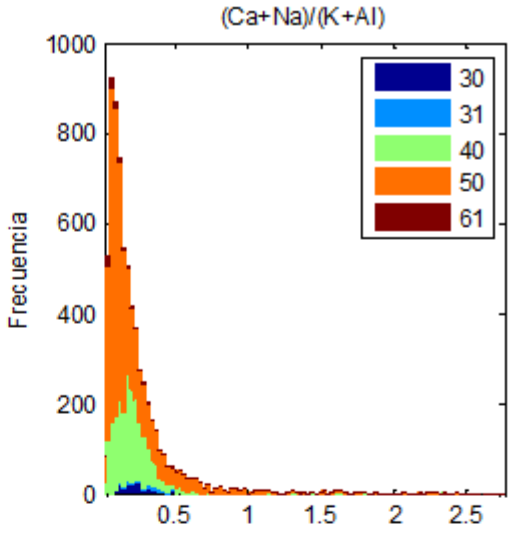

Na/AI
1.5
1
0.5
0
30
31
40
50
61

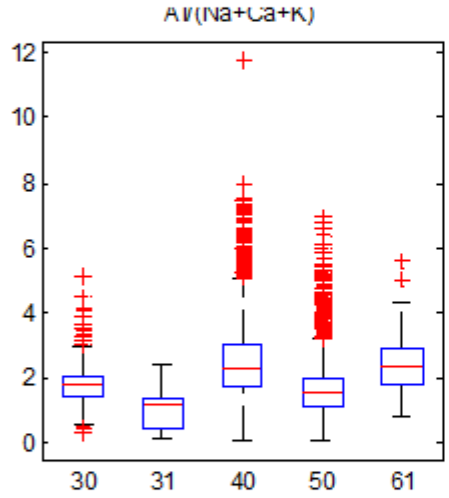

Na/AI
Frecuencia
200
150
100
50
0
0.05
0.1
0.15
0.2
30
31
40
50
61

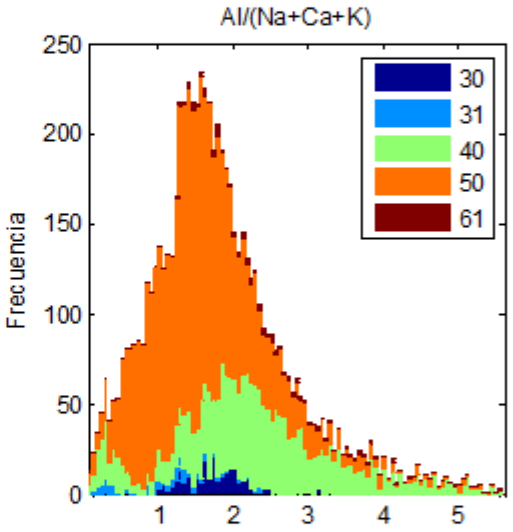

(AI+K)/(Na+Ca+Mg)
25
20
15
10
5
0
30
31
40
50
61

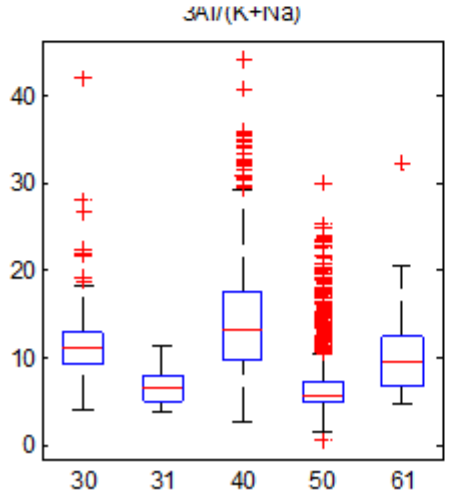

(AI+K)/(Na+Ca+Mg)
Frecuencia
600
500
400
300
200
100
0
2
4
6
8
10
12
30
31
40
50
61

