Proyecto 1: Fundamentos y Preparación de Datos
Introducción a la Ciencia de Datos

Integrantes:	Avendaño, Joksan; Canché, Elías; Rodríguez, Juan
Programa Educativo:	Maestría en Probabilidad y Estadística
Institución:	Centro de Investigación en Matemáticas
Profesor:	Dr. Marco Antonio Aquino López

Resumen

Se presentan a través del desarrollo de un ejemplo las técnicas de limpieza de datos de manejo de datos faltantes y outliers. Los datos utilizados son de concentraciones de isótopos de carbono-13 en distintas regiones de Europa a lo largo de 400 años. [illegible] los criterios de limpieza se contextualizan los datos al comportamiento típico del carbono-13 captado en árboles a lo largo del tiempo.

1. Introducción

En el quehacer estadístico son indispensables los datos, pero con frecuencia los que recibe un analista no son exactamente los mismos que fueron generados en el estudio original. Esto se debe a que las bases de datos suelen presentar estructuras consideradas *irregulares*, que al ser tratadas terminan modificando la información disponible. Una dificultad adicional es que lo que constituye un *comportamiento irregular* depende en gran medida del contexto, lo que hace que el tratamiento de los datos no sea un proceso fácilmente escalable.

Actualmente existen herramientas que nos permiten pre-analizar y preparar información para su uso, todas ellas han sido formalizada en términos probabilísticos. Un ejemplo son los mecanismos que que explican la presencia de datos faltantes, como se describe en [Sch00]. Particularmente, los mecanismos de *faltantes aleatorios* (MAR), *faltantes completamente aleatorios* (MCAR) y *faltantes no aleatorios* (MNAR) justifican por qué es necesario revisar cada caso con atención.

En este contexto, el científico de datos asume el rol de traducir las complicaciones de una disciplina en propuestas para el manejo estadístico de la información. El siguiente trabajo tiene como finalidad ilustrar esta conjunción interdisciplinaria y presenta la preparación de una base de datos sobre isótopos de carbono.

Los principales objetivos de este trabajo son:

- Identificar, mediante métodos numéricas y gráficas, situaciones que pudieran requerir *limpiar* los datos.
- Contextualizar las irregularidades encontradas y proponer decisiones para su tratamiento.

2. Descripción y exploración inicial de los datos

2.1. Descripción de la base de datos

La base de datos [Mem+23] fue generada en el marco del proyecto ISONET con el objetivo de crear una extensa red espaciotemporal de isótopos estables hallados en anillos de los árboles en toda Europa, iniciativa financiada por el quinto Programa Marco de la Comisión Europea "Energía, Medio Ambiente y Desarrollo Sostenible". Consiste de 400 años de reconstrucciones anuales de la variabilidad climática europea mediante una red isotópica de alta resolución; se han construido 24 cronologías europeas de isótopos estables con resolución anual a partir de la celulosa de los anillos de los árboles durante los últimos 400 años (1600 d. C. – 2003 d. C.) para el carbono y el oxígeno, y durante los últimos 100 años para el hidrógeno, dando así 25 cronologías.

Para obtener los datos fueron perforados quince o más ejemplares codominantes de *Pinus sylvestris*, *Quercus robur/petraea* o *Cedrus atlantica* a aproximadamente 1.5 metros de altura respecto al suelo desde dos posiciones opuestas utilizando un perforador de 5 mm de diámetro, todos de edad similar y para cada sitio (Suunto, Finlandia o Mora, Suecia).

Las primeras 9 líneas de cada columna contienen información sobre el sitio donde fue tomada la muestra (ver tabla 1); mientras que la fila 10 se utiliza como encabezado para los datos 13CVPDB, los cuales se miden

en unidades $^{13}C/^{12}C$, este, es una relación de $^{13}C/^{12}C$ por milésimas de pulgada (mm) frente a la densidad de polvo de Viena (DPEV). Estas son medidas continuas e indexadas por el año y se describen en la tabla 2.1; no hay variables categóricas.

