

Geoquímica de rocas a escala mundial

Estadística y machine learning sobre 159.760 análisis de roca total de EarthChem PetDB

Informe del proyecto: la base de datos, la estadística descriptiva y la documentación técnica de cómo se construyó todo, incluida la página web.

Universidad Central del Ecuador

Facultad de Ingeniería en Geología, Minas, Petróleos y Ambiental (FIGEMPA) · Carrera de Ingeniería en Minas

Asignatura: Estadística y Probabilidades · Docente: Dr. Christian Mejía E., Ph.D.

Integrantes

Cristoffer Joel Rumiguano Chariguaman

Ana Gabriela Fernández Chuquimarca

Evelyn Maritza Bustamante Armijos

Jessica Jaqueline Haro Simba

Sitio web: <https://crisrumiguano-art.github.io/geoquimica-earthchem-ml/>

Repositorio: <https://github.com/crisrumiguano-art/geoquimica-earthchem-ml>

Quito, octubre de 2026

Contenido

Guía rápida para el equipo	3
Dónde está cada cosa	3
Qué tiene la página web	3
Estado por lección	3
Cinco ideas clave para la exposición	4
Cómo trabajar con los archivos en tu computadora	4
Glosario	4
Parte I. La base de datos	5
1. Qué es esta base	5
2. De dónde salieron los datos	5
3. Cómo se preparó	5
4. Variables	6
5. Requisitos del trabajo	7
6. Limitaciones que conviene conocer	7
7. Acceso y cita	7
Parte II. Estadística descriptiva	8
8. Por qué no sirvió la agrupación por defecto	8
9. Outliers: otras rocas, no errores	8
10. ¿Alcanzan los datos para las regresiones?	9
Parte III. Cómo se construyó: documentación técnica	11
11. Arquitectura general	11
12. Herramientas y librerías	11
13. Limpieza de la base (scripts/limpiar.py)	11
14. Mapa de ubicación (Python y QGIS)	12
15. Estadística en R (carpeta R/)	13
16. La página web	13
17. Publicación y control de versiones	15
18. Documentación del proyecto	15
19. Cómo reproducir todo	16
20. Limitaciones técnicas y mantenimiento	16
Anexo A. Diccionario de variables	17
Anexo B. Estructura del repositorio	18

Guía rápida para el equipo

Esta guía resume, en pocas páginas, qué está hecho, dónde está cada cosa y qué hay que saber para la exposición. El resto del informe explica todo con detalle.

Dónde está cada cosa

Qué	Dónde
Página web del proyecto	https://crisrumiguano-art.github.io/geoquimica-earthchem-ml/
Estadística descriptiva (tablas y gráficas)	https://crisrumiguano-art.github.io/geoquimica-earthchem-ml/descriptiva/
Tabla de variables (dominio, rango, etc.)	https://crisrumiguano-art.github.io/geoquimica-earthchem-ml/descriptiva/#variables · data/tabla_variables.xlsx
Este informe (PDF y Word)	https://crisrumiguano-art.github.io/geoquimica-earthchem-ml/informe_proyecto.pdf · https://crisrumiguano-art.github.io/geoquimica-earthchem-ml/informe_proyecto.docx
Base limpia (159.760 × 75)	data/earthchem_limpio.csv.gz en el repositorio
Mapa en QGIS	data/gis/mapa_ubicacion.qgz (abrir junto con mapa_ubicacion.gpkg)
Código en R	carpeta R/ del repositorio
Repositorio (todo el código)	https://github.com/crisrumiguano-art/geoquimica-earthchem-ml
Datos originales de EarthChem (~600 MB)	Google Drive de Dalember: TRABAJOS ESTADISTICA › 10-26 GDRIVE GEOQUIMICA ROCAS › datos_crudos_earthchem (pedir acceso)

Qué tiene la página web

- **Inicio:** introducción, planteamiento del problema, mapa de ubicación (QGIS, Python e interactivo) y objetivos.
- **Descriptiva (Unidad 1):** población, individuo y muestra; tabla de variables; tablas de frecuencia y gráficas de las variables cualitativas, discretas y continuas; tratamiento de outliers; gráficos interactivos.
- **Datos:** de dónde salieron los datos, cómo se prepararon y todas las descargas.
- **Equipo:** integrantes y docente.
- **Inferencial, Regresión y Machine learning:** aparecen como «pronto»; se completan en las próximas unidades.

Estado por lección

Lección	Qué pide	Estado
L1–L2	introducción, problema, objetivos; población, individuo y muestra	Hecho (web: Inicio y Descriptiva 01)
L3	muestreo y mapa de ubicación en QGIS	Mapa hecho (QGIS y Python). Falta el informe de muestreo estratificado
L4	tabla de variables: tipo, subtipo, dominio, rango, instrumento, unidad	Hecho (web y Excel, formato del curso)
L5	tablas de frecuencia (TDF) y gráficas	Hecho: 4 cualitativas, 2 discretas, 3 continuas (SiO ₂ , MgO, Cr)
L6	indicadores de posición, dispersión y forma	Hecho para esas variables; se puede ampliar a más
L7	exposición de la etapa descriptiva y conclusiones	Pendiente: redactar conclusiones y preparar la exposición
Unidades 2–4	inferencia, regresión, red neuronal	Pendiente

Cinco ideas clave para la exposición

1. **La muestra no es aleatoria.** PetDB reúne lo que los investigadores decidieron estudiar durante casi un siglo, con más rocas del fondo del mar que de los continentes. Para la inferencia se usarán submuestras aleatorias por estrato.
2. **Una celda vacía significa «no se midió», no cero.** Por eso cada análisis usa solo las filas que tienen medidas las variables de ese análisis, y siempre se informa cuántas son (n).
3. **La agrupación por defecto no servía.** Con Sturges sobre el rango completo, el 99,9 % del cromo caía en un solo intervalo. Se usaron intervalos sobre el rango típico y dos clases abiertas para los extremos.
4. **Los outliers no se borraron.** Casi todos son otras rocas (riolitas, peridotitas, cromititas), no errores. Borrarlos habría eliminado poblaciones enteras.
5. **Los elementos traza se miran en escala logarítmica.** Así se ve su forma real y aparecen dos poblaciones en el cromo: basaltos y rocas ultramáficas.