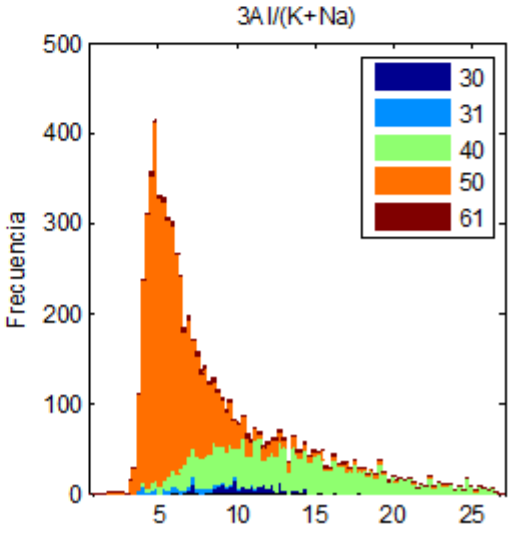

(Ca+Na)/(K+Al)
0
2
4
6
8
30
31
40
50
61

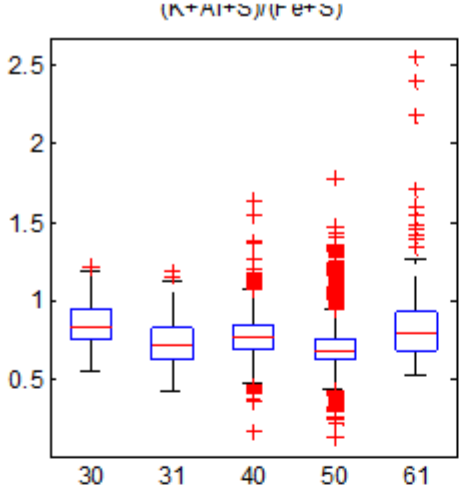

(Ca+Na)/(K+Al)
Frecuencia
0
200
400
600
800
1000
0.5
1
1.5
2
2.5
30
31
40
50
61

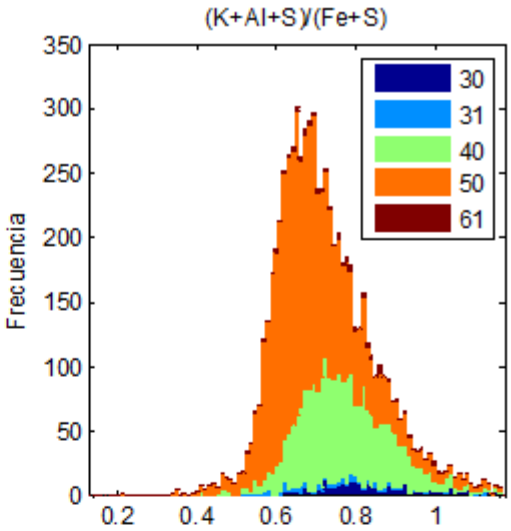

Al/(Na+Ca+K)
0
2
4
6
8
10
12
30
31
40
50
61

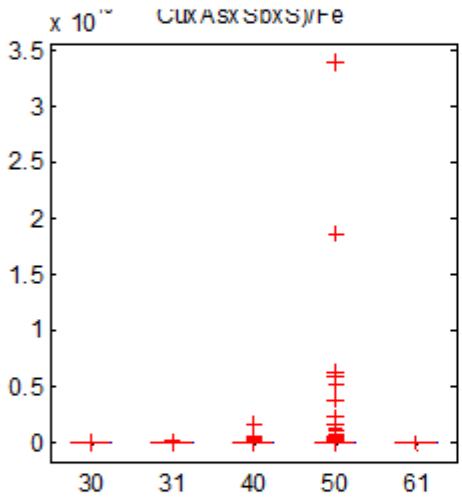

Al/(Na+Ca+K)
Frecuencia
0
50
100
150
200
250
1
2
3
4
5
30
31
40
50
61