Característica	Descripción
Código del sitio	Código de 3 letras para cada sitio
Nombre del sitio	Nombre del sitio forestal o de la localidad más cercana
País	País al que pertenece el sitio
Latitud	Coordenadas geográficas, latitud en grados decimales
Longitud	Coordenadas geográficas, longitud en grados decimales
Especie	Nombre en latín de la especie arbórea
Primer año	Primer año de registro
Último año	Último año de registro
Elevación	Elevación promedio de la ubicación de los árboles respecto al nivel del mar
Los valores faltantes se indican con NA	

Cuadro 1: Descripción del encabezado

Código	Sitio	País	Latitud	Longitud	Especie	Primer año	Último Año	Elevación
COL	Col Du Zad	Morocco	32.97	-5.07	Cedrus atlantica	1600	2000	2200
SER	Monte Pollino	Italy	39.93	16.21	Pinus leucodermis	1604	2003	1900
CAZ	Cazorla	Spain	37.93	-2.97	Pinus nigra	1600	2002	1820
POE	Poellau	Austria	47.31	15.81	Pinus nigra	1600	2002	500
LIL	Pinar de Lillo	Spain	43.07	-5.25	Pinus sylvestris	1600	2002	1600
ILO	Sivakkovaara	Finland	62.98	31.27	Pinus sylvestris	1600	2002	200
INA	Inari	Finland	68.93	28.31	Pinus sylvestris	1600	2002	150
GUT	Gutuli	Norway	62.00	12.18	Pinus sylvestris	1600	2003	800
NIE2	Niepolomice	Poland	50.03	20.35	Pinus sylvestris	1627	2003	190
PAN	Panemunes	Lithuania	54.09	23.96	Pinus sylvestris	1816	2002	45
SUW	Suwalki	Poland	53.95	23.25	Pinus sylvestris	1600	2004	160
VIG	Vigera	Switzerland	46.05	8.77	Pinus sylvestris	1675	2003	1400
WIN	Windsor	United Kingdom	51.43	-0.61	Pinus sylvestris	1763	2003	80
WOB	Woburn	United Kingdom	51.98	-0.59	Pinus sylvestris	1604	2003	10
PED	Pedraforca	Spain	42.23	1.70	Pinus uncinata	1600	2003	2120
VIN	Vinuesa	Spain	42.00	2.75	Pinus uncinata	1850	1999	1950
DRA	Dransfeld	Germany	51.51	9.78	Quercus petraea	1776	1999	320
CAV	Cavergno	Switzerland	46.35	8.60	Quercus petraea	1637	2002	900
FON	Fontainebleau	France	48.38	2.67	Quercus petraea	1600	2000	100
AHI	Perchtoldsdorf Wehrturm, Gaaden Glockenturm, Klosterneuburg Glockenturm	Austria	48.25	16.77	Quercus petraea	1600	1883	n.s.
LAI	Lainzer Tiergarten	Austria	48.18	16.20	Quercus petraea	1812	2003	300
LOC	Lochwood	United Kingdom	55.27	-3.43	Quercus petraea	1749	2003	175
BRO	Bromarv	Finland	60.00	23.08	Quercus robur	1901	2002	5
NIE1	Niepolomice	Poland	50.03	20.35	Quercus robur	1627	2003	190
REN	Renn	France	48.02	-1.83	Quercus robur	1611	1998	100

Cuadro 2: D e de las variables

En el contexto geológico, los datos de $\delta^{13}C$ se utilizan para determinar si en las regiones de medición ha crecido cierto tipo de pasto. Valores positivos de $\delta^{13}C$ indican un incremento en carbón orgánico en las rocas sedimentarias; mientras que valores negativos indican decremento. En la paleoceanografía se ha utilizado para estudiar la historia de concentraciones de algas [Lib92]. Una posible aplicación de interés económico [1] de esta información es como criterio para determinar si una región puede ser fuente de carbón vegetal.

[1] Otra aplicación, presentada en [Mei18], es en ciencia forense. Se ha considerado que los valores para carbono y nitrógeno de δ, pueden utilizarse para encontrar fuentes de marihuana. Se ha intentado hacer estas mismas técnicas para el rastreo de fuentes de morfina, sólo que no ha resultado tan eficiente como con la marihuana.

3. Detección de problemas en los datos

La base de datos tiene algunas características que, de no manejarse con cuidado, podrían llevar a problemas al momento del procesamiento de los datos.