Cómo trabajar con los archivos en tu computadora

Ver la base en Excel. Descarga `data/earthchem_limpio.csv.gz` desde la página Datos, descomprímela con 7-Zip y en Excel usa Datos > Desde texto/CSV, con origen UTF-8 y, en Transformar datos, configuración regional Inglés (Estados Unidos) para que el punto sea el separador decimal.

Correr los scripts de R. Instala R (4.4 o más nuevo) y RStudio. En GitHub, en el repositorio, usa Code > Download ZIP y descomprime. Abre RStudio, ve a Session > Set Working Directory y elige la carpeta del repositorio. La primera vez ejecuta `install.packages(c("e1071", "openxlsx"))`. Después, por ejemplo, `source("R/03_continuas.R")`: las gráficas quedan en `docs/img/r/` y las tablas en `R/salidas/`.

Abrir el mapa en QGIS. Descarga `mapa_ubicacion.qgz` y `mapa_ubicacion.gpkg` (página Datos) en la misma carpeta y abre el `.qgz`. La composición de impresión se llama «Mapa de ubicación A4».

Glosario

Término	Qué significa
TDF	tabla de distribución de frecuencias: cuántos datos caen en cada clase o intervalo
Li, Ls, MC	límite inferior y superior del intervalo, y su marca de clase (el punto medio)
ni, hi	frecuencia absoluta (cuántos datos) y relativa (qué porcentaje)
Ni, Hi ↑ / ↓	frecuencias acumuladas ascendentes y descendentes; con ellas se dibuja la ojiva
Sturges	regla para el número de intervalos: $k = 1 + 3,3 \log n$
IQR (rango intercuartílico)	$Q3 - Q1$; un dato es atípico si cae fuera de $Q1 - 1,5 \cdot IQR$ o de $Q3 + 1,5 \cdot IQR$
Clase abierta	intervalo sin uno de sus límites («< 36» o «≥ 66») que agrupa los valores extremos
Error de agrupación	diferencia entre la media (o desviación) calculada con la TDF y con los datos sin agrupar
Escala log10	trabajar con el logaritmo del valor; separa bien datos que van de 0,01 a 100.000
Estrato	subgrupo homogéneo de la muestra (por ejemplo, solo rocas volcánicas máficas)
Dominio / rango	valores que la variable puede tomar en teoría / valores que tomó en la muestra (ver sección 4.1)
wt% / ppm	porcentaje en peso / partes por millón (1 wt% = 10.000 ppm)

Parte I. La base de datos

1. Qué es esta base

La base reúne 159.760 análisis químicos de rocas tomados de muestras recolectadas en todo el mundo. Cada fila corresponde al análisis de una muestra de *roca total*, es decir, de la roca molida y medida en conjunto, no de un mineral en particular. Para cada muestra se informa su composición en elementos mayores (sílice, aluminio, magnesio, calcio, sodio, potasio y otros, expresados como óxidos en porcentaje en peso), en elementos traza (en partes por millón) y en tierras raras, junto con el tipo de roca, el lugar donde se tomó, el ambiente geológico en que se formó y el artículo científico que publicó el dato.

Predominan las rocas ígneas, y en particular los basaltos (72.307 análisis), seguidos por peridotitas, tefras, harzburgitas, diabasas, lherzolitas, gabros y andesitas. Cerca del 65% de las muestras proviene del fondo oceánico: PetDB nació para reunir los datos de las dorsales oceánicas y de las perforaciones de los programas internacionales de exploración del océano (DSDP, ODP e IODP).

2. De dónde salieron los datos

Los datos provienen de **PetDB** (Petrological Database), una de las bases que integran **EarthChem**, la infraestructura de datos geoquímicos financiada por la National Science Foundation de Estados Unidos. Hoy PetDB se consulta a través de la Astromaterials Data Synthesis (search.astromat.org) y del portal de EarthChem (www.earthchem.org/petdb).

PetDB no genera datos propios: su equipo los transcribe de artículos científicos revisados por pares y los ordena en un formato común. La exportación usada reúne datos de 2.348 artículos publicados entre 1932 y 2025; algo más de la mitad (58%) corresponde a trabajos publicados desde el año 2000. Cada fila conserva la cita del artículo de origen y un enlace a la ficha de la muestra en PetDB, de modo que cualquier valor se puede rastrear hasta su publicación.

La descarga se hizo el **13 de octubre de 2026** con un único filtro: material analizado igual a roca total (incluido el registrado como roca total o vidrio sin especificar). Se pidieron todas las familias de variables disponibles: elementos mayores, trazas, tierras raras, edades, miembros extremos, moda mineral, gases nobles y volátiles. La consulta exacta quedó guardada en `data/raw/earthchem_export_oct13_26_142339.txt`, así que la descarga se puede repetir. Licencia **Creative Commons Atribución 4.0 (CC BY 4.0)**: los datos pueden usarse y redistribuirse citando la fuente.

3. Cómo se preparó

La exportación original venía en dos archivos que suman unos 600 MB: uno con los resultados químicos, de 2.318 columnas, y otro con los datos descriptivos de cada muestra. Las columnas son tantas porque cada variable química viene acompañada de tres columnas de texto (método analítico, comentario y laboratorio). La preparación consistió en:

6. unir los dos archivos, verificando que las muestras coincidieran fila por fila;
7. quitar las columnas de método, comentario y laboratorio;
8. conservar las variables químicas con datos en al menos el 5 % de las muestras (quedan 52; se descartan, entre otras, las relaciones isotópicas, presentes en menos del 2 % de las filas);
9. convertir las coordenadas a grados decimales, con el sur y el oeste en negativo;

10. dejar vacía, y no en cero, toda celda química sin medición, y también las concentraciones negativas, que son lecturas bajo el límite de detección (salvo la pérdida por calcinación, que puede ser negativa de verdad);
11. corregir cinco muestras de la dorsal de Gakkel que venían con latitud y longitud invertidas;
12. construir cinco variables nuevas: año de publicación, número de analitos, clase de sílice, grado de alteración y ambiente tectónico agrupado.