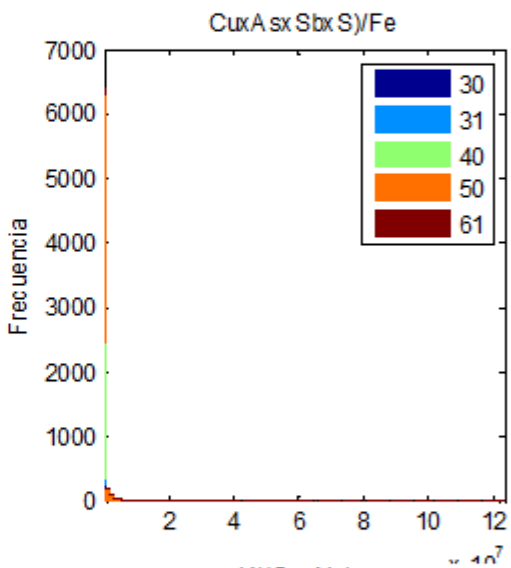

3Al/(K+Na)
0
10
20
30
40
30
31
40
50
61

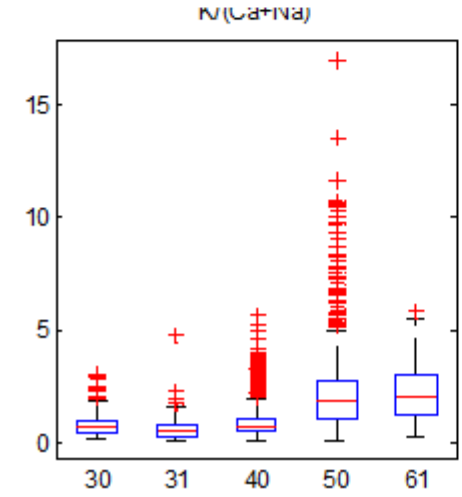

3Al/(K+Na)
Frecuencia
0
100
200
300
400
500
5
10
15
20
25
30
31
40
50
61

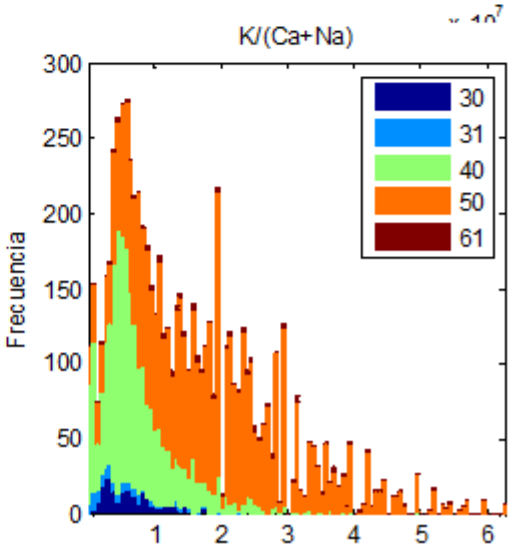

(K+Al+S)/(Fe+S)
2.5
2
1.5
1
0.5
30
31
40
50
61

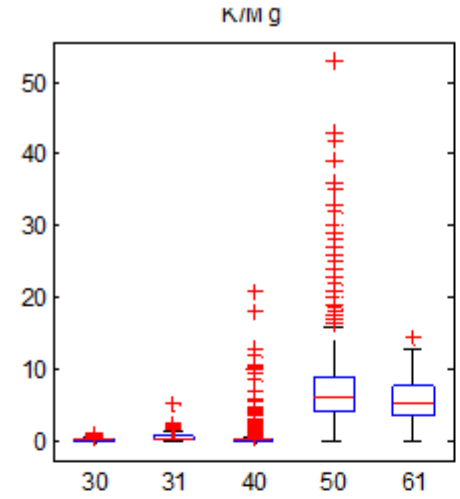

(K+Al+S)/(Fe+S)
Frecuencia
350
300
250
200
150
100
50
0
0.2
0.4
0.6
0.8
1
30
31
40
50
61

