Introducción a Ciencia de Datos

Analisis a las relaciones de isótopos de carbono estables de la celulosa de los anillos de los árboles de la red de sitios del proyecto europeo ISONET

Jessica Rubí Lara Rosales jessica.lara@cimat.mx
Eric Ernesto Moreles Abonce eric.moreles@cimat.mx
Luis Erick Palomino Galván luis.palomino@cimat.mx

Objetivo

El objetivo es reconstruir la variabilidad climática en Europa durante los últimos siglos, para ello usaron como indicador los isótopos estables en los anillos del crecimiento de los árboles; buscando cambios en el clima, impactos de factores ambientales en el crecimiento de los árboles y la reconstrucción de la variabilidad del clima en Europa a partir de registros naturales. Los datos provienen del proyecto ISONET, el cual reunió información de 24 cronologías de isótopos estables de árboles europeos, con una cobertura de 400 años (1600 - 2003) para carbono y oxígeno, y de 100 años para hidrógeno. Donde las variables principales son $\delta 13C$ que se relaciona con la fotosíntesis y la disponibilidad de agua, es decir, indica sobre un estrés hídrico y la eficiencia en el uso de agua en los arboles; $\delta 180$ que habla sobre la transpiración y la fuente de agua, refleja condiciones climáticas como la humedad y temperatura; por ultimo δD asociado con la composición isotópica del agua en los arboles.

Una posible aplicación es en el estudio de cambio climático, relacionando los cambios locales en Europa con eventos climáticos globales. En el proceso poder ver como distintas especies responden a variaciones climáticas.

Esperamos ver que los cambios en $\delta 180$ y δD reflejan la variación de la precipitación a lo largo del tiempo. También ver si la diferencia de altitud, temperatura y humedad afectan a las muestras isotópicas.

Introducción

El parámetro 13C-VPDB es la notación completa que especifica el estándar de referencia contra el cual se compara la proporción de isótopos de carbono 13 y carbono 12 de una muestra. VPDB hace referencia a las siglas Vienna Perr Dee Belemnite el cual es un estándar de referencia internacional basado en los restos fósiles de un belemnoide (un animal marino similar a un clamar) de la formación geológica Pee Dee en Carolina del Sur EE. UU.
En palabras simples: Es la regla.º el ”punto ceroüniversal que se utiliza para medir y expresar si una muestra tiene más o menos Carbono-13 en relación con Carbono-12.

Los análisis de isótopos estables se dan como una desviación de la muestra con respecto al estándar VPDB. La formula es

$$\delta 13C(\,\%) = [(\text{Rsample}/\text{Rstandard}) - 1] \cdot 1000$$

donde Rsample es la proporción 13C/12C de la muestra y Rstandard es la proporción 13C/12C del estándar VPDB.

Descripción datos

En total se tienen 11 variables.

1. **Site Code:** Código de 3 letras para cada sitio, escala nominal.

2. **Site name:** nombre del sitio forestal o pueblo más cercano, escala nominal.
3. **Country:** el país donde se saco la muestra, escala nominal.
4. **Latitude:** coordenadas geográficas, latitud en grados decimales, escala de intervalo.
5. **Longitude:** coordenadas geográficas, longitud en grados decimales, escala de intervalo.
6. **Species:** nombre en latín de las especies de árboles, escala nominal.
7. **First year CE:** primer año (el más antiguo) del registro del sitio de δ13C, escala de intervalo.
8. **Last year CE:** último año (el más joven) del registro del sitio de δ13C, escala de intervalo.
9. **elevation a.s.l. :** elevación promedio del sitio de árboles en metros sobre el nivel del mar, escala de razón.
10. **Year CE:** fecha del anillo de árbol, escala de intervalo.

Las variables que son nombres de pueblos, ciudades o codificación de sitios claramente son datos nominales. Para el caso de escala por intervalo como lo son latitud, longitud y los años de medición son de intervalo, pues el cero no representa la ausencia de lo que se esta midiendo. Finalmente la elevación sobre el nivel de mar es de razón pues este variable puede tomar valores en todos los reales y el cero absoluto si existe.

En total se tienen 25 mediciones de las cuales solo las variables de los años no esta completos.