4. Variables

La base final tiene 159.760 filas y 75 columnas, con todos los tipos de variable que se usan en estadística:

Tipo	Variables	Comentario
Nominales	Sample Type, Sample SubType, Tectonic Setting, Tectonic Group, Geographic Features, Rock Texture, Expedition, Citation	Categorías sin orden. El ambiente tectónico tiene 59 valores; Tectonic Group los resume en tres grandes ambientes más los mixtos.
Nominal binaria	Analyzed Material	Dos valores: roca total y roca total o vidrio sin especificar.
Ordinales	SiO2 Class, Alteration Grade, Geologic Age, Sample Alteration	SiO2 Class: ultrabásica (< 45 %), básica (45–52 %), intermedia (52–63 %), ácida (≥ 63 %). Alteration Grade: 0 (fresca) a 5 (totalmente alterada).
Numéricas discretas	Publication Year, N Analytes	Solo valores enteros.
Numéricas continuas	52 químicas, Latitude, Longitude, elevaciones, edades (Ma)	Óxidos mayores en % en peso; trazas y tierras raras en ppm.
Identificadores	Sample Name, Sample URL	No se analizan; sirven para rastrear cada muestra.

La tabla de variables completa sigue el formato de la plantilla del curso: nombre, descripción, tipo, subtipo, dominio (como definición y como conjunto, por ejemplo $D=\{x/x \text{ es concentración} \wedge x \in \mathbb{R}\}$ y $D=\{x/x \in \mathbb{R}^+ \cup \{0\}\}$), rango observado (por ejemplo $R=\{y/ 0.0011 \leq y \leq 98.9\}$ para el SiO₂), instrumento analítico, unidad y escala de medición. Está en la página web (Estadística descriptiva) y en data/tabla_variables.xlsx. El diccionario resumido está en el Anexo A.

4.1 Dominio y rango: qué significan

La idea es la misma que con las funciones en matemáticas. El **dominio** son todos los valores que la variable podría tomar en teoría, por cómo es la variable. El **rango** son los valores que tomó de verdad en nuestros datos. La plantilla del curso escribe el dominio dos veces: en palabras (la definición) y como conjunto numérico. Por ejemplo, una concentración puede ser cualquier número real desde cero hacia arriba (dominio $\mathbb{R}^+ \cup \{0\}$), pero en esta base el SiO₂ va de 0,0011 % a 98,9 % (rango).

Variable	Dominio (definición)	Dominio (conjunto)	Rango
SiO ₂ (wt%)	$D=\{x/x \text{ es concentración} \wedge x \in \mathbb{R}\}$	$D=\{x/x \in \mathbb{R}^+ \cup \{0\}\}$	$R=\{y/ 0.0011 \leq y \leq 98.9\}$
Publication Year	$D=\{x/x \text{ es tiempo} \wedge x \in \mathbb{N} \wedge \text{Tipo}(x)=\text{años}\}$	$D=\{x/x \in \mathbb{Z}^+\}$	$R=\{y/ 1932 \leq y \leq 2025\}$
Alteration Grade	$D=\{x/x \text{ es grado de alteración} \wedge x \in \mathbb{N} \wedge 0 \leq x \leq 5\}$	$D=\{x/x \in \mathbb{Z}^+ \cup \{0\}\}$	$R=\{y/y \in \{0, 1, 2, 3, 4, 5\}\}$
Sample Type	$D=\{x/x \text{ es roca} \wedge \text{Clasificación}(x)=\text{petrográfica}\}$	$D=\{x/x \in \text{Clase de roca}\}$	$R=\{y/y \in \text{Clase de roca}\}$ (27 categorías)

En las variables cualitativas el rango son las categorías que aparecen en la base; cuando son pocas se enumeran y cuando son muchas se indica cuántas hay.

5. Requisitos del trabajo

Requisito	Exigido	Base
Filas	al menos 10.000	159.760
Columnas	al menos 30	75
Tipos de variables	nominales, ordinales, discretas y continuas	todos presentes

6. Limitaciones que conviene conocer

No todas las muestras tienen todas las mediciones. Cada artículo midió lo que necesitaba, así que la base tiene muchos huecos: solo 46.308 filas (29%) tienen completos los seis óxidos principales, y las tierras raras aparecen en una de cada cinco filas. Una celda vacía significa que el valor no se midió; conviene filtrar las filas completas para cada análisis en lugar de rellenar huecos sin criterio.

Falta el TiO₂. El óxido de titanio prácticamente no vino en la exportación (cuatro valores). Si el análisis lo requiere, habrá que volver a descargarlo desde PetDB.

Unidades y formas distintas. Níquel, cobalto, cobre y zinc aparecen en % en peso y en ppm (1 % = 10.000 ppm), y el hierro se informa de cuatro maneras (FeO, Fe₂O₃, FeOT y Fe₂O₃T).

Sesgos de origen. Hay muchos más basaltos de fondo oceánico que rocas continentales y los métodos de laboratorio cambiaron en casi un siglo. El ambiente más frecuente, spreading center, reúne 26.621 muestras, mientras otros tienen apenas decenas. La muestra no es probabilística: la inferencia deberá hacerse con submuestras aleatorias estratificadas.

7. Acceso y cita

La base limpia está en `data/earthchem_limpio.csv.gz` (CSV comprimido, 26 MB). Para abrirla en Excel: descomprimir (por ejemplo con 7-Zip) e importar desde Datos › Desde texto/CSV con codificación UTF-8 y configuración regional Inglés (Estados Unidos), para que el punto se lea como separador decimal. En R o Python se lee directo.