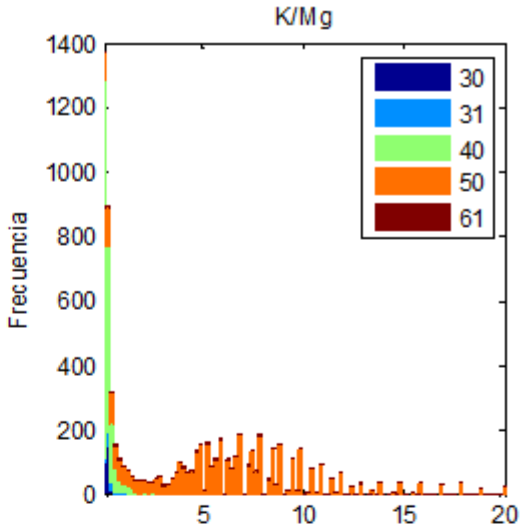

x 10
CuxAsxSbxS)/Fe
3.5
3
2.5
2
1.5
1
0.5
0
30
31
40
50
61

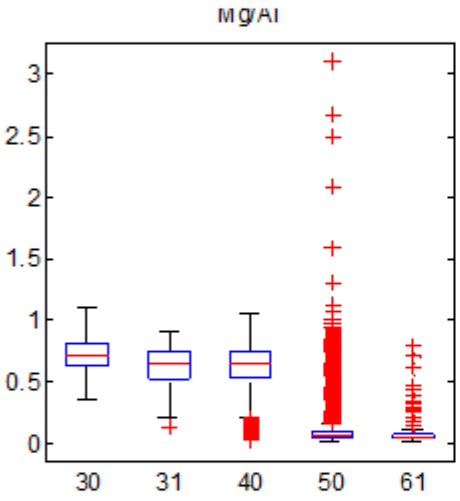

CuxAsxSbxS)/Fe
Frecuencia
7000
6000
5000
4000
3000
2000
1000
0
2
4
6
8
10
12
30
31
40
50
61

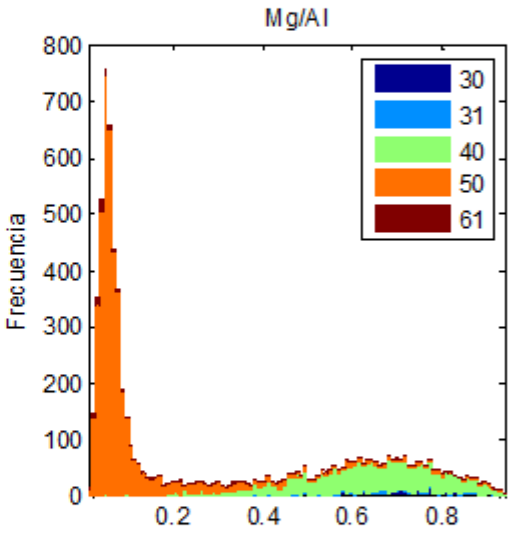

K/(Ca+Na)
15
10
5
0
30
31
40
50
61

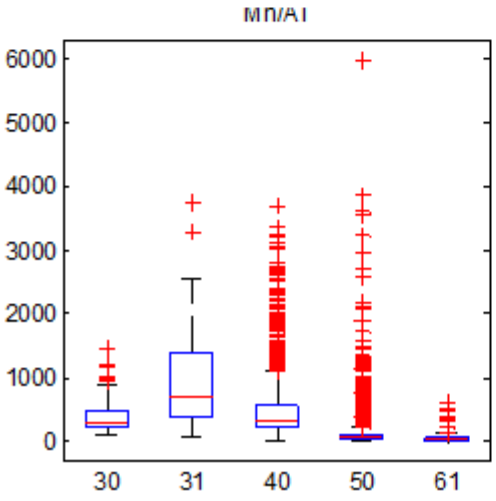

K/(Ca+Na)
Frecuencia
300
250
200
150
100
50
0
1
2
3
4
5
6
30
31
40
50
61

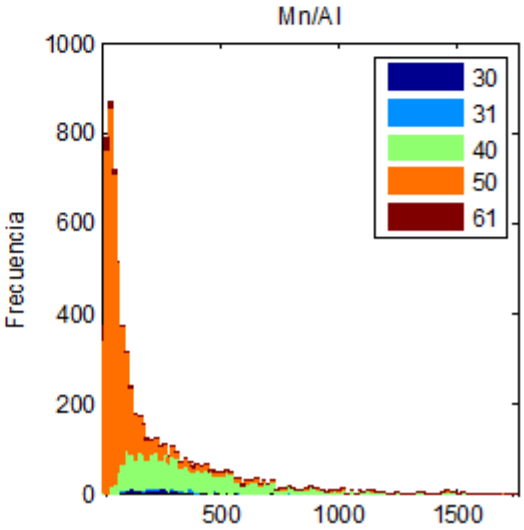

K/Mg
50
40
30
20
10
0
30
31
40
50
61


K/Mg
1400
1200
1000
800
600
400
200
0
Frecuencia
5
10
15
20
30
31
40
50
61


Mg/Al
3
2.5
2
1.5
1
0.5
0
30
31
40
50
61


Mg/Al
800
700
600
500
400
300
200
100
0
Frecuencia
0.2
0.4
0.6
0.8
30
31
40
50
61