Sobre la organización del archivo de datos:

- Si bien son informativas, las primeras nueve filas del archivo complican la selección de los datos numéricos que podrían ser el principal objeto de interés.
- La documentación de la base de datos en [Mem+23] indica que hay datos sobre otros elementos además del carbono como el hidrógeno y el oxígeno. Este hecho no es claro de la base de datos, ya que todas las columnas con valores numéricos tienen el mismo nombre, correspondiente al método de medición del carbono-13.

Sobre los datos faltantes:

- Muchos de los valores que en la base son reportados como faltantes, se encuentran al inicio de las columnas. Esto se debe a que los estudios que conforman la base de datos comenzaron en fechas distintas.
- Algunos de los valores que se leen como `NA`, por cuestión del modo de captura de datos, tienen espacios al final. Aunque se lean todos de la misma manera, cuando se procesan los datos la computadora podría llegar a identificar un caracter extra haciendo que *como cadenas* sean distintas el `NA` sin espacio y el `NA` con espacio.

Sobre la manera en que los datos están registrados:

- Al igual que con los valores faltantes con `NA`, algunos valores numéricos están registrados con espacios adicionales. La complicación que trae esto es el que el procesador de datos en ocasiones considera los *numéricos* con espacios como *cadenas de texto*.

Ya mencionados los detalles detectados que pueden obstaculizar el análisis, se pueden mencionar detalles *estructurales* de los datos. Con esto se hace referencia a condiciones de los datos que puedan afectar directamente procesos de inferencia sobre los datos.

- La base de datos **no** presenta problemas de datos duplicados.
- En algunas columnas, existen años dentro del periodo de estudio donde se tienen valores de `NA`. Las propociones de estos datos faltantes se describe en la tabla 3
- Al visualizar los datos como series de tiempo e histogramas, aparecen puntos que se comportan de maneras que no son consistentes con el comportamiento de otros puntos adyacentes en el tiempo. Estos posiblemente podrían considerarse como *outliers*

En la figura 1 se presenta una gráfica del comportamiento regular de las columnas. Esta regularidad se atribuye a que no hay datos faltantes en el periodo de estudio y tampoco hay datos que se desvíen de la tendencia global. En las figuras 2, 4 y 5 se presenta una misma columna de datos que presenta problemas tanto de datos faltantes como de outliers. En la visualización como serie de tiempo de la figura 2 son evidentes los datos faltantes, mientras que en la visualización a través de un histograma en la figura 4 es evidente la presencia de outliers, lo cual se refuerza con el boxplot de la figura 5.

Sitio	Periodo	N. Datos	Faltantes	% Faltantes
BRO	(1901–2002)	102	0	0.00 %
CAV	(1637–2002)	366	1	0.27 %
CAZ	(1600–2002)	403	0	0.00 %
COL	(1600–2000)	401	121	30.17 %
DRA	(1776–1999)	224	0	0.00 %
FON	(1600–2000)	401	118	29.43 %
GUT	(1600–2003)	404	1	0.25 %
ILO	(1600–2002)	403	0	0.00 %
INA	(1600–2002)	403	0	0.00 %
AHI	(1600–1883)	284	0	0.00 %
LAI	(1812–2003)	192	1	0.52 %
LIL	(1600–2002)	403	4	0.99 %
LOC	(1749–2003)	255	0	0.00 %
NIE1	(1627–2003)	377	0	0.00 %
NIE2	(1627–2003)	377	0	0.00 %
PAN	(1816–2002)	187	0	0.00 %
PED	(1600–2003)	404	3	0.74 %
POE	(1600–2002)	403	0	0.00 %
REN	(1611–1998)	388	21	5.41 %
SER	(1604–2003)	400	0	0.00 %
SUW	(1600–2004)	405	0	0.00 %
VIG	(1675–2003)	329	1	0.30 %
VIN	(1850–1999)	150	0	0.00 %
WIN	(1763–2003)	241	9	3.73 %
WOB	(1604–2003)	400	5	1.25 %