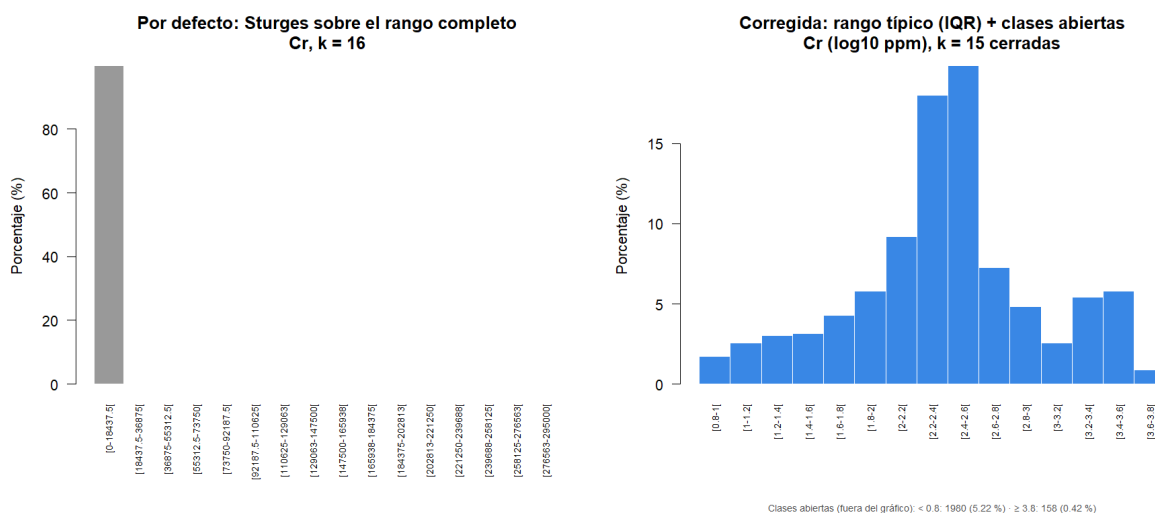
Cita sugerida: *Datos descargados de EarthChem PetDB / Astromaterials Data Synthesis (https://search.astromat.org/) el 13 de octubre de 2026; material analizado = roca total. Licencia CC BY 4.0.*

Parte II. Estadística descriptiva

La estadística se calcula en R, como exige la asignatura, siguiendo la estructura de las plantillas del curso: tablas de distribución de frecuencias (TDF) con límites, marca de clase, frecuencias absolutas y relativas y sus acumuladas ascendentes y descendentes; histograma con polígono de frecuencias, ojiva y diagrama de caja; e indicadores de posición, dispersión y forma. Los resultados completos están en la página web.

8. Por qué no sirvió la agrupación por defecto

Las plantillas calculan k por Sturges ($k = 1 + 3,3 \log n$) y la amplitud como $a = R / k$. En esta base el recorrido está dominado por unos pocos valores extremos: el cromo va de 0,002 a 295.000 ppm, cada intervalo mide 18.437 ppm y el primero se queda con el 99,9 % de los datos. Las reglas $\sqrt{n}$ y $2\sqrt{n}$ dan más de 190 intervalos.

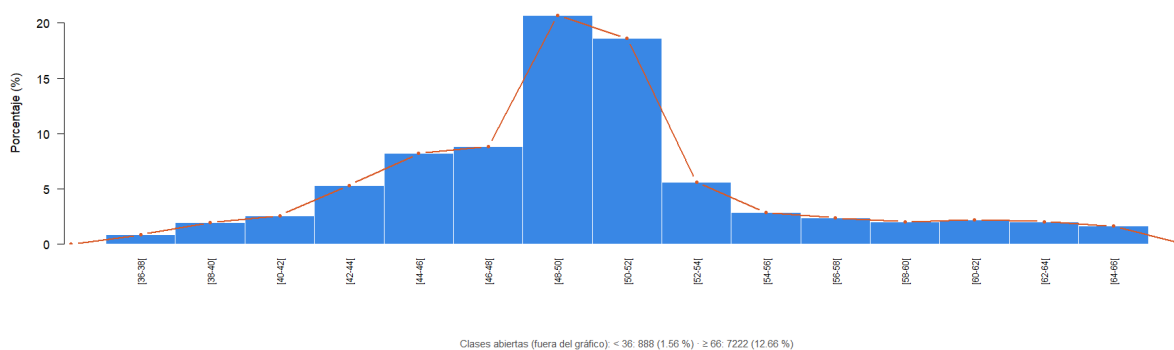


Cromo: agrupación por defecto (izquierda) frente a la corregida en log10 (derecha).

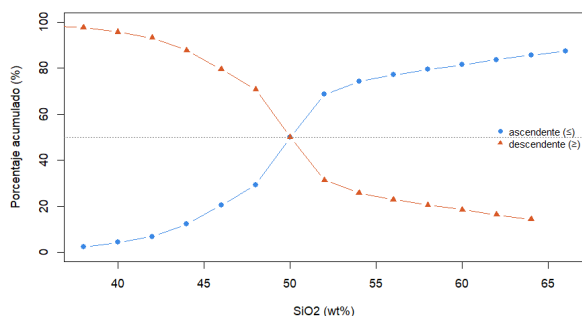
9. Outliers: otras rocas, no errores

La regla intercuartílica marca entre el 8 y el 17 % de los datos, y casi todos son rocas reales: tefras y rocas ácidas en la sílice, peridotitas en el magnesio, cromititas en el cromo. La base mezcla poblaciones, y la regla supone una sola. Por eso no se elimina ningún dato. La TDF corregida usa intervalos de Sturges sobre el rango típico (los límites intercuartílicos), con amplitud redonda, y dos clases abiertas para los extremos. El error de agrupación (media y desviación agrupadas frente a datos totales) queda por debajo del 5 % en todos los casos. Los elementos traza se analizan en log10 y todo el análisis se repite en un estrato homogéneo (rocas volcánicas máficas).

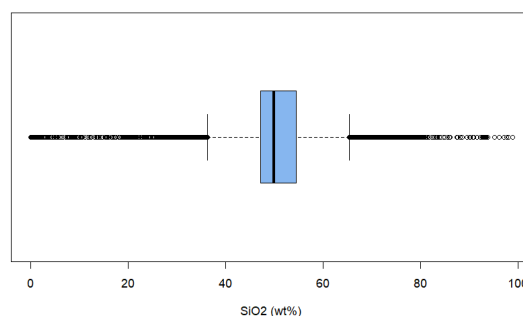
Gráfica No. 1: Histograma y polígono de frecuencias de SiO₂ (wt%)



Gráfica No. 2: Ojiva ascendente y descendente



Gráfica No. 3: Diagrama de caja (todos los datos)



SiO₂: histograma con polígono, ojiva y diagrama de caja de la TDF corregida.

El razonamiento completo, con todas las cifras, está en R/analisis_agrupacion_outliers.md.

