Rev. Inst. investig. Fac. minas metal. cienc. geogr. vol 26 n° 51, 2023: e24493 ISSN-L:1561-0888
https://doi.org/10.15381/iigeo.v26i51.24493

Artículos: Ingeniería metalúrgica

Automatización del análisis exploratorio de datos y procesamiento geoquímico univariado empleando Python

Automation of exploratory data analysis and univariate geochemical processing using Python

Brayan Jarry Castillo Requiz[1,a], Jesús Daniel Tarazona Silva[1,b], Cristian Eugenio Tarazona Silva[2,c], Christian Hurtado Enriquez[1,d], Félix Abraham Cornelio Orbegoso[1,e]

Recibido: 13/01//2023 - Aprobado: 27/03/2023 – Publicado: 02/06/2023

RESUMEN

La automatización de procesos viene siendo implementada en distintas disciplinas de las ciencias geológicas, ello se ve en el desarrollo de librerías como Pyrolite, PyGeochemCalc, dh2loop 1.0, NeuralHydrology, GeoPyTools entre otros. El presente trabajo aborda una metodología para automatizar el análisis geoquímico univariado mediante el uso de paquetes de código abierto en Python como Pandas, Seaborn, Matplotlib, Statsmodels y Scipy, las cuales serán integrados a un script en un entorno de trabajo local como Jupyter Notebook o en un entorno online como Google Colaboratory. El Script está diseñado para procesar cualquier tipo de datos geoquímicos, permitiendo remover los outliers, realizar cálculos y gráficos de los elementos con su respectivo dominio geológico. Los resultados incluyen gráficos como el box-plot, cuantil-cuantil, cálculos de las pruebas de normalidad y de los parámetros geoquímicos, lo que permite determinar el valor de fondo o background y el umbral o threshold de los elementos trabajados. El resultado de los parámetros geoquímicos será procesado posteriormente en softwares de información geográfica, la cual permite generar mapas de anomalías metálicas univariadas y de las cuencas anómalas.

Palabras claves: Análisis exploratorio de datos, análisis univariado, automatización, Python, Script.

ABSTRACT

Process automation is being implemented in different disciplines of earth sciences, as seen in the implementation of libraries such as Pyrolite, PyGeochemCalc, dh2loop 1.0, NeuralHydrology, GeoPyToo among others. The present work addresses a methodology to automate the geochemical univariate analysis by using Python and open-source packages such as pandas, seaborn, matplotlib, statsmodels which will be integrated into a script in a local work environment such as Jupyter notebook or in an online environment such as Google Collaboratory. The script is designed to process any type of geochemical data, allowing to remove outliers, perform calculations and graphs of the elements and their respective geological domain. The results include graphics such as boxplot, quantile-quantile and calculations of normality tests and geochemical parameters, allowing to determine the background and threshold of the elements worked. The result of the geochemical parameters will be further processed in geographic information software which allows to generate the univariate anomaly map and the anomalous basins.

Keywords: Exploratory data analysis, univariate analysis, automation, Python, script.

1 Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería Geológica, Minera, Metalúrgica y Geográfica, Lima, Perú.
2 Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería Eléctrica y Electrónica, Lima, Perú.
a Autor para correspondencia: brayan.castillo@unmsm.edu.pe - ORCID: https://orcid.org/0000-0003-3751-2133
b E-mail: jesus.tarazona4@unmsm.edu.pe - ORCID: https://orcid.org/0000-0001-7985-4786
c E-mail: cristian.tarazona@unmsm.edu.pe - ORCID: https://orcid.org/0000-0002-5340-0468
d E-mail: churtadoe@unmsm.edu.pe - ORCID: https://orcid.org/0000-0001-7474-2451
e E-mail: fcornelioo@unmsm.edu.pe - ORCID: https://orcid.org/0000-0001-5532-6209

I. INTRODUCCIÓN

La aplicación de la geoquímica en la exploración tiene su importancia debido al papel que tiene en la vectorización de zonas de interés económico en base a los metales principales y sus *pathfinders*, cuya naturaleza y distribución dependen del sistema mineral a prospectar. Para considerar una población de datos como anomalía se determinan los valores de *background* o valor de fondo que indica un rango de valores en lugar de un valor absoluto por la naturaleza de la concentración de los elementos en la corteza. El límite superior del valor de fondo se denomina el *threshold* o umbral y las concentraciones superiores al umbral indicarían la presencia de una anomalía geoquímica que podría representar la presencia de un potencial depósito mineral (Carranza 2009).

En los últimos 20 años el procesamiento de los datos geoquímicos viene siendo desarrollado de forma automatizada mediante softwares estadísticos; sin embargo, en la actualidad para el campo de la geología se vienen implementando librerías de código abierto en el lenguaje de *Python* como es el caso del paquete *Pyrolite* para el procesamiento litogeoquímico (Williams et al. 2020), *PyGeochemCalc* para cálculos termodinámicos (Awolayo & Tutolo 2022), *dh2loop 1.0*, que es una biblioteca para el procesamiento y la clasificación automatizados de registros geológicos (Joshi et al. 2020), *NeuralHydrology,* una librería para el análisis y cálculos hidrogelógicos (Roberts et al. 2018), *GeoPyTools,* una librería para cálculos y gráficos geológicos (Yu et al. 2019), las cuales facilitan y mejoran la eficacia en el procesamiento de grandes bases de datos.

En el presente trabajo se describe una metodología para automatizar el procesamiento del análisis exploratorio de datos y el cálculo de los parámetros geoquímicos a través de medidas de tendencia central, mediante el desarrollo de un *script* usando el lenguaje de programación de código abierto Python y librerías como *Pandas*, *Numpy, Seaborn, Matplotlib, Statsmodels* a través de ciclos y condicionales que automatizará la gestión de los elementos con sus respectivos dominios geológicos. El objetivo de la aplicación de estos métodos es reducir el tiempo de procesamiento que muchas veces corresponde a un procedimiento mecánico o rutinario y no permite al geólogo centrar sus esfuerzos y disponer de un tiempo adecuado para la conceptualización e interpretación geológica de los resultados.

II. MÉTODO

La metodología de trabajo se compone de tres etapas: preprocesamiento, automatización y la elaboración de mapas temáticos (Figura 1). En la primera etapa, se realiza la limpieza de datos mediante Excel. Una vez completada la limpieza de datos, se procede con la implementación de la automatización en el entorno web de desarrollo de código abierto Jupyter Notebook (Figura 2), que es compatible con el lenguaje de programación Python. El código desarrollado realiza el análisis exploratorio de datos, seguido del cálculo de las pruebas de normalidad y los parámetros geoquímicos filtrando los elementos y dominios geológicos para finalmente ser plasmados en mapas temáticos a través de un software de información geográfica. Para la elaboración de los mapas temáticos univariados y el mapa de las cuencas anómalas, se empleó el software ArcMap 10.8.

Por otro lado, un entorno de trabajo en programación es un software que asiste a los programadores para el desarrollo de softwares (Kerrigan et al. 2007). Algunos ejemplos de entornos de trabajo son: Eclipse, PyCharm, Visual Studio, Jupyter Lab, Jupyter Notebook, Google Colaboratory entre otros. Cada uno de estos entornos tiene sus propias características, funcionalidades específicas y buscan proporcionar un ambiente de trabajo de fácil uso para el programador.

En este proyecto, la secuencia de comandos fue desarrollada en el entorno web de Jupyter Notebook (Figura 2) debido a su particularidad de proporcionar conexiones a objetos de investigación del tipo "*open source*" como conjuntos de datos, códigos, flujos de trabajo y publicaciones que residen en otros lugares (Randles et al. 2017). La secuencia de comandos es un archivo de texto que se implementará en las celdas de código (Figura 3) que contiene una serie de instrucciones para llevar a cabo las tareas mencionadas anteriormente.