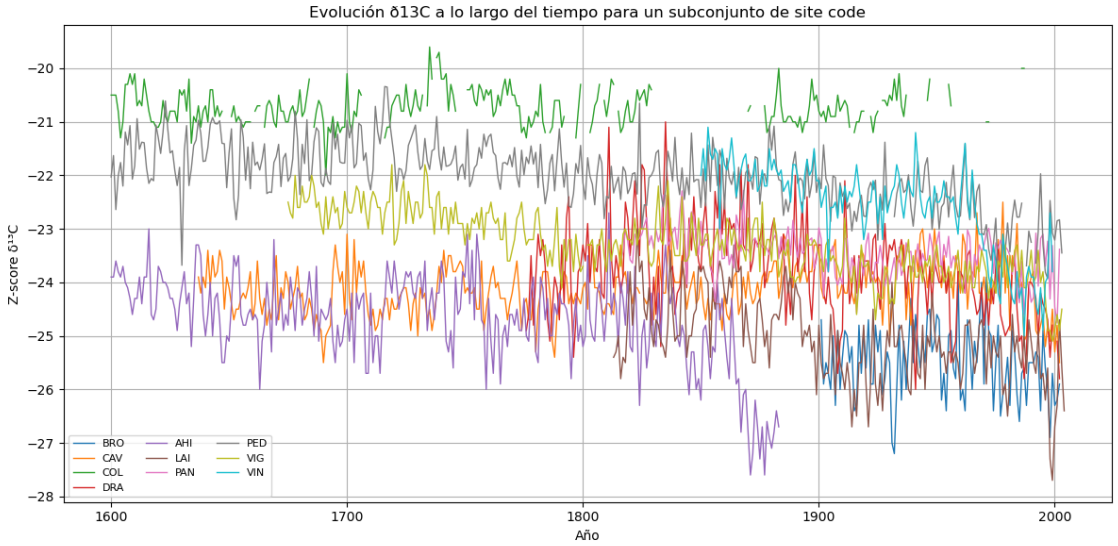


Figura 1: Variación de $\delta 13C$ en los árboles registrados a los largo del los años 1600-2005

Notemos que la Figura 1 nos muestra como los datos no son claros y que por lo mismo, podemos analizarlos de dos maneras diferentes: Si consideramos la muestra que se recolecto de un año XXXX donde sus características serian donde fueron tomadas cada muestra figura 2 o bien considerar la muestra que se recolecto de un árbol en un sitio especifico y sus características serían el año figura 3. Intentaremos analizar los datos de ambas maneras, ya que ambas tienen diferentes objetivos. Analizar las muestras por años nos permite tener registro de cierta manera como

el clima fue cambiando en Europa a lo largo del tiempo, los valores del mismo país se pueden promediar para tener una medición global de cada país. Así mismo si solo quisiéramos medir la variación del clima mediante carbono 13 obtenido de un árbol de cierto país, cambiaríamos el enfoque. Ahora en nuestras muestras las características serian los años.

Características en fila azul, muestras en columna rosa.

Site Code	BRO	CAV	CAZ	COL	DRA	FON	GUT	ILO	INA
Site name	Bromarv	Cavergno	Cazorla	Col Du Zad	Dransfeld	Fontainebleau	Gutuli	Sivakkovaara	Inari
Country	Finland	Switzerland	Spain	Morocco	Germany	France	Norway	Finland	Finland
Latitude	60.00	46.35	37.93	32.97	51.51	48.38	62.00	62.98	68.93
Longitude	23.08	8.6	-2.97	-5.07	9.78	2.67	12.18	31.27	28.31
Species	*Quercus robur*	*Quercus petraea*	*Pinus nigra*	*Cedrus atlantic*	*Quercus petrae*	*Quercus petraea*	*Pinus sylvestri*	*Pinus sylvestris*	*Pinus sylvestri*
First year CE	1901	1637	1600	1600	1776	1600	1600	1600	1600
Last year CE	2002	2002	2002	2000	1999	2000	2003	2002	2002
elevation a.s.l.	5	900	1820	2200	320	100	800	200	150
Year CE	13CVPDB	13CVPDB	13CVPDB	13CVPDB	13CVPDB	13CVPDB	13CVPDB	13CVPDB	13CVPDB
1600	NA	NA	-21.1	-20.5	NA	-23.65	-23.16	-23.8	-26.1
1601	NA	NA	-21.0	-20.5	NA	-23.60	-22.29	-22.6	-26.3
1602	NA	NA	-20.9	-20.5	NA	-23.60	-23.30	-23.4	-25.8
1603	NA	NA	-20.9	-20.8	NA	-24.58	-22.60	-23.3	-25.5
1604	NA	NA	-21.7	-21.3	NA	-24.40	-23.14	-23.9	-25.6
1605	NA	NA	-21.3	-21.0	NA	-24.33	-24.43	-23.3	-25.1
1606	NA	NA	-20.5	-20.3	NA	-24.06	-23.69	-23.2	-24.5
1607	NA	NA	-20.9	-20.3	NA	-23.54	-23.99	-23.8	-24.9
1608	NA	NA	-21.0	-20.1	NA	-24.62	-24.03	-23.6	-25.6
1609	NA	NA	-21.1	-20.3	NA	-23.32	-23.35	-23.3	-25.7
1610	NA	NA	-21.1	-20.1	NA	-23.53	-23.05	-23.1	-25.5
1611	NA	NA	-21.2	-20.7	NA	-23.78	-23.23	-23.2	-25.2
1612	NA	NA	-21.7	-20.6	NA	-23.93	-23.40	-23.0	-24.9