Cuadro 3: Porcentaje de valores faltantes por periodo de estudio

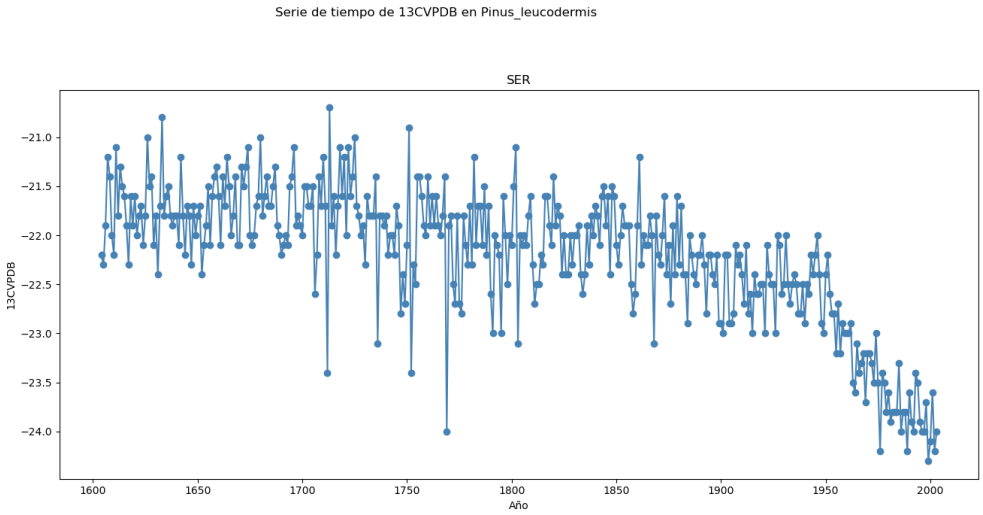

Figura 1: Serie de tiempo correspondiente a la columna SER con un comportamiento que se considera regular.

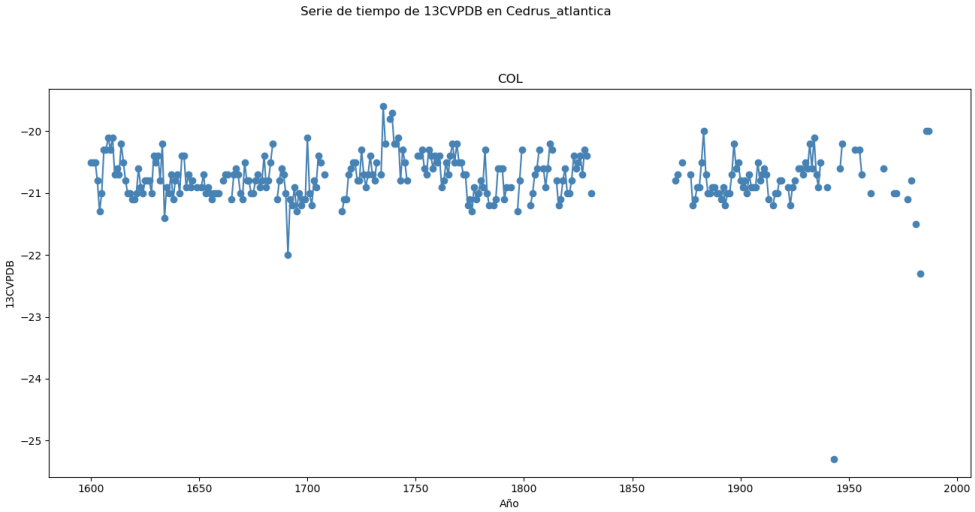

Figura 2: Serie de tiempo correspondiente a la columna COL con un periodo largo de `NA` y datos menores que los demás.