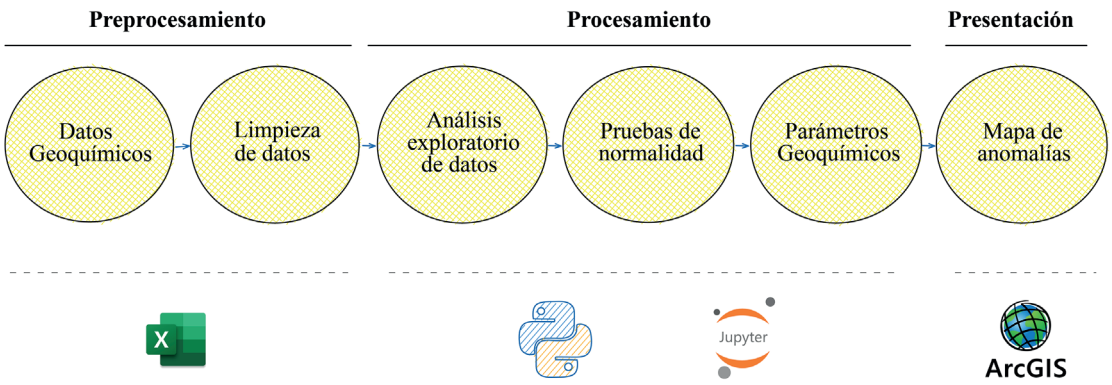

Figura 1. Secuencia de etapas del tratamiento de datos para un análisis geoquímico univariado desde el preprocesamiento, automatización y la elaboración de mapas temáticos.

Figura 2. Entorno de trabajo de Jupyter Notebook. Se visualiza el árbol de archivos que indica la ubicación de la carpeta de trabajo y el cuaderno de programación donde se implementará el código

Figura 3. Cuaderno de programación de Jupyter Notebook, donde se observa la barra de menú, la barra de herramientas y la celda de código

2.1 Datos geoquímicos

Para este estudio se usó la geoquímica de sedimentos de quebradas (*stream sediments*) del INGEMMET correspondiente al proyecto "GE-1 Prospección Geoquímica Regional entre los Paralelos 9° y 10° Latitud Sur", que está conformada por 730 muestras (Figura 3), con una densidad de muestreo promedio de 1 muestra por 10 Km^2, las muestras tienen una digestión con agua regia y un análisis ICP-MS para 52 elementos y ensayo al fuego-absorción atómica para el oro (Chira et al. 2008). La Figura 4 muestra la ubicación geográfica de las 711 muestras analizadas en este estudio, las cuales fueron recolectadas en diferentes cuencas, incluyendo Chucpín, Mosna, Torres Vizcarra, Marañón Medio y Marañón Alto.

2.2 Limpieza de datos

De acuerdo con Sahoo et al. (2019), el proceso de limpieza de datos consiste en eliminar los errores y validar los datos. En este primer paso se analizan las variables de la base de datos emitida por el laboratorio, el cual está compuesta por datos numéricos y no numéricos. Un ejemplo de estos últimos son los valores por debajo del límite de detección "< LD", los cuales representan un problema para el análisis estadístico debido a que obstaculizan el procesamiento estadístico ya que requieren la sustitución o en su defecto la eliminación de estos datos.

El proceso de limpieza de datos se llevó a cabo en Excel, donde se reemplazaron los valores no numéricos por la mitad del límite de detección inferior de los elementos correspondientes. Posteriormente, se importaron los datos al software ArcMap 10.8 para llevar a cabo la intersección entre la ubicación de las muestras y los dominios geológicos (Figura 4). Este paso permite asignar la etiqueta de dominio geológico a las muestras, la cual será empleada en los análisis posteriores.

2.3 Análisis exploratorio de datos

Posterior a la limpieza de datos, se llevó a cabo el análisis exploratorio de datos, para lo cual se emplearon los módulos *Pandas, Seaborn* y *Matplotlib;* donde *pandas*, es una librería que brinda herramientas para la manipulación de datos, a través de sus submódulos con funciones básicas que permiten leer, modificar y filtrar archivos tipo xlsx, csv, json y otros formatos de texto (Cabrera 2020). La librería de *Matplotlib* se especializa en el análisis, visualización y modelamiento de grandes conjuntos de datos (Lemenkova 2020) y *Seaborn* en generar gráficos estadísticos de alto nivel que se integra directamente con la librería de *Pandas* (Waskom 2021). De esta forma S*eaborn* amplía la biblioteca *Matplotlib* para la creación de gráficos en Python.

El proceso comienza con la importación de las librerías mencionadas en el entorno de Juyter Notebook (Figura 5). Posterior a ello se procede con la lectura del archivo Excel, para lo cual se utiliza la librería *Pandas* (Figura 5). A continuación, se procede con el análisis exploratorio de datos, que consiste en aplicar una serie de herramientas estadísticas descriptivas y gráficas que permite comprender la estructura de la base de datos, definir las variables significativas, determinar valores atípicos, sugerir y probar modelos conceptuales y finalmente identificar el mejor tratamiento e interpretación posible de los datos (Carranza 2009). Para llevar a cabo el análisis exploratorio de datos en Python, se pueden utilizar las funciones integradas *describe* y *type* (Figura 5). La función *describe* se utiliza para obtener una descripción estadística de los datos numéricos y nos proporciona datos como el número de valores no vacíos, el promedio, el valor mínimo, el percentil 25, percentil 50, percentil 75 y el valor máximo. Por su parte, la función *type* se utiliza para determinar el tipo de variable en Python. Por ejemplo, si se desea conocer el tipo de una variable llamada "cu_ppm", se puede ejecutar *type*

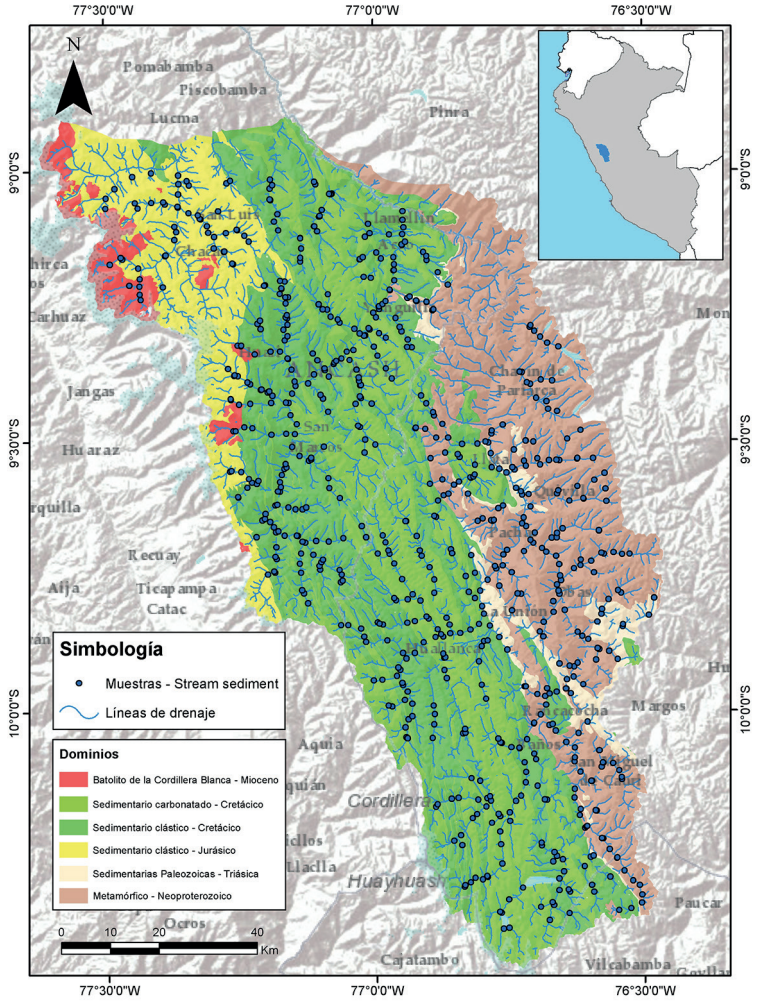

Figura 4. Mapa de los dominios geológicos que incluye la ubicación de las muestras de sedimentos de quebrada. Modificado de INGEMMET

("cu_ppm") y se mostrará si la variable es de tipo entero (*int*), decimal (*float*), una cadena de caracteres (*str*), entre otros. En resumen, *type* en Python es útil para identificar el tipo de dato que se está manipulando, siendo *int* para números enteros, *float* para números decimales y *str* para cadenas de texto y el comando *describe* proporciona información estadística descriptiva de los datos numéricos. Ambos comandos son útiles para comprender la estructura de los datos y su tipo de contenido.