Figura 2

Site Code	1600	1601	1602	1603	1604	1605	1606	1607	1608	1609	1610	1611	1612	1613
BRO	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CAV	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CAZ	-21.1	-21.0	-20.9	-20.9	-21.7	-21.3	-20.5	-20.9	-21.0	-21.1	-21.1	-21.2	-21.7	-21.4
COL	-20.5	-20.5	-20.5	-20.8	-21.3	-21.0	-20.3	-20.3	-20.1	-20.3	-20.1	-20.7	-20.6	-20.7
DRA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
FON	-23.65	-23.60	-23.60	-24.58	-24.40	-24.33	-24.06	-23.54	-24.62	-23.32	-23.53	-23.78	-23.93	-24.58
GUT	-23.16	-22.29	-23.30	-22.60	-23.14	-24.43	-23.69	-23.99	-24.03	-23.35	-23.05	-23.23	-23.40	-23.08
ILO	-23.8	-22.6	-23.4	-23.3	-23.9	-23.3	-23.2	-23.8	-23.6	-23.3	-23.1	-23.2	-23.0	-23.1
INA	-26.1	-26.3	-25.8	-25.5	-25.6	-25.1	-24.5	-24.9	-25.6	-25.7	-25.5	-25.2	-24.9	-25.6
AHI	-23.9	-23.9	-23.6	-23.8	-23.9	-23.7	-24.0	-24.1	-24.3	-24.6	-24.3	-24.3	-23.9	-24.4
LAI	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
LIL	-21.4	-21.7	-21.7	-23.7	-22.4	-22.0	-21.9	-21.9	-21.6	-21.4	-20.7	-22.2	-22.2	-22.1
LOC	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NIE1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NIE2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PAN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PED	-22.03	-21.63	-22.64	-22.01	-21.76	-22.20	-21.19	-21.43	-21.07	-22.07	-21.82	-20.94	-21.51	-21.38
POE	-23.9	-23.9	-23.4	-23.7	-24.1	-23.5	-23.5	-23.1	-22.9	-23.9	-24.3	-22.8	-23.8	-23.1
REN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	-23.07	-22.64	-23.36
SER	NA	NA	NA	NA	-22.2	-22.3	-21.9	-21.2	-21.4	-22.0	-22.2	-21.1	-21.8	-21.3
SUW	-23.5	-23.5	-23.4	-23.6	-23.7	-23.4	-23.3	-23.8	-23.9	-23.5	-23.5	-23.6	-23.0	-22.6
VIG	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
VIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
WIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
WOB	NA	NA	NA	NA	-23.9	-24.2	-24.1	-22.7	-23.6	-23.2	-23.5	-23.1	-23.4	-24.7

Figura 3

Detección de problemas en los datos

Para tratar el problema de datos faltantes o outliers, podemos seguir diferentes estrategias para mejorar el análisis sin meter sesgos. ¿Qué podemos hacer con los años con muchos NA? Si hay pocos NA, podemos interpolar usando lineal o la media; si hay muchos NA, no conviene imputar porque introducimos sesgos, por lo que trabajaremos solo con los periodos donde si se solapen todos los países.

Una pregunta interesante es: ¿Por qué bajo MCAR la eliminación de casos completos es insesgada, pero menos eficiente? La pérdidad de eficiencia viene porque al eliminar casos completos estamos usando menos información, es decir, el número efectivo de observaciones es aleatorio y en promedio menor que los datos. Por lo que, bajo MCAR, no introduces sesgo porque la ausencia de datos no está relacionada con los valores, pero pierdes precisión porque el tamaño efectivo de la muestra disminuye, lo que incrementa la varianza del estimador respecto a tener la muestra completa. Lo anteriormente dicho es una consecuencia de lo probado en el Ejercicio 5 de esta Tarea.

Análisis de datos faltantes

Analizándolo de ambas maneras en las siguientes tablas se muestra el porcentaje y la cantidad de valores faltantes.

	BRO	CAV	CAZ	COL	DRA	FON	GUT	ILO	INA	AHI	LAI	LIL	LOC
No.	304	41	3	126	180	123	3	3	3	122	214	7	151
%	74.87	10.09	0.73	31.03	44.33	30.29	0.73	0.73	0.73	30.04	52.7	1.72	37.19

	NIE1	NIE2	PAN	PED	POE	REN	SER	SUW	VIG	VIN	WIN	WOB
No.	2	2	3	5	3	39	6	1	78	256	171	7
%	0.49	0.49	0.73	1.23	0.73	9.6	1.47	0.24	19.21	63.05	42.11	1.72

De ello podemos resaltar que hay 8 muestras que superan el 30 % de datos faltantes.

Para el caso de años entre 1600 y 2005 se tienen un total de 406 años. Así que solo mostraremos aquellos años en los que tienen 10 o más muestras faltantes, recordemos que en total de cada año deberían de haber 25 datos capturados.