10. ¿Alcanzan los datos para las regresiones?

Sí. Aunque hay muchos huecos, en una regresión solo cuentan las filas que tienen medidas las dos variables, y quedan miles. La tabla muestra algunos pares útiles para la Unidad 3, con el número de filas completas y la correlación (r), en escala original y en escala logarítmica:

Variable X	Variable Y	Filas completas	r	r (log10)
MgO (wt%)	Ni (ppm)	9.379	0,90	0,85
La (ppm)	Ce (ppm)	31.690	0,91	0,98
Zr (ppm)	Hf (ppm)	20.521	0,15	0,97
Nb (ppm)	Ta (ppm)	18.439	-0,00	0,92
Rb (ppm)	K ₂ O (wt%)	18.485	0,73	0,90
Th (ppm)	U (ppm)	23.481	0,65	0,89
MgO (wt%)	Cr (ppm)	20.074	0,34	0,78
SiO ₂ (wt%)	Na ₂ O+K ₂ O	47.868	0,55	0,47

Para una regresión múltiple con los seis óxidos mayores quedan 46.308 filas completas. Tres precauciones:

- **Usar escala logarítmica en los elementos traza.** Zr y Hf parecen no relacionarse en escala normal (r bajo), pero en log10 la relación es casi perfecta: unos pocos valores extremos arruinan la regresión si no se transforma.
- **No rellenar huecos con ceros ni promedios.** Trabajar con las filas completas de cada análisis e informar su n .
- **Estratificar.** Mezclar basaltos con peridotitas puede inventar o esconder relaciones; conviene repetir cada regresión por tipo de roca o por grupo tectónico.

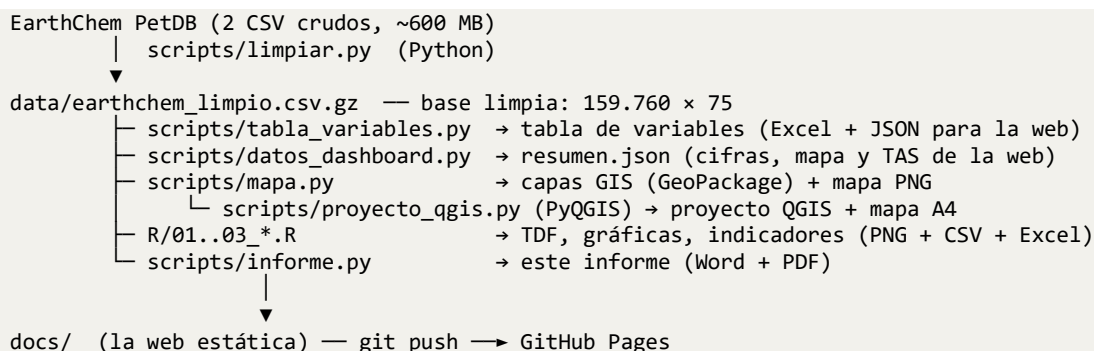
Pares sugeridos por tipo de modelo: lineal simple, MgO frente a Ni; potencial o log-log, Zr frente a Hf y Nb frente a Ta; múltiple, los óxidos mayores.

Parte III. Cómo se construyó: documentación técnica

Esta parte explica cómo se hizo cada producto del proyecto, con qué herramientas y por qué, para que cualquier integrante (o el docente) pueda entenderlo, reproducirlo y modificarlo. Todo el código está en el repositorio.

11. Arquitectura general

El proyecto es una cadena de pasos donde cada uno lee el resultado del anterior. No hay servidor ni base de datos en línea: todo se calcula en la computadora y el resultado se publica como archivos estáticos.



12. Herramientas y librerías

Herramienta	Versión	Para qué se usó
Python	3.13	limpieza, tabla de variables, mapa, agregados de la web, informe
pandas / numpy	2.3 / 2.4	lectura y resumen de la base limpia
csv, gzip, re (biblioteca estándar)	—	lectura por streaming del CSV crudo de 2.318 columnas
openpyxl	3.1	exportar la tabla de variables a Excel
geopandas, shapely, pyproj, pyogrio	1.1 / 2.1 / 3.7 / 0.12	capas GIS, reproyección a Robinson, GeoPackage
matplotlib + matplotlib-scalebar	3.10 / 0.9	mapa de ubicación con escala, norte y grilla
QGIS (PyQGIS)	3.44	proyecto .qgz y composición de impresión A4
Natural Earth 1:110m	—	límites de países (dominio público)
R	4.6.1	estadística descriptiva (exigencia de la asignatura)
e1071 / openxlsx (R)	1.7 / 4.2	asimetría y curtosis; exportar TDF a Excel
HTML5, CSS3, JavaScript	—	la página web, sin framework
Plotly.js (CDN)	2.35.2	gráficos interactivos: mapa, barras, diagrama TAS
Google Fonts	—	Fraunces, Instrument Sans y JetBrains Mono
Pillow	12.2	recorte de fotos, optimización de imágenes, favicon
gflow-cli (Google Flow)	0.83	imágenes ilustrativas generadas con IA
python-docx + Microsoft Word	1.2	este informe y su versión PDF
Git, GitHub, GitHub Pages, gh CLI	—	control de versiones y publicación de la web

13. Limpieza de la base (scripts/limpiar.py)

El CSV crudo tiene 2.318 columnas y celdas con saltos de línea, así que no se carga entero con pandas: se recorre fila por fila con el módulo csv de la biblioteca estándar, en dos pasadas. La primera cuenta cuántos valores numéricos tiene cada columna química y comprueba que el nombre de cada muestra coincida con el del archivo de metadatos (si el orden se rompe, el script se detiene). La segunda escribe la base limpia comprimida con gzip.