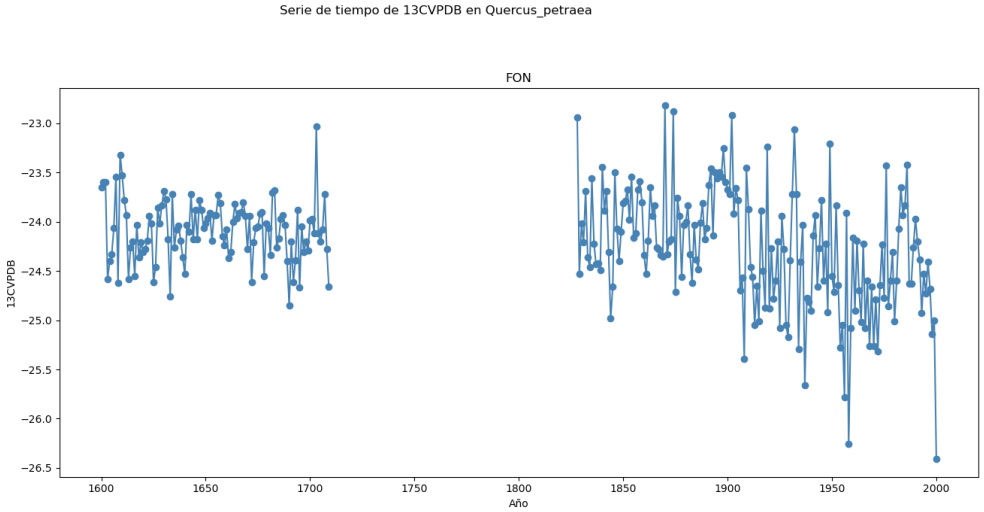

Figura 3: Serie de tiempo correspondiente a la columna FON con un periodo largo de `NA` y datos menores que los demás.

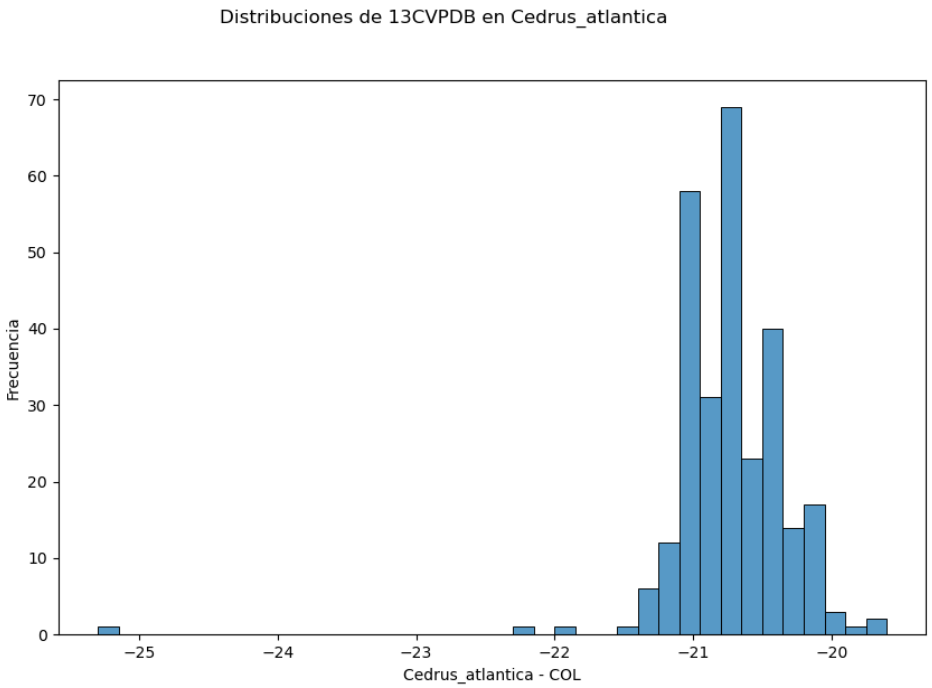

Figura 4: Histograma de la serie presentada en la figura 2.