Con el propósito de comparar el conjunto de datos entre poblaciones, generamos el diagrama de caja y bigotes (Box plot) mediante los módulos de *Seaborn* o *Matplotlib* (Figura 6). El diagrama de cajas y bigotes representa gráficamente los principales aspectos de una distribución de frecuencias, tales como la posición, dispersión, asimetría, longitud de las colas, puntos anómalos. Las cajas del diagrama indican los cuartiles, que corresponden al 25% y 75% de los datos, y contienen un punto o una línea horizontal que representa la mediana (50% de los datos). Además, el bigote superior se extiende hasta el valor del tercer cuartil (Q3) más 1.5 veces el rango intercuartil (RI), mientras que el bigote inferior se extiende hasta el valor del primer cuartil (Q1) menos 1.5 veces el rango intercuartil. Cualquier valor que esté fuera de estos bigotes se considera un valor atípico o *outlier* y estarán representado por puntos. Los diagramas elaborados por *Seaborn* o *Matplotlib* siguen las consideraciones mencionadas previamente.

2.4. Pruebas de normalidad

Luego del análisis exploratorio de datos, se procede a realizar las pruebas de normalidad de los mismos. Esto debido a que muchos de los procedimientos estadísticos para las pruebas paramétricas se basan en la suposición de que los datos siguen una distribución normal o gaussiana (Ghasemi & Zahediasl 2012). No existe un procedimiento estándar para determinar si una muestra sigue una distribución normal. No obstante, un procedimiento razonable para llevar a cabo esta tarea es (1) realizar una evaluación cualitativa inicial de la gráfica del histograma, ajustándose mediante una función de densidad de probabilidad normal, (2) inspeccionar un gráfico cuantil-cuantil y (3) aplicar las pruebas estadísticas pertinentes para determinar la normalidad (Petrelli 2021).

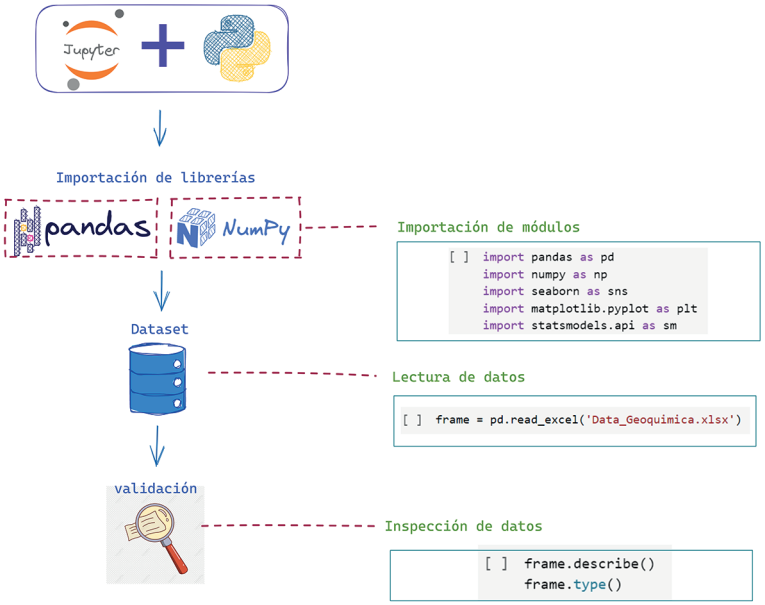

Figura 5. Secuencia de comandos para importar las librerías a Jupyter Notebook, leer un archivo Excel y el posterior análisis exploratorio de datos

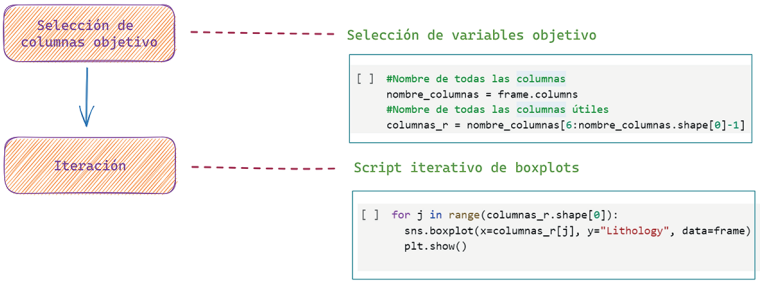

Figura 6. Secuencia de comandos para obtener el gráfico de cajas y bigotes

En el presente estudio para evaluar las pruebas de normalidad se emplea el gráfico cuantil-cuantil (*Q-Q plot*) y las pruebas de Shapiro-Wilk y Kolmogorov Smirnov con la corrección de Lillifors. El diagrama de cuantil-cuantil (*Q-Q plot*) se basa en la disposición bidimensional de los datos, , que permite comparar la distribución de frecuencias de los datos observados con respecto a un modelo de distribución normal (Alperin 2013). Si las muestras comparadas provienen de la misma distribución, entonces la gráfica debería mostrar aproximadamente una línea recta a lo largo de y = x (King & Eckersley 2019). La prueba de Shapiro-Wilk y la de Kolmogorov-Smirnov con la corrección de Lilliefors son pruebas estadísticas que se utilizan para verificar si una muestra de datos sigue una distribución normal (Shapiro & Wilk 1965; Lilliefors 1967). La hipótesis nula de ambas pruebas es que los datos siguen una distribución normal.

El código nos dará como resultado el p valor de la prueba, el cual es la máxima probabilidad de rechazar la hipótesis nula cuando es verdadera. Se utiliza α para indicar el nivel de significancia (Alperin 2013). Si el valor p obtenido es menor que el nivel de significancia establecido (por ejemplo, α=0.05), se rechaza la hipótesis nula y se concluye que los datos no siguen una distribución normal.

Es importante destacar que las librerías de Python incluyen los fundamentos matemáticos para realizar estos análisis. Las librerías usadas y el flujo del trabajo del código para realizar el gráfico cuantil-cuantil, las pruebas de normalidad de Shapiro-Wilk y Kolmogorov Smirnov con la corrección de Lillifors están detalladas en la Figura 7, Figura 8 y Figura 9.

```python
import statsmodels.api as sm
import matplotlib.pyplot as plt

for i in range(dominio.shape[0]):
    filtro = data_elementos[(data_elementos['DOMINIO GEO'] == dominio[i])]
    for j in range(columnas_e.shape[0]):
      sm.qqplot(filtro[columnas_e[j]],fit = True, line='45')
      plt.title("Q-Q plot "+str(dominio[i])+" "+str(columnas_e[j]))
plt.close()
```

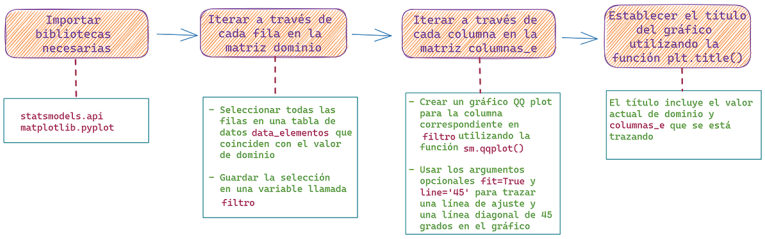

Figura 7. Secuencia de comandos para obtener el gráfico cuantil-cuantil

```python
from scipy.stats import shapiro


for i in range(dominio.shape[0]):
    filtro = data_elementos[(data_elementos['DOMINIO GEO'] == dominio[i])]
    for j in range(columnas_e.shape[0]):
      stat, pval = shapiro(filtro[columnas_e[j]])
      if pval > 0.05:
        print('It is normally distributed')
      else:
        print('It is not normally distributed')
```