Mn/Al
6000
5000
4000
3000
2000
1000
0
30
31
40
50
61

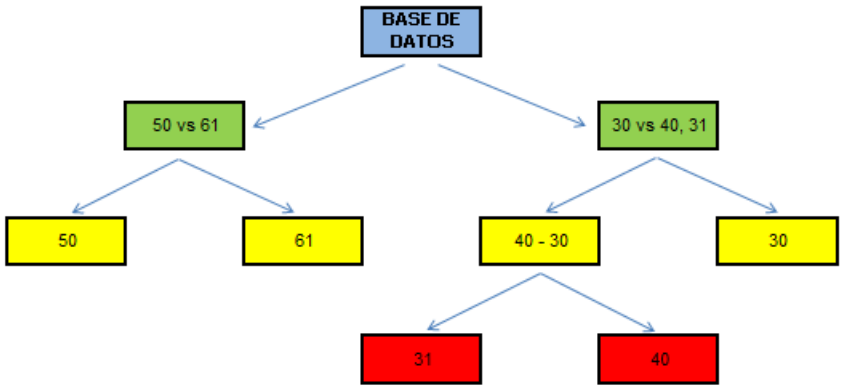

Mn/Al
1000
800
600
400
200
0
Frecuencia
500
1000
1500
30
31
40
50
61

Apéndice D

Anexos: Modelos

La sección «D» de los anexos incluye los modelos que fueron realizados para estudiar la salida del clustering de categorías y corroborar que guiándose por este no se construiría un modelo que podría ser mejorado alterando su orden de clasificación en el árbol de decisión.

A continuación se muestran modelos alternativos en donde se estudio una separación primero de la alteración 30 en vez de la alteración 31 como fue estipulado según el dendograma de categorías



SEPARACION SIMPLE

Alteración 30 vs 31 40

La variable de interés es: Rb, el cual tiene un umbral de corte mayor que: 19.9, para pertenecer a 30. Los errores que presento el modelo fueron para la base de datos de Ajuste de: 14.6 % y de entrenamiento: 9.1 %.

Alteración 31 vs 40

La variable de interés es: 3*Al/(K+Na) , el cual tiene un umbral de corte mayor que: 8.4, para pertenecer a 3'. Los errores que presento el modelo fueron para la base de datos de Ajuste de: 16.5 % y de entrenamiento: 9.1 %.

Modelo final Separación Simple

Bajo estos parámetros se obtuvo la siguiente tabla, en la cual podemos ver cómo fueron clasificados las mediciones de la base de datos de validación, teniendo un acierto de un 71 %.

	Alt. Predicha	30	31	40	50	61
Alt. Mapeada	Cantidad de Datos	564	633	1851	3160	902
30	261	81%	1%	18%	0%	0%
31	115	12%	69%	17%	2%	0%
40	2169	15%	12%	71%	1%	1%
50	4474	1%	7%	5%	70%	18%
61	91	1%	2%	10%	24%	63%
Acierto Promedio:				71%		

Cluster

Alteración 30 vs 31 40

Para crear la división de las alteraciones 50 y 61 del resto, a través de la metodología forward se utilizó la variable, con sus respectivos errores asociados al momento de su selección:

- Sc: Entrenamiento: 11.7 %, Ajuste: 16.9 %
- Ag: Entrenamiento: 7.2 %, Ajuste: 14.7 %

El algoritmo se detuvo con la variable: Sn que presento los siguientes errores: Entrenamiento: 7.2 %, Ajuste: 14.6 %

Los centroides fueron situados en la siguiente posición (Sc, Ag):

- Centroide 1: (1.67, 4.88)
- Centroide 2: (8.55, 1.92)

El error en validación final correspondió a 15 % en esta etapa.

Alteración 31 vs 40

Para crear la división de las alteraciones 50 y 61 del resto, a través de la metodología forward se utilizó la variable, con sus respectivos errores asociados al momento de su selección:

- Al: Entrenamiento: 28.0 %, Ajuste: 20.1 %

● K/Mg: Entrenamiento: 25.1 %, Ajuste: 16.3 %

El algoritmo se detuvo con la variable: Re que presento los siguientes errores: Entrenamiento: 24.5 %, Ajuste: 16 %

Los centroides fueron situados en la siguiente posición (Al, K/Mg):

● Centroide 1: (0.66, 0.99)

● Centroide 2: (1.83, 0.22)

El error en validación final correspondió a 16.92 % en esta etapa.

Modelo final K-mean clúster

Bajo estos parámetros se obtuvo la en la cual podemos ver cómo fueron clasificados las mediciones de la base de datos de validación, teniendo un acierto de un 71 %.