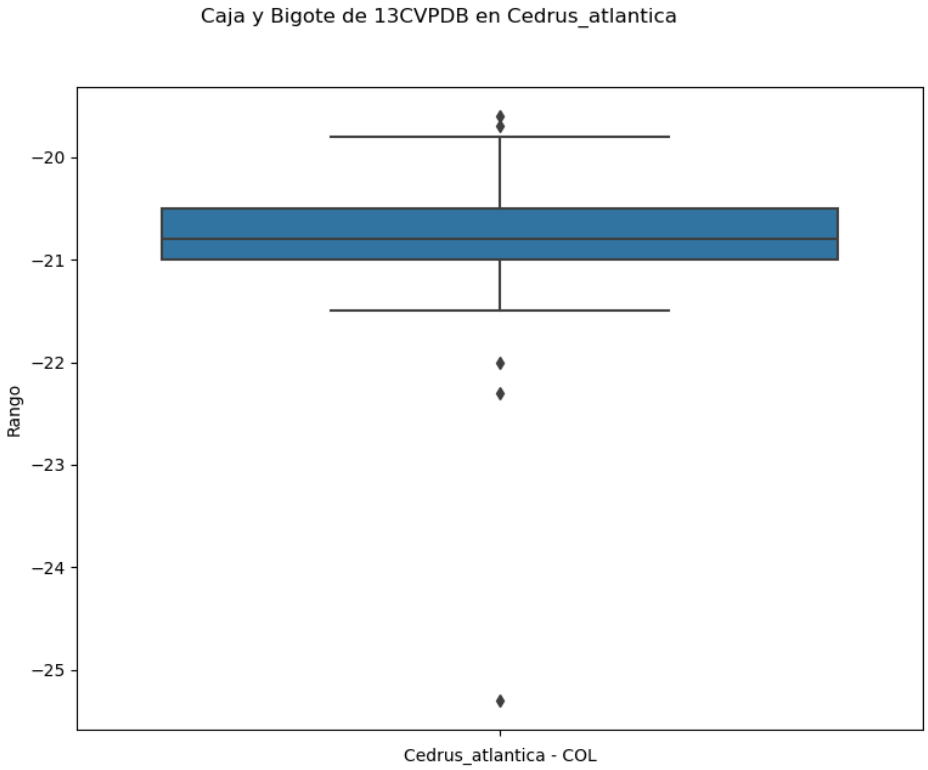

Figura 5: Box-plot de la serie presentada en la figura 2.

4. Manejo de datos faltantes y de *outliers*

4.1. Datos faltantes

Estos datos son fáciles de identificar ya que son entradas vacías o con `NA` dentro de la base de datos. Se sugiere en general que los datos faltan de manera *aleatoria* (como un mecanismo *Missing-at-random*) y no *completamente aleatoria*, pues la mayoría de los datos faltantes se deben a que el estudio aún no comenzaba y esto hace tener alguna dependencia temporal (a la variable *año*).

Descartando lo anterior aún quedan datos faltantes. En la tabla 3 se presentan un resumen de los datos faltantes durante el periodo del estudio. Para estos datos se sigue considerando que faltan de manera aleatoria. Esta decisión se toma considerando situaciones como las de las figuras 1 y 2. El periodo en que faltan los datos es continuo, y en vista del comportamiento regular del fenómeno, sería poco probable un cambio estructural drástico en los datos, por lo que es natural suponer que ocurrió algo que imposibilitó recabar los datos.

Como ya se comentó, se espera regularidad en los datos a lo largo del tiempo. Para imputar datos se proponen métodos que no alteren o sesguen esta regularidad, tales como:

- Imputar la media de los valores adyacentes. El problema de este método es que hace imposible la opción de que el dato no registrado estuviera en realidad fuera del rango de los adyacentes
- Imputar un valor, tomando un nuevo valor con probabilidad uniforme entre los cuartiles 1 y 3. La idea detrás de esto es permitir que se tomen valores no necesariamente cercanos a los valores adyacentes pero que sea representativo del resto de datos

Por ejemplo, para el volumen de datos faltantes de 2 se decide no hacer ningún proceso de imputación. Al hacer falta tantos datos, el proceso de imputación sesgaría cualquier inferencia que se buscara hacer sobre este conjunto de datos. Por lo tanto, dependiendo del problema que se quiera resolver, se sugeriría realizar inferencias por separado, una para cada periodo.

4.2. *Outliers*

Además de con los recursos gráficos, para identificar posibles *outliers* se hicieron diferentes reescalados de datos: Z-score, min-max (ambos sensibles a outliers) y robusto con rango intercuartil. A partir de éstos se estableció una regla de decisión para decidir si una observación es un *outlier* o no. En la tabla 4 se presenta el número de outliers por columna obtenidos con cada método propuesto.

En caso que un dato sea clasificado como *outlier*, se descarta y se imputa un nuevo dato con alguno de los métodos propuestos. Un factor contextual que ayuda a decidir si un dato es *outlier* o no es la regularidad ya discutida en puntos anteriores (en el contexto de concentraciones de isótopos de carbono-13), por la cual se esperaría que los cambios en tendencia no fueran tan bruscos y así los *outliers* corresponderían a *errores de medición.*

Es importante mencionar que los outliers también se pueden decidir a partir de alguna pregunta de investigación, en efecto, al definir el objetivo del estudio, según la definición clásica, los *outlier* son datos que levantan sospechas de haber sido generados o provenir de otro mecanismo distinto al principal [Haw80].