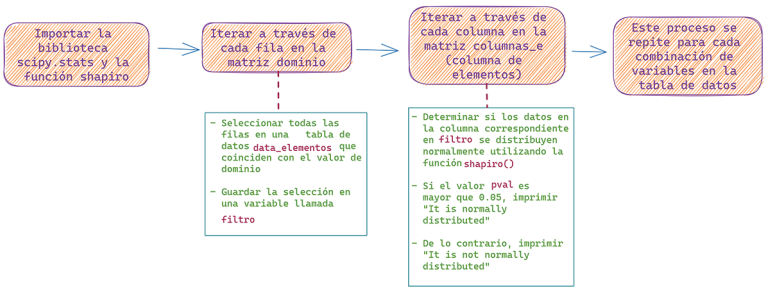

Figura 8. Secuencia de comandos para la prueba de Shapiro-Wilk

```python
import statsmodels.api as sm

for i in range(dominio.shape[0]):
    filtro = data_elementos[(data_elementos['DOMINIO GEO'] == dominio[i])]
    for j in range(columnas_e.shape[0]):
      #Compared with standard normal distribution
      ksl_stat,ksl_p_value=sm.stats.diagnostic.lilliefors(filtro[columnas_e[j]],'norm',pvalmethod='approx')
      if ksl_p_value > 0.05:
        s1 = 'It is normally distributed.'
      else:
        s1 = 'It is not normally distributed'
      lista_ksl = [dominio[i], columnas_e[j], ksl_p_value, s1]
```

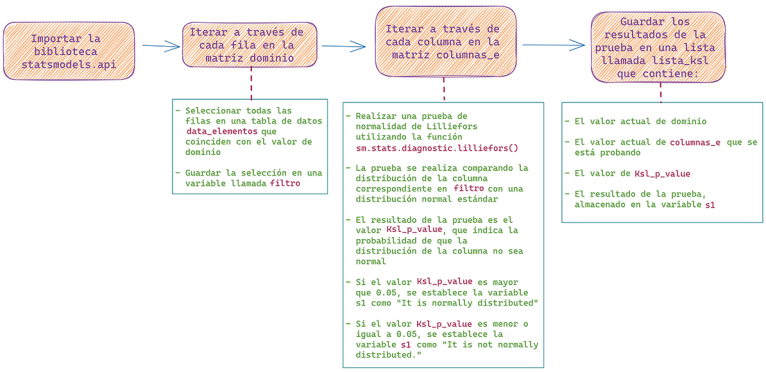

Figura 9. Secuencia de comandos para la prueba de Kolmogorov-Smirnov con la corrección de Lillifors

Posterior a las pruebas de normalidad, aquellos elementos que no pasen la prueba son sometidos a una segunda iteración en la cual se inicia la eliminación de valores atípicos. La remoción de *outliers* es un subproceso dentro del análisis de la normalidad de los datos que se realiza con el fin de ver la distribución sin la presencia de ellos. Para este análisis se toma como base teórica el proceso de construcción de un diagrama de caja y bigotes (ver Figura 10) los cuales representan el límite de los datos atípicos con respecto al resto de la población.

Es a partir de este fundamento que el código remueve los datos atípicos automáticamente y realiza el análisis posterior sin la presencia de ellos.

Posterior a la remoción de los valores atípicos se realiza nuevamente el cálculo de las pruebas de bondad, dando como resultado un nuevo conjunto de elementos que satisfacen las pruebas de normalidad. Para aquellos elementos que no pasen estas pruebas se recomienda usar métodos de estadística no paramétrica o buscar maneras alternas de trabajar los geoquímicos univariados como el índice de enriquecimiento relativo local – LREI (Zuo 2014), los cuales no dependen de la suposición de modelos gaussianos.

2.5 Parámetros geoquímicos

Finalmente, el código realiza el cálculo de los parámetros geoquímicos de los elementos que han pasado las pruebas de normalidad. Es importante tener en cuenta que los parámetros geoquímicos representan los niveles de concentración que diferencian los valores de fondo con los valores que pueden ser considerados anómalos. El concepto de anomalía puede ser complejo y su reconocimiento dependerá de un valor de fondo establecido como referencia para determinar anomalías geoquímicas (Grunsky 2010). Por ejemplo (Hawkes & Webb 1962; Howarth 1983) definieron a este umbral como la media de 2 desviaciones estándar mientras que (Filzmoser & Hron 2008; Zuo 2014) enfocan el análisis de estos umbrales a partir de métodos robustos y técnicas adaptativas.

Con el cálculo de los parámetros geoquímicos correspondientes finaliza la segunda iteración. El flujo de trabajo del código, incluyendo las iteraciones y pasos previamente mencionados, se puede apreciar en detalle en la Figura 11. Además, la Figura 12 muestra los parámetros que se utilizaron en este estudio para determinar las anomalías geoquímicas para lo cual se empleó la media y la desviación estándar.

2.6 Mapas temáticos

Una vez completadas las pruebas de normalidad y el cálculo de los parámetros geoquímicos obtenidos en el paso anterior, se procede a elaborar los mapas temáticos. Para ello, se utiliza el *software* ArcGIS 10.8, que permite ubicar espacialmente las muestras y asignar un atributo de color y tamaño específico que corresponde a los valores de cada parámetro (Figura 12). Además, se generaron cuencas hidrográficas a partir de un modelo digital del terreno (DEM) a una escala de 1000 hectáreas, las cuales serán asignados con un atributo de color tomando en cuenta los parámetros de la Figura 12, que nos permitirá identificar las principales cuencas anómalas.