Cada regla de limpieza es una función pequeña y probada: el bloque principal tiene asserts que verifican, por ejemplo, que «39.244°N» se convierta en 39,244, que una latitud mayor que 90 se intercambie con la longitud y que un ambiente combinado como «FRACTURE ZONE,SPREADING CENTER» caiga en el grupo correcto:

```
def grupo_tectonico(ambiente):
    """Valores combinados toman el grupo común; si difieren, 'mixto'."""
    if not ambiente:
        return ""
    grupos = {g for a in ambiente.upper().split(",")
               for g, miembros in GRUPO_TECTONICO.items() if a.strip() in miembros}
    return grupos.pop() if len(grupos) == 1 else "mixto"
```

La tabla de variables (`scripts/tabla_variables.py`) clasifica cada columna y toma como «instrumento» el método analítico más frecuente de cada variable, leído de las columnas Method del archivo crudo (por ejemplo, el SiO₂ se midió con fluorescencia de rayos X en el 64 % de los casos).

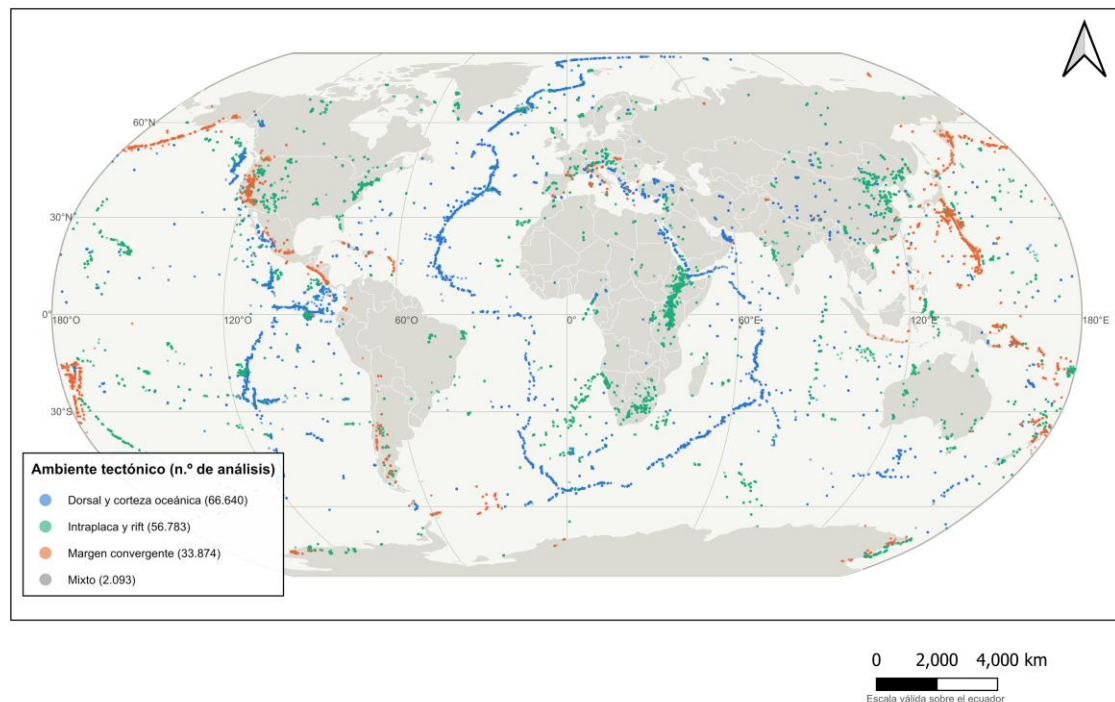
14. Mapa de ubicación (Python y QGIS)

`scripts/mapa.py` convierte las coordenadas en puntos con geopandas, descarga los límites de países de Natural Earth y genera una grilla de 30° y el contorno del globo. Todo se guarda en un GeoPackage (`data/gis/mapa_ubicacion.gpkg`) en WGS84, con nombres de campo cortos para la tabla de atributos de QGIS. Luego dibuja el mapa en proyección Robinson (ESRI:54030) con matplotlib: título, leyenda con el número de análisis por ambiente tectónico, escala gráfica, flecha de norte, coordenadas y fuente.

`scripts/proyecto_qgis.py` se ejecuta con el Python que trae QGIS (PyQGIS). Carga las mismas capas, aplica una simbología categorizada con los mismos colores, arma una composición de impresión A4 y la exporta a PNG. El proyecto se guarda con rutas relativas, así que el .qgz funciona en cualquier computadora si está junto al .gpkg. Un detalle técnico: QGIS no logra ubicar etiquetas junto a los polos en la proyección Robinson, por eso las longitudes se rotulan sobre el ecuador.

Mapa de ubicación de las muestras de roca total — EarthChem PetDB

Proyección Robinson (ESRI:54030) · Datum WGS84 · Estadística y Probabilidades, FIGEMPA – Universidad Central del Ecuador



Fuente: EarthChem PetDB / Astromaterials Data Synthesis (descarga 13/10/2026, CC BY 4.0); límites: Natural Earth. Elaboración: Rumiguano, Fernández, Bustamante y Haro, 2026.

Composición A4 generada con PyQGIS.

15. Estadística en R (carpeta R/)

R/funciones.R reúne las funciones comunes; los scripts 01_cualitativas.R, 02_discretas.R y 03_continuas.R las aplican. La función central arma la TDF; sin argumentos reproduce exactamente la plantilla del curso, y con un rango típico agrega las clases abiertas:

```
tdf_def <- tdf_continua(x, redondear = FALSE) # plantilla: a = R / k
lim <- limites_iqr(x) # Q1 - 1,5·IR ; Q3 + 1,5·IR
tdf_fin <- tdf_continua(x, lim = lim) # Sturges en el rango típico + clases
abiertas
error_agrupacion(x, tdf_fin) # media y desviación agrupadas vs. totales
```

Cada script produce tres salidas: las gráficas en PNG (docs/img/r/), las tablas en CSV para la web (docs/data/tdf/) y un Excel con todas las tablas (R/salidas/). Un problema encontrado: en R 4.6 para Windows, read.csv sobre el archivo .gz se cerraba con un error de segmentación; se resolvió abriéndolo con gzfile y codificación UTF-8 explícita (función leer_base).