Cuando se deciden quitar datos por ser outliers, se pueden imputan datos en su lugar generados con alguna de las metodologías descritas en la sección 4.1.

Sitio	Outliers (IQR)	Outliers (Z-score)
BRO	0	0
CAV	9	1
CAZ	18	2
COL	5	2
DRA	3	2
FON	6	3
GUT	8	3
ILO	17	9
INA	11	0
AHI	14	4
LAI	3	1
LIL	4	2
LOC	3	2
NIE1	5	4
NIE2	34	7
PAN	7	1
PED	10	3
POE	1	1
REN	0	0
SER	20	1
SUW	0	0
VIG	1	1
VIN	8	0
WIN	0	0
WOB	3	3

Cuadro 4: Número de outliers detectados en cada sitio con los métodos de rango intercuartil (IQR) y de puntaje estándar (Z-score).

5. Conclusiones

Al revisar cuidadosamente una base de datos, con datos recabados por grupos de trabajo distintos, se puede ver que la limpieza de datos es una necesidad antes de cualquier proceso de inferencia. Al ser la limpieza de datos un proceso basado en propuestas y decisiones, es muy importante que se documente el tratamiento que se haga a los datos, ya que esto facilitaría la cuantificación de incertidumbre manejada en la inferencia.

Para el análisis exploratorio de esta base de datos y generación de gráficas se utilizó un código en python; el cual resuelve (limpia) los problemas con la siguiente estructura: 1) Primera visualización, 2) Acomodo de valores y nuevas columnas, 3) Valores faltantes, 4) Reescalamiento y detección de outlier, 5) Gráficas. Al final algunas gráficas se presentan por especie, ya que, intuitivamente, el comportamiento debería ser similar independientemente del sitio.

Si la especie del árbol no es relevante para algún objetivo de estudio, entonces otro método de análisis exploratorio que podría utilizarse sería por localidades, ya que algunas de las columnas corresponden a regiones cercanas geográficamente.

También para los datos faltantes o los outliers podrían revisarse eventos históricos que pudieran afectar la toma de datos. Esto sustentado por la naturaleza de los datos que al tratarse de mediciones de materia orgánica, podrían ser sensibles a actividades industriales, pandemias, guerras, etc.

Sería conveniente conocer algún objetivo específico para alguna investigación, con la finalidad de hacer una mejor limpieza en pro de atacar de mejor manera dichos objetivos.

Referencias

[Haw80] Douglas M. Hawkins. *Identification of Outliers*. Springer, 1980.

[Lib92] Susan M. Libes. *Introduction to Marine Biogeochemistry*. Wiley, 2992.

[Mei18] Wolfram Meier-Augenstein. *Stabel Isotope Forensics: Methods and Forensic Applications of Stable Isotope Analysis*. Wiley, 2018.

[Mem+23] ISONET Project Members, Gerhard Hans Schleser, Laia Andreu-Hayles, Zdzislaw Bednarz, Frank Berninger, Tatjana Boettger, Isabel Dorado-Liñán, Jan Esper, Michael Grabner, Emilia Gutiérrez, Gerhard Helle, Emmi Hilasvuori, Högne Jugner, Maarit Kalela-Brundin, Marek Krapiec, Markus Leuenberger, Neil J. Loader, Valérie Masson-Delmotte, Sławomira Pawełczyk, Anna Pazdur, Rūtilė Pukienė, Katja T. Rinne-Garmston, Antonio Saracino, Matthias Saurer, Eloni Sonninen, Michel Stiévenard, Vincent R. Switsur, Elżbieta Szychowska-Krapiec, M. Szczepanek, Luigi Todaro, Kerstin Treydte, Adomas Vitas, John S. Waterhouse, Martin Weigl-Kuska y Rupert Wimmer. *Stable carbon isotope ratios of tree-ring cellulose from the site network of the EU-Project 'ISONET'*. GFZ Data Services, 2023. DOI: 10.5880/GFZ.4.3.2023.002. URL: https://doi.org/10.5880/GFZ.4.3.2023.002.

[Sch00] Joseph L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman Hall, 2000.