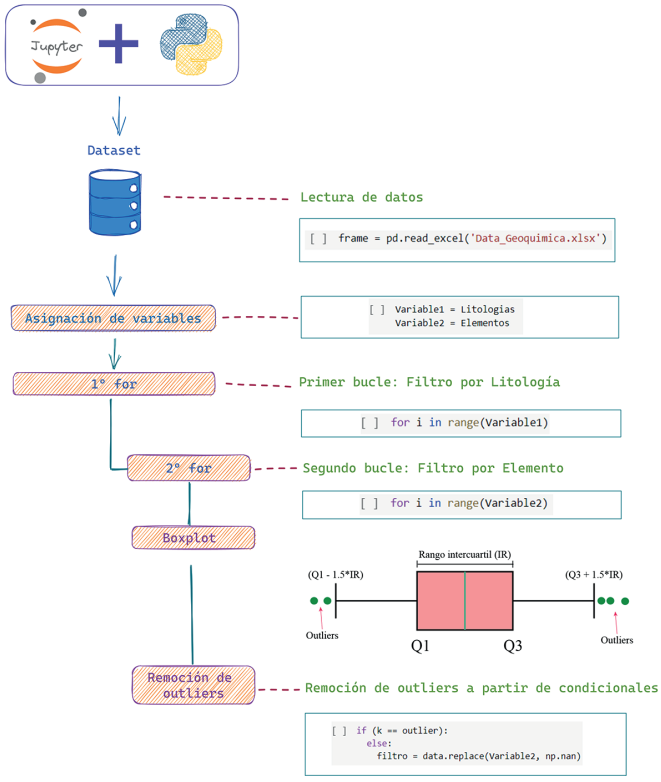

Figura 10. Secuencia de comandos para la remoción de los valores atípicos

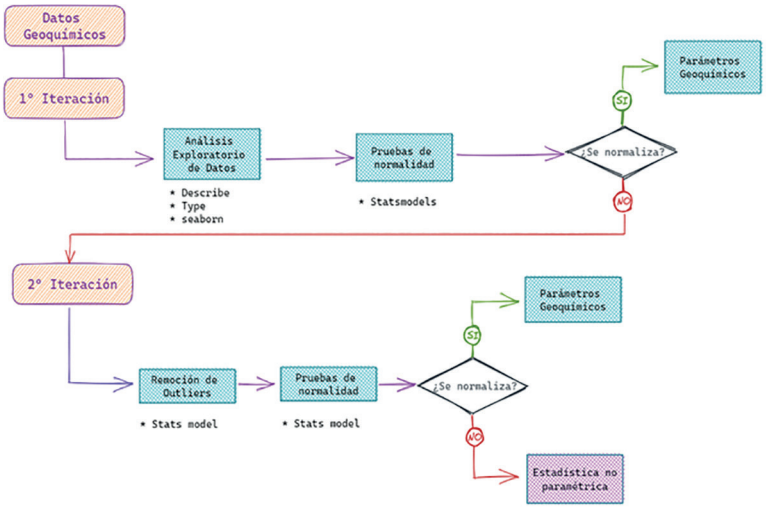

Figura 11. Flujo de trabajo del proceso de automatización para el análisis geoquímico univariado

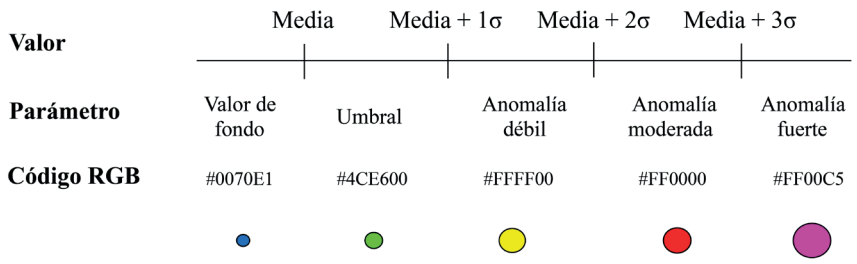

Figura 12. Parámetros geoquímicos calculados a partir de la media y desviación estándar (σ) y los atributos de color RGB y tamaño que se empleará en la elaboración de los mapas temáticos

III. RESULTADOS

Los resultados del presente estudio se representan de la siguiente manera:

Los resultados obtenidos de la selección de dominios litológicos (Tabla 1) y el resumen estadístico con los valores necesarios para el análisis estadístico de datos del comando *describe* se muestra en la Tabla 2.

Los valores obtenidos del comando describen muestras de los principales estadísticos de tendencia central para la suite de los siguientes elementos: Ag, As, Au, Ba, Be, Bi, Ca, Cd, Ce. Co, Cr, Cu, Fe, Ga, K, Mg, Mn, Mo, Ni, P, Pb, Sb, Sn, Sr, Tb, Th, U, V, Y, Yb, Zn calculando sus valores valor mínimo, máximo, media y mediana.

Tabla 1. Resumen del número de muestra por cada dominio litológico

Dominio	Número de muestras
Sedimentario clástico - Cretácico	262
Metamórfico - Neoproterozoico	192
Sedimentario carbonatado - Cretácico	156
Sedimentario clástico - Jurásico	59
Sedimentario Paleozoico - Triásico	28
Batolito de la Cordillera Blanca - Mioceno	14

Tabla 2. Resumen de estadísticos descriptivos obtenidos con describe

N = 711	Unidad	Mínimo	Q1	Median	Q3	Máximo	Mean
Ag	ppm	0.005	0.06	0.09	0.16	10	0.246
As	ppm	1	8	14	26	633	25.051
Au	ppb	2.5	2.5	5	9	456	13.052
Ba	ppm	9	68	108	174	3916	171.391
Be	ppm	0.2	0.4	0.6	0.7	3.3	0.624
Bi	ppm	0.02	0.16	0.22	0.34	10.5	0.334
Ca	%	0.03	0.23	0.69	5.72	15	3.655
Cd	ppm	0.005	0.18	0.37	0.78	45.2	0.813
Ce	ppm	1.69	12.6	22.6	40.5	179.2	30.557
Co	ppm	1.2	8.8	12	17.3	150.4	14.782
Cr	ppm	0.5	8	13	21	265	18.461
Cu	ppm	3.3	15.5	23.4	34	1038	29.509
Fe	%	0.45	2.08	2.71	3.65	15	3.018
Ga	ppm	0.5	1.7	2.7	4.1	14	3.219
K	%	0.02	0.08	0.11	0.15	0.42	0.122
Mg	%	0.03	0.17	0.36	0.7	3.84	0.524
Mn	ppm	78	480	640	934	10000	861.315
Mo	ppm	0.43	1.01	1.36	2.26	120.9	3.571
Ni	ppm	4.2	16.8	24.4	34.5	128.6	28.914
P	ppm	65	388	605	851	2721	679.909
Pb	ppm	3.7	16.5	22.8	32.4	8685	49.077
Sb	ppm	0.12	0.57	0.92	1.87	85.2	2.061
Sn	ppm	0.15	0.6	0.8	1.3	105.3	1.358
Sr	ppm	2.8	14.2	24.3	64.3	684.9	59.96
Tb	ppm	0.08	0.29	0.4	0.54	1.5	0.439
Th	ppm	0.5	2.4	4	7	27.4	5.22
U	ppm	0.13	0.5	0.78	1.11	28.8	1.222
V	ppm	1	14	23	38.75	424	31.244
Y	ppm	1.54	5.51	8.67	11.9	30.1	9.11
Yb	ppm	0.1	0.4	0.7	0.9	2.1	0.712
Zn	ppm	8	73	104	164	3120	169.868

La primera serie de diagramas de cajas y bigotes (*box plot*) generados a partir de los comandos *seaborn, matplotlib y statmodels,* muestran la distribución de datos sin la remoción de *outliers.* En una segunda iteración, el *script* elimina los valores atípicos (Figura 10) y proporciona un nuevo diagrama de cajas y bigotes, los resultados del diagrama de la primera y segunda iteración se muestran en la Figura 13.

Los resultados obtenidos mediante el uso del paquete de *statsmodels* para evaluar si la distribución de los datos sigue una distribución normal a través de los diagramas cuantil-cuantil o *Q-Q plot* se muestran en la Figura 14. Se presentan los resultados de las poblaciones geoquímicas de los sedimentos clástico del cretácico y Metamórfico-Neoproterozico para los elementos de Cu, Pb y sus respectivos valores logarítmicos.

Los diagramas Q-Q plot (Figura 14) permiten observar que para la población de sedimentarios clástico - cretácico la distribución de las concentraciones de Cu y Pb se ajustan mejor a una tendencia de 45° en sus valores logarítmicos; sin embargo para la población Metamórfico – Neoproterozoicos, ambos elementos se ajustan tanto a la distribución normal y lognormal. Esta diferencia entre el tipo de distribución de un dominio litológico y otro puede ser explicada por la naturaleza de la roca misma y su distribución espacial y sirven para calcular los parámetros geoquímicos utilizando los estadísticos adecuados para cada población y elemento.

Posterior a la generación de los gráficos cuantil-cuantil, el código proporciona el resultado de las pruebas de normalidad. Esta prueba se realiza en dos iteraciones: la primera iteración con la data inicial (datos de entrada) y la segunda con la data de los *outliers* removidos (datos procesados), de esta forma se agregan elementos que pasan la prueba de normalidad (Tabla 3).