	Alt. Predicha	**30**	**31**	**40**	**50**	**61**
Alt. Mapeada	Cantidad de Datos	565	775	1342	3711	717
30	261	79%	3%	17%	0%	0%
31	115	3%	84%	4%	9%	1%
40	2169	16%	22%	55%	6%	2%
50	4474	0%	5%	2%	79%	14%
61	91	1%	2%	3%	34%	59%
Acierto Promedio:		71.1%				

REGRESIÓN LOGÍSTICA

Alteración 30 vs 31 40 Para crear la división de las alteraciones 50 y 61 del resto, a través de la metodología forward se utilizaron las siguientes variables, junto con sus errores asociados al momento de su selección:

● Sc: Entrenamiento: 6.35 %, Ajuste: 14.03 %

• Ba: Entrenamiento: 4.63 %, Ajuste: 13.03 %

El algoritmo se detuvo con la variable: Sb, la cual presento los siguientes errores: Entrenamiento: 4.63 %, Ajuste: 12.61 %

El modelo quedo descrito por la siguiente ecuación aproximadamente:

$$Ln[p(x)/(1-p(x))] = -9.21 + 1.44 * Sc + 0.06 * Ba$$

El modelo presenta un error de validación de 12.5 %.

Alteración 31 vs 40

Para crear la división de las alteraciones 50 y 61 del resto, a través de la metodología forward se utilizaron las siguientes variables, junto con sus errores asociados al momento de su selección:

• 3Al/(K+Na): Entrenamiento: 15.8 %, Ajuste: 17.38 %

• U: Entrenamiento: 10.86 %, Ajuste: 12.22 %

• (K+Al+S)/(Fe+S): Entrenamiento: 10.27 %, Ajuste: 11.44 %

• (Ca+Na)/(K+Al): Entrenamiento: 7.86 %, Ajuste: 10.7 %

El algoritmo se detuvo con la variable: In, la cual presento los siguientes errores: Entrenamiento: 5.45 %, Ajuste: 10.32 %. El modelo quedo descrito por la siguiente ecuación aproximadamente:

$$Ln[\frac{p(x)}{(1-p(x)}] = 21.52 - 2.05 * \frac{3 * Al}{K + Na} - 19.14 * U + 3.21 * \frac{K + Al + S}{Fe + S} + 3.11 * \frac{Ca + Na}{K + Al}$$

El modelo presenta un error de validación de 18.1 %.

Modelo final regresión logística

En la tabla se puede apreciar de manera compacta los resultados obtenidos al realizar el árbol binario de clasificación con las distintas regresiones logísticas realizadas, presentando un acierto de un 73 %.

	Alt. Predicha	30	31	40	50	61
Alt. Mapeada	Cantidad de Datos	622	467	1750	3000	1271
30	261	86%	2%	12%	0%	0%
31	115	4%	73%	17%	4%	2%
40	2169	17%	11%	68%	3%	2%
50	4474	1%	3%	5%	65%	26%
61	91	1%	0%	8%	20%	71%
Acierto Promedio:				72.5%		

Redes Neuronales

Alteración 30 vs 31 40

El modelo de RNA utilizado para la separación de las alteraciones tipo 31 del resto, consta de 4 neuronas, utilizando las siguientes variables en orden descendente de importancia, junto a ellas se encuentran los errores asociados al momento de su selección:

- Sc: Entrenamiento: 6.05 %, Tunning: 14.61 %
- Sr: Entrenamiento: 3.82 %, Tunning: 12.12 %

El algoritmo se detuvo con la variable: Al/(Na+Ca+K), la cual presento los siguientes errores: Entrenamiento: 6.71 %, Ajuste: 12.03 %1 Neuronas usadas: 3

El error Validación que presento el modelo es de 11.3 %

Modelo final redes neuronales

La tabla siguiente muestra los resultados utilizando RNA de esta clasificación que alcanza un 74 % de acierto.

	Alt. Predicha	30	31	40	50	61
Alt. Mapeada	Cantidad de Datos	655	525	2114	4531	1097
30	234	91%	1%	8%	0%	0%
31	69	15%	64%	16%	6%	0%
40	2740	15%	10%	68%	5%	2%
50	5837	0%	3%	4%	75%	18%
61	42	0%	2%	7%	17%	74%
Acierto Promedio:				74.3%		

Análisis Multivariable De Alteraciones

Santiago, 2015

Universidad de Chile

Facultad de Ciencias Físicas y Matemáticas

Departamento de Ingeniería de Minas