	1600	1601	1602	1603	2003	2004	2005
No.	10	10	10	10	14	23	25

Para ayudar más el análisis, se presenta el siguiente heatmap para ver la dispersión de los datos faltantes:

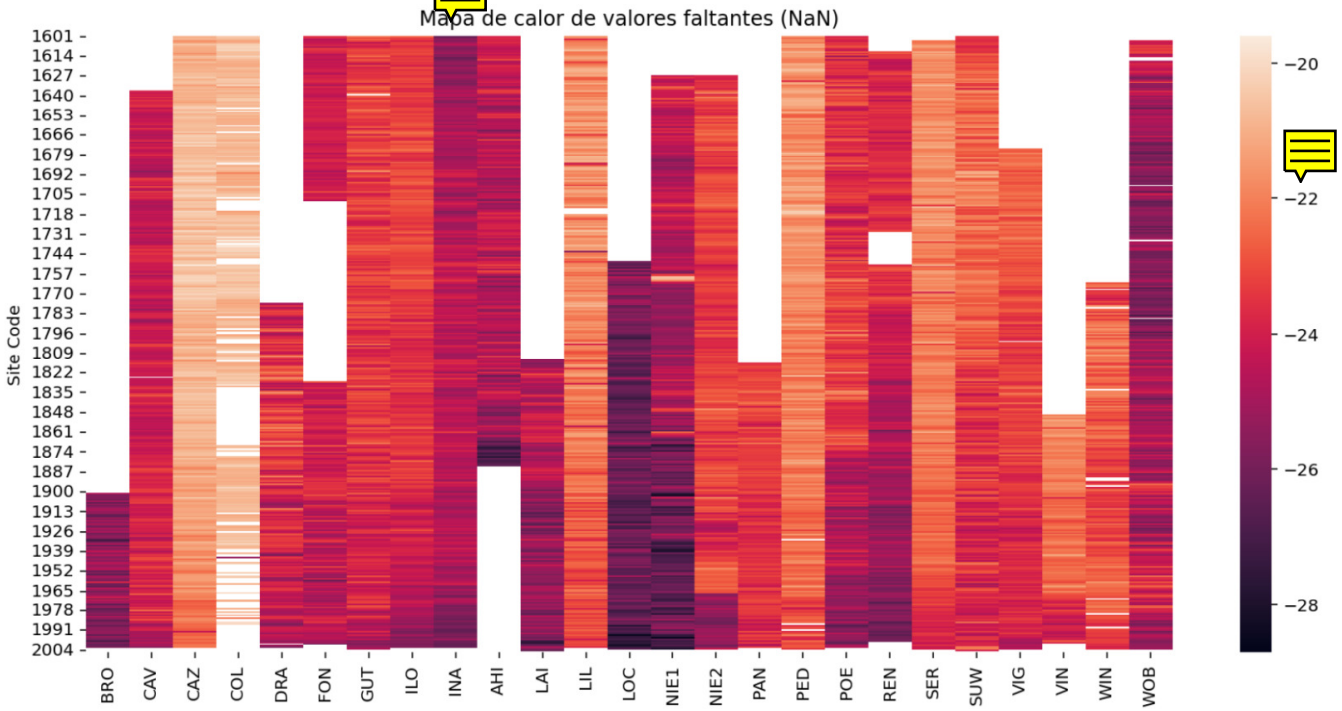


Figura 4: Heatmap de datos faltantes

De manera superficial pareciera que hay muchos datos faltantes de todos los sitios, pero hay otra manera de ver esto: Para cada sitio se da un año inicial y un año final de donde se toman mediciones, y si tomamos esto en cuenta, podemos reescribir nuestra tabla, donde se muestra el porcentaje de datos faltantes y el total de años en los periódos establecidos:

	BRO	CAV	CAZ	COL	DRA	FON	GUT	ILO	INA	AHI	LAI	LIL	LOC
No.	101	365	402	400	223	400	403	402	402	283	191	402	254
%	0	0.27	0	30.17	0	29.43	0.25	0	0	0	0.52	0.99	0

	NIE1	NIE2	PAN	PED	POE	REN	SER	SUW	VIG	VIN	WIN	WOB
No.	376	376	186	403	402	387	399	404	328	149	240	399
%	0	0	0	0.74	0	5.41	0	0	0.3	0	2.49	1.25

Con este panorama, vemos que técnicamente hablando no hay muchos datos faltantes, y salvo por los sitios COL, FON y REN, no se notan grandes vacíos de mediciones. Sería difícil atribuir un solo tipo de datos faltantes a la base de datos, por ejemplo, en los sitios FON y REN hay considerables períodos de tiempo donde no tenemos mediciones, y esto sugiere algún fallo estructural en la recolección de datos. Mientras tanto, el sitio COL le faltan aproximadamente el 30 % de sus datos, y parecen estar esparcidos durante todo el periodo de tiempo que para el cual se tomaron datos, y probablemente los datos faltantes para este sitio en particular sean MAR. Para todo los demás, tenemos pocos si no es que nada de datos faltantes, y parecen estar relativamente uniformemente esparcidos, y uno podría pensar que estos son MCAR.