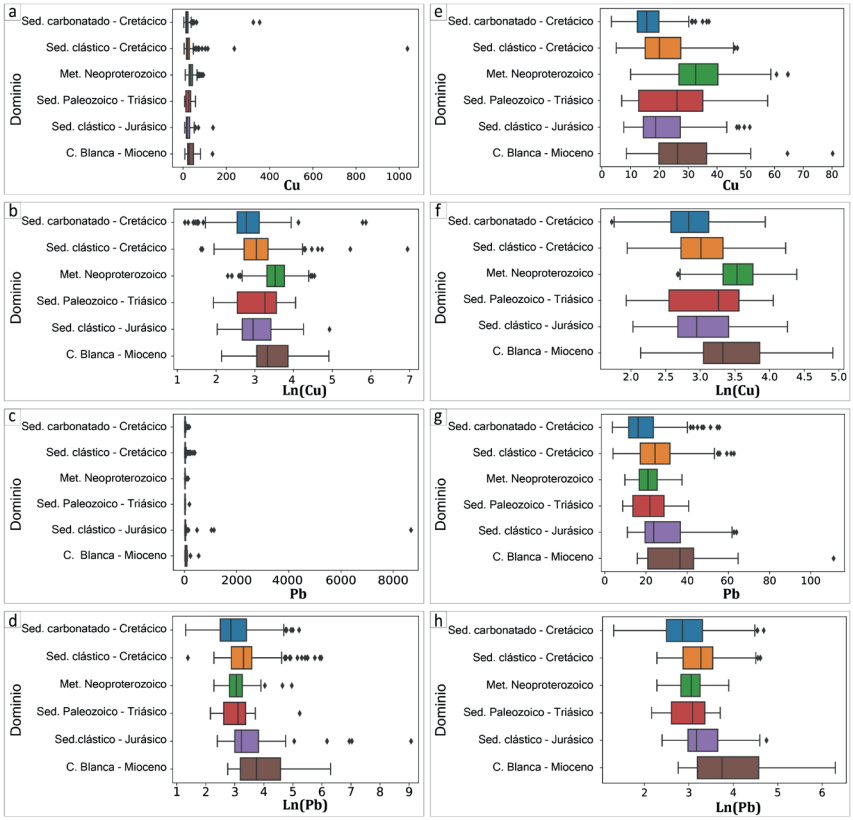

Figura 13. Comparación de los diagramas de cajas y bigotes entre la primera iteración (a-d) y la segunda iteración (e-h)

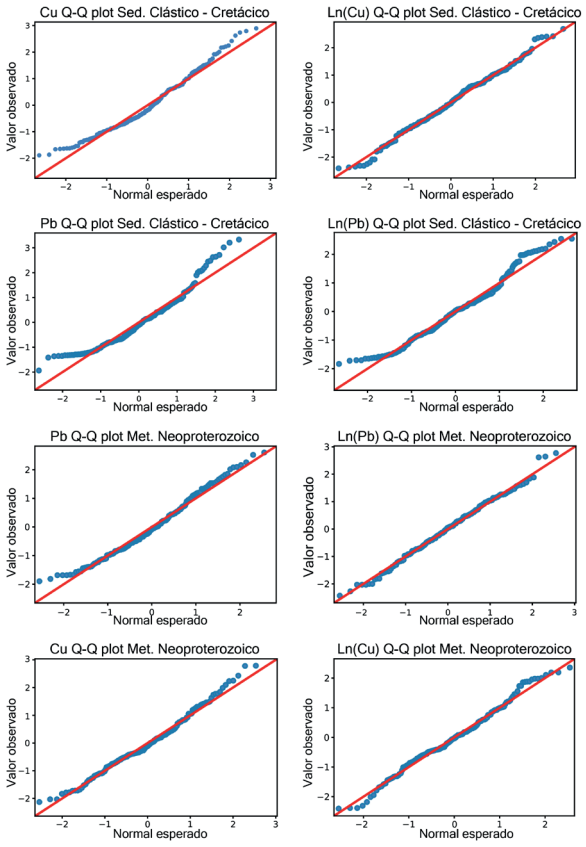

Figura 14. Diagrama cuantil-cuantil para el dominio sedimentario clástico - cretácico y para el dominio Metamórfico – Neoproterozoico de los datos con outliers removidos

Tabla 3. Resultados de las pruebas de normalidad para Cu, Pb, Ln(Cu) y Ln(Pb) después de remover los valores atípicos, donde Ln representa los logaritmos neperianos

Elemento	Dominio	n	Test de normalidad					
			Kolmogorov Smirnov - Lillifors			Shapiro-Wilk		
			Estadístico	pval	Distribución	Estadístico	pval	Distribución
Cu	Sedimentario clástico - Cretácico	262	0.08497	0.00017	-	0.96980	0.00004	-
	Metamórfico - Neoproterozoico	192	0.06869	0.04080	-	0.98627	0.08122	Normal
	Sedimentario carbonatado - Cretácico	156	0.08943	0.00736	-	0.96831	0.00223	-
	Sedimentario clástico - Jurásico	59	0.20175	0.00001	-	0.85624	0.00001	-
	Sedimentario Paleozoico - Triásico	28	0.12312	0.34414	Normal	0.95216	0.22441	Normal
	Batolito de la Cordillera Blanca - Mioceno	14	0.21967	0.08646	Normal	0.88428	0.08151	Normal
Pb	Sedimentario clástico - Cretácico	262	0.07805	0.00189	-	0.94271	0.00000	-
	Metamórfico - Neoproterozoico	192	0.06033	0.09603	Normal	0.97982	0.00862	-
	Sedimentario carbonatado - Cretácico	156	0.15566	0.00000	-	0.87286	0.00000	-
	Sedimentario clástico - Jurásico	59	0.20814	0.00001	-	0.85964	0.00002	-
	Sedimentario Paleozoico - Triásico	28	0.11088	0.52801	Normal	0.94938	0.20704	Normal
	Batolito de la Cordillera Blanca - Mioceno	14	0.28416	0.01332	-	0.80087	0.00966	-
Ln(Cu)	Sedimentario clástico - Cretácico	262	0.04460	0.31890	Normal	0.99352	0.34492	Normal
	Metamórfico - Neoproterozoico	192	0.04388	0.57167	Normal	0.98879	0.15838	Normal
	Sedimentario carbonatado - Cretácico	156	0.08229	0.01908	-	0.98051	0.03930	-
	Sedimentario clástico - Jurásico	59	0.12806	0.01899	-	0.95465	0.02989	-
	Sedimentario Paleozoico - Triásico	28	0.18752	0.01289	-	0.93487	0.08188	Normal
	Batolito de la Cordillera Blanca - Mioceno	14	0.13354	0.70542	Normal	0.96637	0.82473	Normal
Ln(Pb)	Sedimentario clástico - Cretácico	262	0.06303	0.01882	-	0.97154	0.00007	-
	Metamórfico - Neoproterozoico	192	0.02794	0.98024	Normal	0.99412	0.66044	Normal
	Sedimentario carbonatado - Cretácico	156	0.09128	0.00367	-	0.97267	0.00418	-
	Sedimentario clástico - Jurásico	59	0.14485	0.00646	-	0.94157	0.01085	-
	Sedimentario Paleozoico - Triásico	28	0.12427	0.35630	Normal	0.95923	0.35496	Normal
	Batolito de la Cordillera Blanca - Mioceno	14	0.23314	0.03776	-	0.89881	0.10838	Normal

Finalmente, con los elementos que pasaron las pruebas de normalidad se calcularon los parámetros geoquímicos en base a la media y la desviación estándar, estos valores sirven para la elaboración de los mapas de anomalías univariadas y sus cuencas anómalas en el software ArcGIS 10.8 las cuales se muestran a continuación (Figura 15 y 16).