16. La página web

16.1 Estructura y tecnología

La web es un sitio estático: archivos HTML, CSS y JavaScript que el navegador lee directamente, sin framework ni paso de compilación. Se eligió así para que GitHub Pages la sirva gratis tal cual y para que cualquier integrante pueda editarla con un editor de texto. Tiene una página por unidad del curso, que comparten un archivo de estilos y uno de código:

Archivo	Contenido
docs/index.html	portada, introducción, problema, mapa (con pestañas QGIS / Python / interactivo), objetivos

docs/descriptiva/index.html	Unidad 1: población y muestra, tabla de variables, TDF y gráficas de R
docs/datos/index.html	origen de los datos, preparación y descargas
docs/equipo/index.html	integrantes y docente
docs/assets/estilo.css	todo el diseño: colores, tipografía, grillas, animaciones, versión móvil
docs/assets/app.js	código común: navegación, animaciones, pestañas, lectura de CSV, estilo de gráficos
docs/data/	resumen.json, variables.json y las TDF en CSV
docs/img/	fotos, mapas, gráficas de R e imágenes ilustrativas

16.2 Cómo carga los datos

La página no contiene los datos escritos en el HTML: los pide al cargarse con fetch y los dibuja. La tabla de variables viene de variables.json; los gráficos interactivos, de resumen.json (254 KB con agregados y muestras de 6.000 puntos, en lugar de la base completa de 26 MB); y las TDF, de los CSV que exporta R. Para leer esos CSV sin librerías se escribió un lector de pocas líneas:

```
async function leerCSV(url) {
  const texto = (await (await fetch(url)).text()).trim();
  const campos = linea => [...linea.matchAll(/"([^\"]*)"|([^\s,]*)/, |$)/g)]
    .map(m => m[2] ?? m[1]).slice(0, -1);
  const [cab, ...filas] = texto.split(/\r?\n/).map(campos);
  return filas.map(f => Object.fromEntries(cab.map((c, i) => [c, f[i]])));
}
```

Por eso la web debe verse a través de un servidor (GitHub Pages, o localmente con `python -m http.server`): abriendo el archivo con doble clic, el navegador bloquea esas lecturas.

16.3 Componentes y comportamiento

- **Navegación entre páginas:** un menú fijo marca la página actual (atributo `data-pagina` del `body`); las unidades aún no hechas aparecen como «pronto».
- **Aparición al hacer scroll:** los bloques con clase `rev` aparecen con un desplazamiento suave usando `IntersectionObserver`, una API del navegador que avisa cuando un elemento entra en pantalla.
- **Cifras animadas:** las cifras de la portada cuentan hacia arriba con `requestAnimationFrame`.
- **Pestañas:** una función genérica muestra un panel y oculta los demás (mapa, variables continuas).
- **Tabla de variables filtrable:** filtros por tipo y un buscador que filtra mientras se escribe.
- **Gráficos interactivos:** `Plotly.js` dibuja el mapa mundial, las barras de ambientes tectónicos y el diagrama TAS, con información al pasar el cursor.

16.4 Diseño

La dirección visual es la de una revista científica sobre un tema oscuro «basalto»: fondo casi negro cálido, texto color hueso y un acento color magma. Los colores se definen una sola vez como variables CSS y el resto del código las usa por nombre, de modo que cambiar un color se hace en un solo lugar:

```
:root {
  --page: #0f0e0c; --surface-1: #171513; --text-primary: #f0ece4;
  --magma: #ff6b2c; --oro: #e9b44c;
  --series-1: #3987e5; --series-2: #d95926; --series-3: #199e70;
  --serif: "Fraunces", Georgia, serif; --sans: "Instrument Sans", sans-serif;
}
```

Los colores de los datos (azul, naranja y verde agua, más una rampa azul para las variables ordinales) son una paleta validada para personas con daltonismo, y son los mismos en la web, en el mapa de Python y en el de QGIS. Nunca se usan más de tres colores de categoría en un mismo gráfico. La

tipografía combina Fraunces para títulos, Instrument Sans para el texto y JetBrains Mono para cifras y fórmulas.

16.5 Accesibilidad, celulares y rendimiento

- Todas las imágenes tienen texto alternativo; las pestañas y botones de filtro usan atributos ARIA.
- Si el sistema pide reducir el movimiento (prefers-reduced-motion), se desactivan las animaciones.
- Debajo de 900 px de ancho las grillas pasan a una o dos columnas y el menú se desplaza horizontalmente; se verificó que ninguna página se desborde en una pantalla de celular de 390 px.
- Las imágenes fuera de pantalla se cargan recién al acercarse (loading="lazy") y los scripts se cargan con defer.
- Las imágenes generadas se comprimieron a JPEG (unos 170–335 KB cada una).

16.6 Imágenes

Las fotos del equipo son reales: se recortaron a 400 × 400 px con Pillow. Las imágenes ilustrativas (lámina delgada de basalto, Andes ecuatorianos, buque de perforación, testigos de roca) se generaron con Google Flow desde la terminal con gflow-cli, por ejemplo:

```
gflow image t2i "Macro photograph of a basalt thin section under cross-polarized light..." \
--model nano2 --aspect 16:9 -n 1 --out <carpeta>
```

La web indica en los pies y en el pie de página que esas imágenes son ilustrativas y fueron generadas con IA. El favicon (el logo «Pt») se dibujó en SVG y se exportó a ICO y PNG con Pillow.

17. Publicación y control de versiones

El proyecto vive en un repositorio Git publicado en GitHub. Cada cambio se registra como un commit con un mensaje que explica qué se hizo, así que el historial documenta cómo evolucionó el trabajo. GitHub Pages está configurado para publicar la carpeta docs/ de la rama main: cada git push que modifica esa carpeta actualiza la web en uno a tres minutos. El repositorio es público porque GitHub Pages gratuito no funciona con repositorios privados; los datos tienen licencia CC BY 4.0, que permite publicarlos citando la fuente.

El repositorio pertenece a la cuenta de Cristoffer Rumiguano (crisrumiguano-art), a quien se transfirió el 15 de octubre de 2026; Dalember (cuenta DalemberV) colabora con permiso de escritura. Cualquier integrante con acceso puede proponer cambios con un pull request.

Límites a tener en cuenta: GitHub rechaza archivos de más de 100 MB, por eso los datos crudos (~600 MB) no están en el repositorio. Los archivos más pesados que sí están son las capas GIS (~35 MB) y la base limpia (~26 MB).