Con lo anterior en mente, uno podría preguntar si se puede hacer imputación de los datos. Para los tres sitios con mas datos faltantes no hay una buena opción, quitar los datos nos deja con bloques muy grandes donde nos falta información, mientras que intentar llenarlos con otros métodos podría desviar nuestro modelo original de manera considerable, cambiando por completo los resultados obtenidos. Para los otros nuestras opciones son más variadas, pero se tiene que tener cuidado. Uno podría estar tentado a hacer imputación con la media, pero incluso en el heatmap antes mencionado se puede ver que parece haber tendencias no lineales en los datos conforme avanza el tiempo, e introducir la media en los valores faltantes podría hacer que perdamos de vista esta relación. Se podría considerar algo más flexible, como regresión estocástica, pero incluso esto podría introducir ruido que no hubiera estado presente en el modelo de donde vinieron los datos. Al final uno tiene que hacer una elección sabiendo que no hay sistema de imputación que permita llenar los datos faltantes como vinieran del modelo original, pues de poder hacer eso no necesitaríamos estudiar los datos.

Outliers

Para la detección de outliers, usaremos dos mecanismos distintos.

Boxplot

Este mecanismo nos permite identificar los posibles outliers de manera visual. En este gráfico los posibles outliers son aquellos que se encuentran más allá de los limites de los bigotes. Para nuestro datos se obtuvieron las siguientes gráficas.

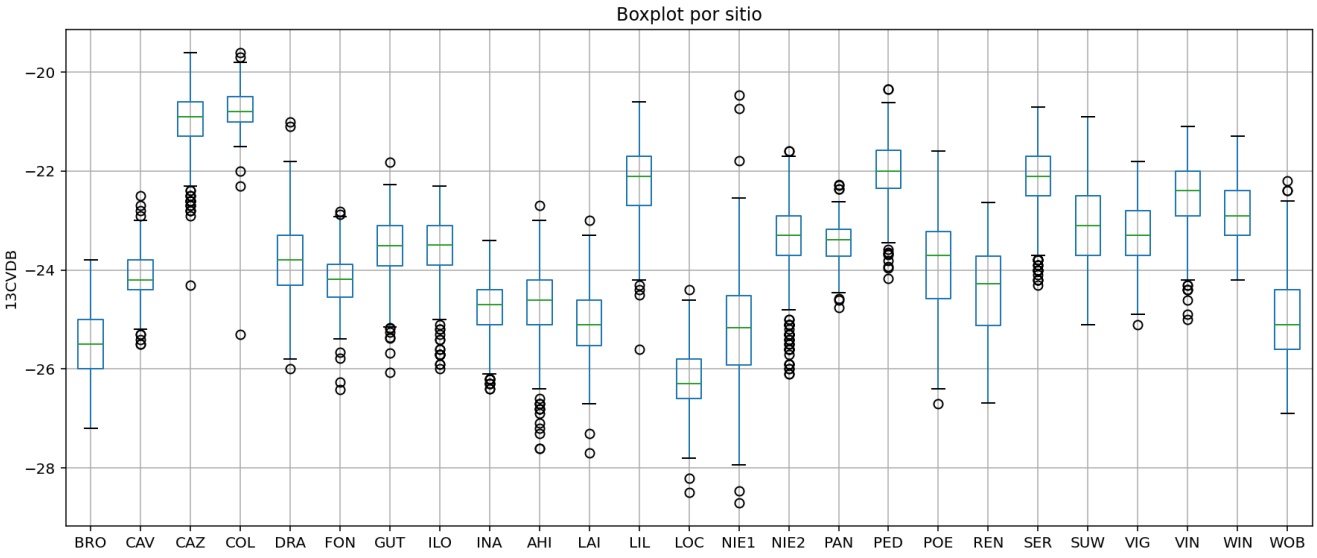


Figura 5: Boxplot de los 25 sitios

En esta gráfica podemos notar como CAV, CAZ, GUT, ILO, INA, AHI, NIE2, PED y SER muestran una gran cantidad de posibles outliers. Dada esta información podríamos verificar si realmente se tratan de outliers con otros métodos mas robustos ya que quitarlos no seria muy viable pues son bastantes. Otra opción seria mantenerlos y para quitar cierta variabilidad en el modelo se podría contrarrestar con alguna transformación.