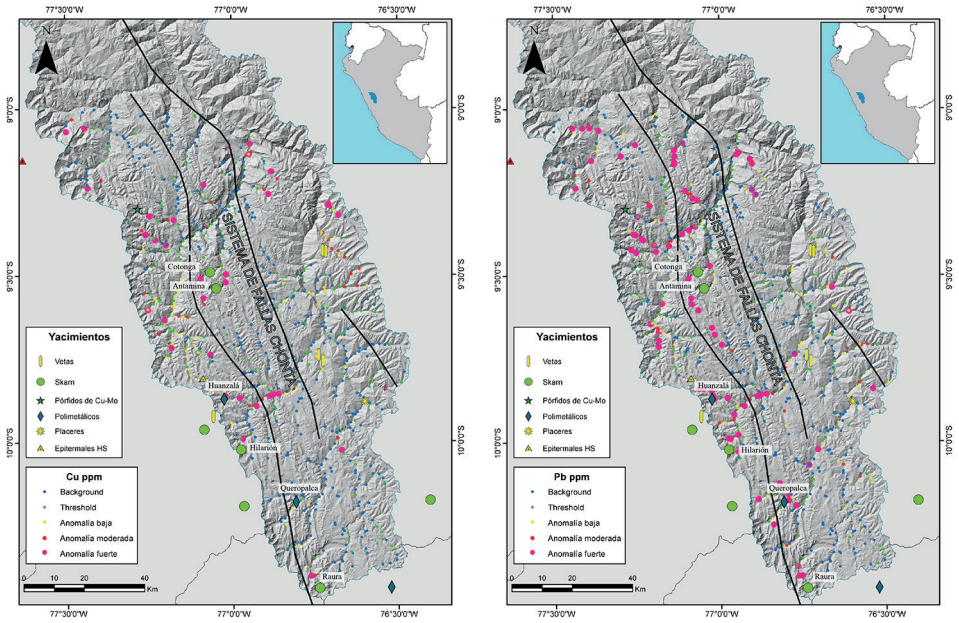

Figura 15. Mapas temáticos el Cu y Pb entre los paralelos 9° - 10° Latitud Sur en la vertiente atlántica

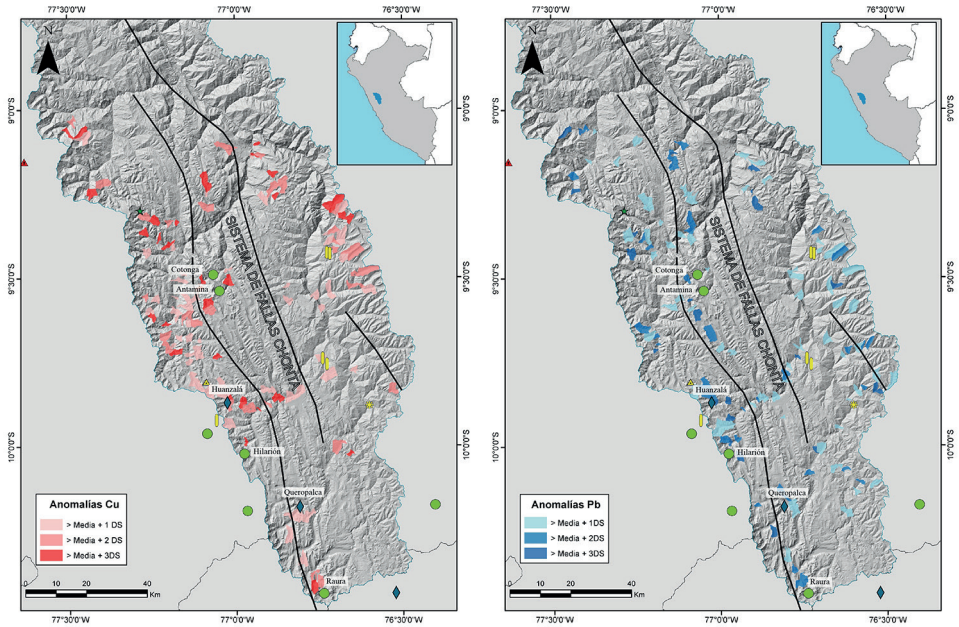

Figura 16. Delimitación de las cuencas anómalas de Cu y Pb entre los paralelos 9° - 10° Latitud Sur en la vertiente atlántica

Los mapas univariados de Cu y Pb se construyeron a partir de los parámetros geoquímicos, donde las áreas anómalas se basan en la media y desviación estándar detalladas en la Figura 12. Los resultados muestran una buena correspondencia de anomalías moderadas a fuertes de Cu y Pb en el sector norte del distrito de San Marcos, cercana a los yacimientos de Cotonga y Antamina. Otro sector donde se han detectado anomalías, que van desde moderadas a fuertes de Cu y Pb es en las quebradas adyacentes de los yacimientos de Huanzalá e Hilarión, siguiendo la dirección de drenaje SW-NE. Al sur del área de estudio se observan anomalías fuertes de Cu y Pb proximales al yacimiento de Raura. El Cu y Pb se extienden predominantemente en el sector occidental del Sistema de Fallas Chonta (SFCH).

Las cuencas anómalas sectorizan 68 cuencas con anomalías fuertes de Cu (valores mayores a la media más 3 veces la desviación estándar) y 77 cuencas con anomalías moderadas de Cu (valores mayores a la media más 2 veces

la desviación estándar) que pueden ser posibles blancos de exploración. Con respecto a las cuencas anómalas para Pb se identificaron 35 cuencas con anomalías fuertes y 54 cuencas con anomalías moderadas.

IV. DISCUSIÓN

El análisis geoquímico univariado suele ser dividido en tres etapas que todo geólogo debe llevar a cabo, independientemente de la metodología a usar (1) preprocesamiento, en el cual se hace una curación de los datos (2) procesamiento, que consiste en los cálculos correspondientes al análisis exploratorio de datos y los cálculos de parámetros geoquímicos y (3) presentación, en la que se elaboran los mapas de anomalías en base los resultados de la etapa anterior.

Sobre el preprocesamiento realizado para el tratamiento de valores no numéricos o censurados es importante resaltar que esta etapa puede ser incluida con un procesamiento automatizado en Python usando las librerías *Pandas* y *Numpy*, sin embargo, no fueron abordadas en este trabajo y se optó por un preprocesamiento en Excel para levantar esos errores previos al procesamiento en Python.

En relación con el procesamiento automatizado, la validez del proceso y de los cálculos estadísticos en Python realizados de las medidas de posición, medidas de dispersión, construcción de diagramas y pruebas de normalidad, el presente trabajo se basa en los fundamentos matemáticos que todos los programas de tratamiento estadístico tienen incorporados. Estos fundamentos pueden ser revisados en las respectivas documentaciones de las librerías utilizadas, tales como *Seaborn* (https://seaborn.pydata.org/), *Matplotlib* (https://matplotlib.org/) y Statsmodels (https://www.statsmodels.org/). En estas documentaciones se pueden encontrar descripciones detalladas de las funciones utilizadas en el trabajo, información sobre los parámetros de entrada y salida requeridos, así como ejemplos de uso.

De los resultados obtenidos a partir de la segunda iteración, las poblaciones que no pasen las pruebas de normalidad, incluso con los valores atípicos removidos, probablemente tengan distribución de datos distinta a la normal o log normal por la naturaleza misma de los datos geológicos y se sugiere probar otros tipos de transformaciones que se ajusten a su distribución (por ejemplo distribución Poisson, distribución Exponencial, distribución Weibull, distribución Fisher, etc.) o en su defecto usar técnicas robustas como el índice de enriquecimiento relativo local – LREI (Zuo 2014) las cuales pueden ser automatizadas pero no forman parte de este trabajo.

Sobre la presentación de los mapas se utilizan los datos exportados de la etapa de procesamiento en Python y son elaborados el software ArcGIS, dado que es un proceso que no requiere automatización de cálculos y construcción de diagramas para cada población estadística.

Finalmente, es posible integrar los resultados de las anomalías metálicas y las cuencas anómalas identificadas y a través de superposición ponderada de ráster con datos de (1) litología favorable, (2) geología estructural, (3) alteraciones hidrotermales (4) anomalías geofísicas y generar un mapa de prospectividad mineral soportado en un enfoque multidisciplinario.