18. Documentación del proyecto

- **README.md**: la portada del repositorio en GitHub, con enlaces a todo.
- **AGENTS.md**: reglas y decisiones del proyecto (requisitos del curso, decisiones de limpieza, trampas de los datos, diseño). Lo leen los asistentes de programación con IA (Claude Code, Gemini, Codex); CLAUDE.md y GEMINI.md lo importan.
- **openwiki/**: documentación explicativa en formato OpenWiki: base de datos, cómo regenerar y publicar, y estructura de la web.
- **R/analisis_agrupacion_outliers.md**: el razonamiento de la estadística descriptiva.
- **Este informe**: se genera con scripts/informe.py, así que se actualiza solo si cambian los datos.

19. Cómo reproducir todo

Los dos CSV crudos están respaldados en Google Drive (carpeta 10-26 GDRIVE GEOQUIMICA ROCAS › datos_crudos_earthchem). En la computadora de Dalember, data/raw/crudos es un enlace (junction) a esa carpeta, y los scripts la usan por defecto. En otra computadora, basta con copiar ahí los CSV o pasar su carpeta como argumento. Desde la raíz del repositorio y en este orden:

```
pip install -r requirements.txt
python -I scripts/limpiar.py          # usa data/raw/crudos
python -I scripts/datos_dashboard.py
python scripts/tabla_variables.py     # usa data/raw/crudos
python -I scripts/mapa.py
"C:\Program Files\QGIS 3.44.8\bin\python-qgis-ltr.bat" scripts/proyecto_qgis.py
Rscript R/01_cualitativas.R
Rscript R/02_discretas.R
Rscript R/03_continuas.R
python scripts/informe.py
git add -A && git commit -m "..." && git push
```

Para revisar la web antes de publicarla: `python -m http.server 8765 -d docs` y abrir `http://localhost:8765/` en el navegador.

20. Limitaciones técnicas y mantenimiento

- El menú de navegación está repetido en cada página: al agregar una página (por ejemplo, Inferencial) hay que actualizarlo en todas.
- Los textos de lectura de las gráficas en la página descriptiva están escritos a mano con las cifras actuales; si cambian los datos, hay que revisarlos.
- El informe usa Microsoft Word para generar el PDF y actualizar el índice; sin Word queda solo el .docx.
- mapa.py necesita internet para descargar los límites de Natural Earth.
- gflow-cli automatiza la interfaz web de Google Flow y puede dejar de funcionar si Google la cambia.

Anexo A. Diccionario de variables

Columna	Tipo	% con dato
Sample Name	identificador	99,5
Sample URL	identificador	100,0
Citation	nominal	100,0
Publication Year	discreta	100,0
Analyzed Material	nominal (binaria)	100,0
Sample Type	nominal	100,0
Sample SubType	nominal	100,0
Latitude	continua	100,0
Longitude	continua	100,0
Geographic Features	nominal	99,8
Tectonic Setting	nominal	99,8
Rock Texture	nominal	24,9
Sample Alteration	ordinal (texto original)	22,1
Elevation Min (m)	continua	53,5
Elevation Max (m)	continua	39,2
Geologic Age	ordinal	10,5
Geologic Age Max (Ma)	continua	5,5
Geologic Age Min (Ma)	continua	7,9
Expedition	nominal	99,9
Tectonic Group	nominal	99,8
Alteration Grade	ordinal	20,8
SiO2 Class	ordinal	35,7
N Analytes	discreta	100,0
SiO2 (wt%)	continua	35,7
Al2O3 (wt%)	continua	35,9
Fe2O3 (wt%)	continua	14,6
Fe2O3T (wt%)	continua	8,1
FeO (wt%)	continua	13,8
FeOT (wt%)	continua	9,1
MgO (wt%)	continua	35,7
MnO (wt%)	continua	34,5
CaO (wt%)	continua	35,8
Na2O (wt%)	continua	34,5
K2O (wt%)	continua	34,0
P2O5 (wt%)	continua	30,2
Co (wt%)	continua	6,9
Cu (wt%)	continua	7,8
LOI (wt%)	continua	15,6
Ni (wt%)	continua	8,7
Zn (wt%)	continua	8,0
Rb (ppm)	continua	27,6
Sr (ppm)	continua	31,1
Ba (ppm)	continua	25,1
Cs (ppm)	continua	11,2
Pb (ppm)	continua	15,2
Y (ppm)	continua	27,5
Zr (ppm)	continua	28,8
Hf (ppm)	continua	16,8
Nb (ppm)	continua	24,4
Ta (ppm)	continua	14,2
Th (ppm)	continua	18,1

U (ppm)	continua	15,9
Ni (ppm)	continua	16,7
Co (ppm)	continua	10,5
V (ppm)	continua	20,8
Cr (ppm)	continua	23,7
Sc (ppm)	continua	18,6
Cu (ppm)	continua	10,3
Ga (ppm)	continua	10,7
Li (ppm)	continua	5,7
Zn (ppm)	continua	12,6
Ce (ppm)	continua	21,8
Dy (ppm)	continua	15,6
Er (ppm)	continua	15,4
Eu (ppm)	continua	19,1
Gd (ppm)	continua	15,6
Ho (ppm)	continua	13,7
La (ppm)	continua	21,6
Lu (ppm)	continua	17,0
Nd (ppm)	continua	20,6
Pr (ppm)	continua	14,5
Sm (ppm)	continua	20,3
Tb (ppm)	continua	16,9
Tm (ppm)	continua	11,2
Yb (ppm)	continua	19,1

Anexo B. Estructura del repositorio

data/	base limpia, tabla de variables, capas GIS y proyecto QGIS
raw/	consulta de descarga y términos de uso
scripts/	Python: limpieza, tabla de variables, agregados web, mapa, QGIS, informe
R/	estadística descriptiva (funciones + 3 scripts) y su análisis
salidas/	Excel con todas las TDF
docs/	la web publicada en GitHub Pages
assets/	estilo.css y app.js
descriptiva/	datos/ equipo/ páginas
data/	resumen.json, variables.json, tdf/*.csv
img/	fotos, mapas, gráficas de R (img/r/) e imágenes generadas
openwiki/	documentación explicativa
AGENTS.md	reglas del proyecto (CLAUDE.md y GEMINI.md lo importan)
README.md	portada del repositorio