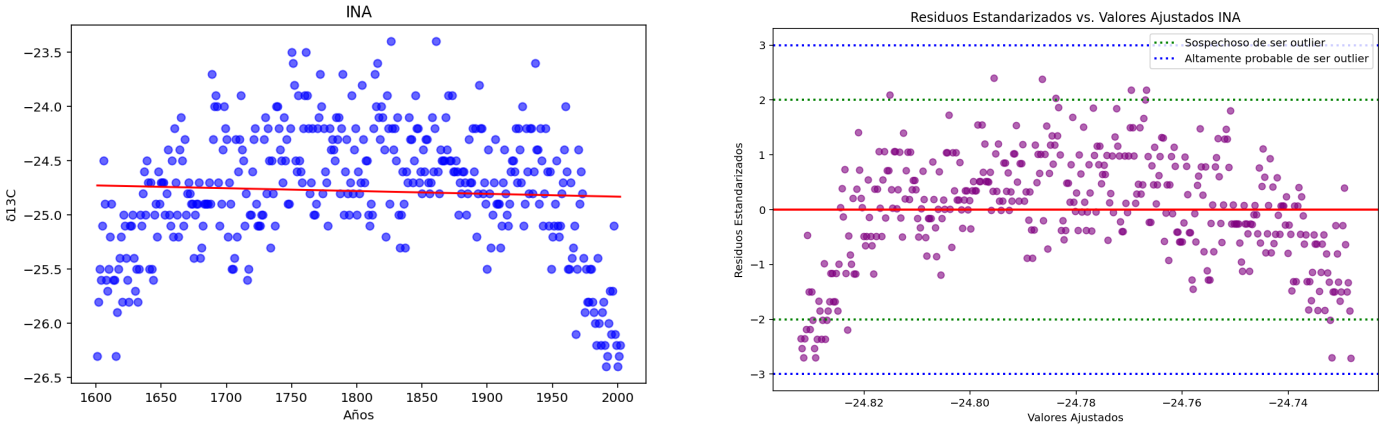
También en estas gráficas es posible medir la variabilidad de las medias en los distintos lugares y con ello podemos visualizar como CAZ y COL pareciera que varían más que la otras. Un posible análisis extra podría ser el ver que tan diferentes son las medias pero clasificando los países con latitudes, longitudes y/o nivel del mar similares.

Regresión lineal

El segundo método que consideramos fue el de regresión lineal el cual nos permite detectar posibles outliers y/o

observaciones influyentes. Esto se puede hacer mediante la gráfica de residuos estandarizados contra valores ajustados y haciendo bandas de confianza de 2 (valores sospechosos de ser outliers) y 3 (altamente probables de ser outliers). Estos son valores críticos obtenidos de una aproximación t-student.

Tomando como variable independiente los años y variable dependiente las mediciones de δ13C de uno de los 25 sitios, obtuvimos 25 regresiones. De entre todas las gráficas podemos rescatar ciertos patrones no lineales que seguían algunas muestras como se ve en la figura 6, de ello que también la clasificación de los datos outliers sean varios y no sea tan creíbles.

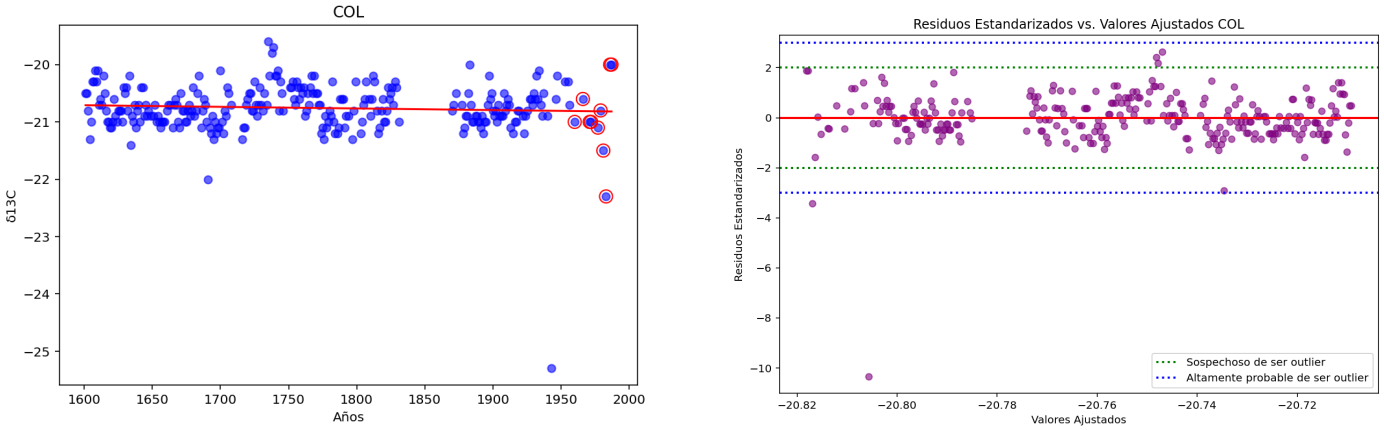


(a) Leverege alto

(b) Estandarizados vs ajustados

Figura 6: Tendencia no lineal

Por otro lado, también resultaron muestras con observaciones influyentes como se ven en la figura 7.



(a) Leverege alto

(b) Estandarizados vs Ajustados

Figura 7: Observaciones influyentes

Codificación y escalamiento

Uno de los problemas que tuvimos al manejar la tabla de datos son las variables categóricas, ya que necesitábamos identificarlas con pandas, como por ejemplo NA, las especies de los arboles, los países y los site code. En nuestro caso, fue necesario usar los métodos label enconding y one- hot. Usamos label enconding para asignar un número entero a cada categoría y así, cuando el orden puede tener sentido o el algoritmo no se ve afectado por la magnitud.