V. CONCLUSIONES

La novedad de este trabajo consiste en automatizar la segunda etapa del análisis univariado (Figura 1) usando lenguaje de programación y librerías de acceso libre para trabajar con grandes bases de datos y un gran número de dominios litológicos, no se realizan cálculos ni algoritmos nuevos que requieran ser comprobados, solo se usa el lenguaje de programación utilizando ciclos y condicionales para automatizar el tiempo del filtrado de datos para cada dominio litológico y cada elemento geoquímico, siendo esta etapa el factor diferencial en el tiempo de procesamiento mecánico.

El código puede ser reutilizable e implementado para cualquier conjunto de datos que requieran un filtrado por categorías de dominios litológicos y de elementos, solo serán necesarios ajustar los parámetros de lectura de datos y renombramiento de etiquetas con las filas y columnas útiles para los cálculos.

Las cuencas anómalas de Cu y Pb presentadas como ejemplo de este trabajo de automatización resaltan zonas asociadas a mineralización adyacente a los yacimientos (Cotonga, Antamina, Huanzalá, Hilarión y Raura).

VI. AGRADECIMIENTOS

Los autores agradecemos al Msc. Giovanni Pedemonte Castro por sus sugerencias durante el desarrollo de este trabajo.

VII. REFERENCIA

Alperin, M. (2013). *Introducción al análisis estadístico de datos geológicos*. Editorial de la Universidad Nacional de La Plata (EDULP).https://doi.org/10.35537/10915/3422

Awolayo, A. N., & Tutolo, B. M. (2022). *PyGeochemCalc: A Python package for geochemical thermodynamic calculations from ambient to deep Earth conditions.* Chemical Geology, 606, 120984. https://doi.org/https://doi.org/10.1016/j.chemgeo.2022.120984

Cabrera, S. (2020). *Magmatismo carbonífero–triásico medio (20-31° S): anomalías de la costa, segmentación y relación con la tectónica*. [Tesis de Licenciatura]. Universidad de Concepción. http://repositorio.udec.cl/xmlui/handle/11594/6318

Carranza, E. J. M. (2009). *Predictive modeling of mineral exploration targets. Handbook of exploration and environmental geochemistry*, 11, 3-21. https://doi.org/10.1016/S1874-2734(09)70005-1

Chira, J.; Gonzáles, R.; Vargas, L.; Rivera, R.; Chero, D. & Guerra, K. (2008) - *Prospección geoquímica regional entre los paralelos 9° - 10° Latitud Sur* (Vertiente Atlántica).

INGEMMET. Boletín, Serie B: Geología Económica, 18, 109 p., 4 mapas. https://hdl.handle.net/20.500.12544/207

Filzmoser, P., & Hron, K. (2008). *Outlier Detection for Compositional Data Using Robust Methods. Mathematical Geosciences,* 40(3), 233–248. https://doi.org/10.1007/s11004-007-9141-5

Ghasemi, A., & Zahediasl, S. (2012). *Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. International Journal of Endocrinology and Metabolism*, 10(2), 486–489. https://doi.org/10.5812/ijem.3505

Grunsky, E. C. (2010). *The interpretation of geochemical survey data. Geochemistry: Exploration, Environment, Analysis,* 10(1), 27–74. https://doi.org/10.1144/1467-7873/09-210

Hawkes, H., & Webb, J. (1962). *Geochemistry in Mineral Exploration*. https://journals.lww.com/soilsci/Citation/1963/04000/Geochemistry_in_Mineral_Exploration.16.aspx

Howarth, R. (1983). *Statistics and data analysis in geochemical prospecting. Handbook of Exploration Geochemistry*, Elsevier. https://www.elsevier.com/books/statistics-and-data-analysis-in-geochemical-prospecting/howarth/978-0-444-42038-1

Joshi, R., Madaiah, K., Jessell, M., Lindsay, M., & Pirot, G. (2021). *dh2loop 1.0: an open-source Python library for automated processing and classification of geological logs. Geoscientific Model Development*, 14(11), 6711–6740. https://doi.org/10.5194/gmd-14-6711-2021

Kerrigan, M., Mocan, A., Tanler, M., & Fensel, D. (2007). *The Web Service Modeling Toolkit - An Integrated Development Environment for Semantic Web Services. In E. Franconi*, M. Kifer, & W. May (Eds.), The Semantic Web: Research and Applications (pp. 789–798). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72667-8_57

King, A. P., & Eckersley, R. J. (2019). *Inferential Statistics IV: Choosing a Hypothesis Test. In Statistics for Biomedical Engineers and Scientists* (pp. 147–171). Elsevier. https://doi.org/10.1016/B978-0-08-102939-8.00016-5

Lemenkova, P. (2020). *Python Libraries Matplotlib, Seaborn and Pandas for Visualization Geo-spatial Datasets Generated by QGIS Analele stiintifice ale Universitatii "Alexandru Ioan Cuza" din Iasi* - seria Geografie, vol. 64(1), pp. 13-32, 2020, Available at SSRN: https://ssrn.com/abstract=3699706

Lilliefors, H. W. (1967). *On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. Journal of the American Statistical Association,* 62(318), 399. https://doi.org/10.2307/2283970

Petrelli, M. (2021). *Introduction to Python in Earth Science Data Analysis*. Springer International Publishing. https://doi.org/10.1007/978-3-030-78055-5

Randles, B. M., Pasquetto, I. v, Golshan, M. S., & Borgman, C. L. (2017). *Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study*. 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 1–2. https://doi.org/10.1109/JCDL.2017.7991618

Roberts, W., Williams, G. P., Jackson, E., Nelson, E. J., & Ames, D. P. (2018). *Hydrostats: A Python Package for Characterizing Errors between Observed and Predicted Time Series*. Hydrology, 5(4). https://doi.org/10.3390/hydrology5040066

Sahoo, K., Samal, A., Pramanik, J., Pani, S. (2013). *Exploratory data analysis using Python. International Journal of Innovative Technology and Exploring Engineering* (IJITEE), vol. 8, no 12, https://www.ijitee.org/wp-content/uploads/papers/v8i12/L35911081219.pdf

Shapiro, S. S., & Wilk, M. B. (1965). *An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52(3/4), 591*. https://doi.org/10.2307/2333709

Waskom, M. (2021). *Seaborn: statistical data visualization. Journal of Open Source Software*, 6(60), 3021. https://doi.org/10.21105/joss.03021

Williams, M., Schoneveld, L., Mao, Y., Klump, J., Gosses, J., Dalton, H., Bath, A., & Barnes, S. (2020). *Pyrolite: Python for geochemistry. Journal of Open Source Software*, 5(50), 2314. https://doi.org/10.21105/joss.02314

Yu, Q.-Y., Bagas, L., Yang, P.-H., & Zhang, D. (2019). *GeoPyTool: A cross-platform software solution for common geological calculations and plots. Geoscience Frontiers*, 10(4), 1437–1447. https://doi.org/https://doi.org/10.1016/j.gsf.2018.08.001

Zuo, R. (2014). *Identification of weak geochemical anomalies using robust neighborhood statistics coupled with GIS in covered areas*. Journal of Geochemical Exploration, 136, 93–101. https://doi.org/https://doi.org/10.1016/j.gexplo.2013.10.011

Contribución de autoría

Conceptualización: B.C. & J.T.; Curación de datos: J.T.; Análisis Formal: C.T.; Investigación: B.C. & J.T.; Metodología: B.C.; Administración del proyecto: B.C. & J.T.; Recursos: B.C., J.T. & C.T.; Software: C.T.; Supervisión: C.H. & F.C.; Validación – Verificación: B.C., J.T., C.H. & F.C.; Visualización: B.C. & J.T.; Redacción - borrador original: B.C., J.T. & C.T.; Redacción - revisión y edición: C.H. & F.C.

Conflicto de intereses

Los autores declaran no tener conflictos de intereses