Usamos one-hot para crear una columna binaria por cada categoría, así al usar regresiones lineales evitamos que el algoritmo interprete los números como orden.

Dado que tenemos series de $\delta 13C$ de diferentes árboles y países, con rangos y medias distintas, aplicar transformaciones sirve para compararlos entre ellos, ya que z-score centra y escala los datos, permitiendo comparar variaciones relativas. También z-score nos permite ver anomalías de manera estandarizada

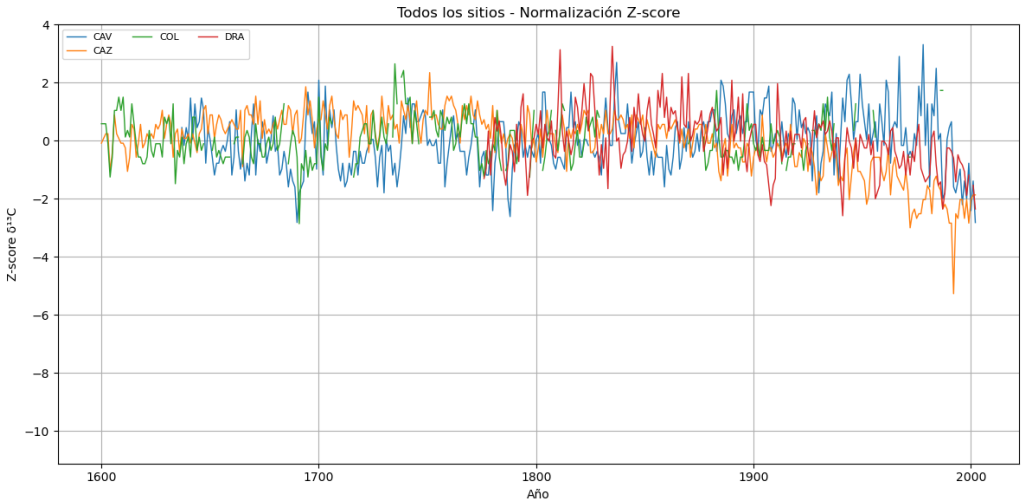


Figura 8: Normalización z-core para CAV, CAZ, COL, DRA

Donde podemos ver como algunos alcanzan valores mayores a 3, indicando un posible outlier. Por otro lado, la normalización Min-Max es útil para la visualización y comparación de patrones, ya que pone las series en $[0, 1]$ resaltando solo la forma de la curva.

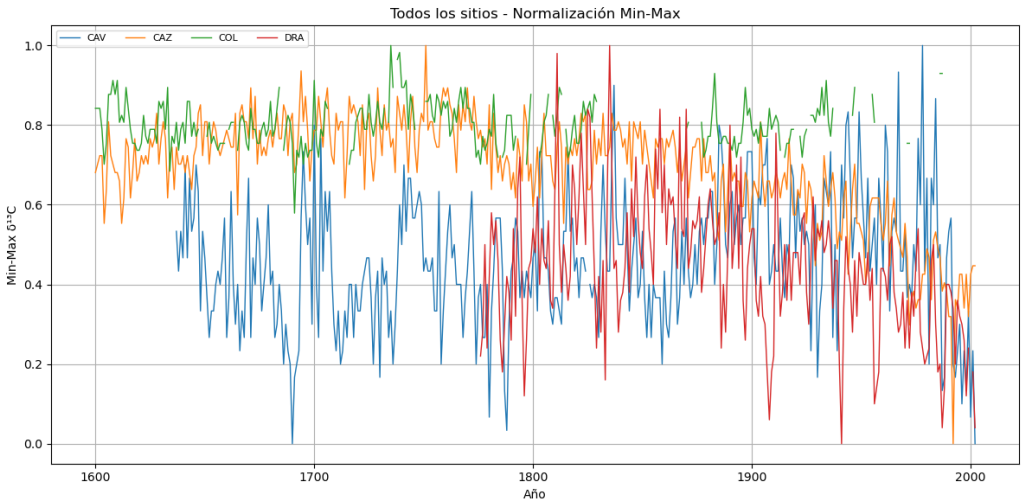


Figura 9: Normalización Min-Max para CAV, CAZ, COL, DRA

Visualización exploratoria

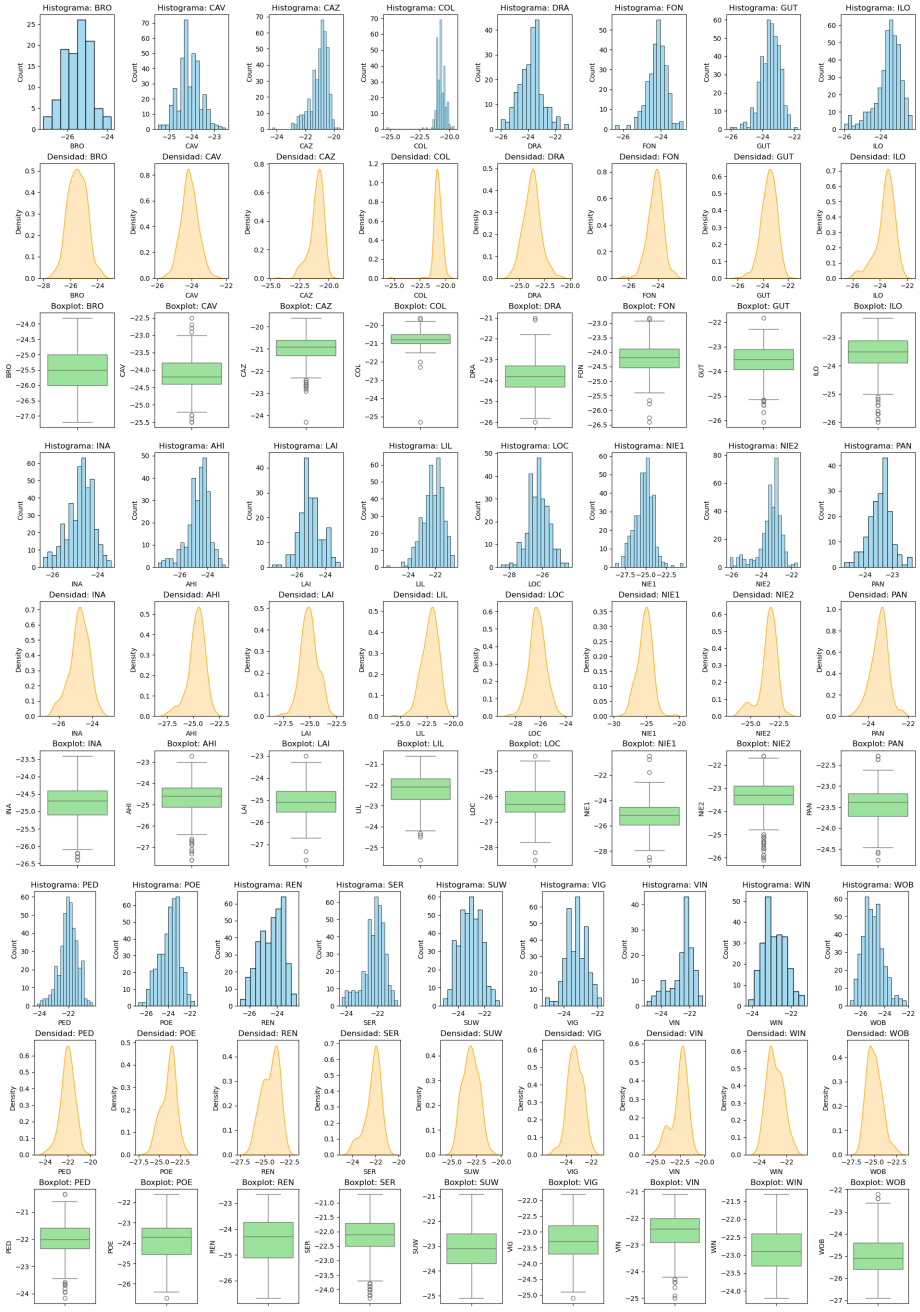


Figura 10: Histogramas, Gráfica de Densidades y Boxplot de todos los datos

Para realizar un análisis exploratorio de los datos, se examinarán los histogramas, las densidades y los boxplots de cada variable. Los histogramas permiten dividir los datos en intervalos y mostrar la frecuencia de valores en cada rango, lo que facilita diagnosticar la distribución de la variable y determinar si es simétrica o presenta sesgos. Asimismo, barras aisladas del resto de la distribución pueden señalar posibles valores atípicos.

Las densidades estiman de manera suave la función de densidad de la variable, lo que permite identificar picos,

colas y posibles subpoblaciones, ofreciendo una visión más detallada de la distribución que los histogramas.

Finalmente, los boxplots representan los cuantiles y la mediana de la variable, mostrando el rango típico de los datos y destacando los puntos fuera de los bigotes como posibles outliers. Esta combinación de gráficos proporciona una visión completa de la distribución, la dispersión y la presencia de valores extremos en los datos.

Reflexión crítica

¿Como influyen las decisiones de limpieza, imputación, codificación y escalamiento en la etapa posterior de modelado estadístico? La limpieza de datos la podemos interpretar como eliminar duplicados, corregir errores de captura y tratar valores extremos-outliers o nulos; por lo que, si no los limpiamos podríamos tener inconsistencias que puedan generar un sesgo o ruido en los modelos. La imputación de datos faltantes cambia la distribución de la variable, lo que afecta la varianza y correlaciones; así, si se hace de forma inapropiada se puede introducir sesgo que afecte predicciones. Por otro lado, el escalamiento de variables ajusta sus rangos, la que nos ayuda en algoritmos sensibles a magnitudes.

Estas decisiones influyen directamente en la calidad del modelo, su capacidad de generación y la interpretación de los resultados, por lo que una preparación de datos es fundamental en el análisis de cualquier modelo.