Análisis de Datos Tarea 1

Stable carbon isotope ratios of tree-ring cellulose from the site network of the EU-Project 'ISONET'

Brain de Jesús Salazar, César Ávila, Iván García

12 de septiembre de 2025

Resumen

En este proyecto se analizan los datos obtenidos de `https://doi.org/10.5880/GFZ.4.3.2023.002`, provenientes del proyecto *Stable carbon isotope ratios of tree-ring cellulose from the site network of the EU-Project 'ISONET'*. Se discute su origen, propósito y se incluye una exploración inicial de los datos, detectando problemas, sugiriendo estrategias para el manejo de datos faltantes, la codificación y escalamiento, y una visualización exploratoria. Además, se hace una breve discusión de cómo es que las distintas decisiones del preprocesamiento de los datos pueden influir en la etapa posterior del modelado estadístico.

1. Exploración inicial de los datos

La base de datos contiene dataciones de la proporción de isótopos de carbono estables (δ^{13}C) presentes en la celulosa de árboles. Estos datos se obtuvieron, con apoyo de la Comisión Europea, en el marco del proyecto ISONET (400 años de reconstrucciones anuales de la variabilidad climática europea utilizando una red isotópica de alta resolución) mediante el análisis de anillos de crecimiento en árboles; se obtuvieron dataciones anuales de δ^{13}C desde el año 1600 hasta el 2004 para 25 ubicaciones distintas distribuidas en la Unión Europea y Marruecos. Los datos se recabaron para crear una amplia red espacio-temporal de isótopos estables en los anillos de los árboles en toda Europa con el fin de mejorar la comprensión de los sistemas climáticos europeos.

Para cada ubicación se cuenta con 9 filas que corresponden a los datos del lugar:

- Un código de tres letras para cada ubicación. Esta es correspondiente a una variable nominal, puesto que son categorías sin orden. En total fueron 25 ubicaciones distintas.
- Nombre del bosque o poblado cercano. Esta también es correspondiente a una variable nominal, y solo se repite en Niepolomice (NIE1 y NIE2), pero los datos son distintos.
- Nombre del país. Una vez más, esta variable es nominal, y son 12 distintos: Finlandia, Suiza, España, Marruecos, Alemania, Francia, Noruega, Austria, Reino Unido, Polonia, Lituania e Italia.
- Dos entradas con las coordenadas geográficas (latitud y longitud) del lugar. Esta escala de medición es intervalar, puesto que no tiene sentido hablar de razones entre ubicaciones geográficas.

- La especie de árbol de la cual se tomaron las mediciones, la cual es una variable nominal, y hay 7 distintas: *Pinus sylvestris*, *Quercus petraea*, *Quercus robur*, *Pinus nigra*, *Pinus uncinata*, *Cedrus atlantica* y *Pinus leucodermis*.

- El primer y último año del cual se tiene datación para el lugar correspondiente, que corresponden a variables en escala de medición intervalar, puesto que no tiene sentido hablar de razón entre fechas.

- Elevación promedio en metros sobre el nivel del mar, que nuevamente es intervalar.

Finalmente, para cada ubicación, tenemos una serie de dataciones de δ^{13}C con su año correspondiente, cuyas cantidades se muestran en la siguiente sección. Los valores faltantes se indican como NA.

Las unidades del marcador isotópico δ^{13}C son un ratio, respecto a un estándar, expresado por mil, es decir,

$$\delta^{13}C = \left(\frac{(C^{13}/C^{12})_{\text{muestra}}}{(C^{13}/C^{12})_{\text{estándar}}} - 1 \right) \times 1000.$$

Para este proyecto, el estándar está dado por Viena PDB (VPDB). En principio δ^{13}C puede tomar valores en los números reales, pero al tratarse de dataciones de árboles usualmente se encuentra entre -19 y -30.

Mediante el análisis de δ^{13}C y otras proporciones de isótopos estables, es posible reconstruir características climáticas como la temperatura, humedad relativa y características de las precipitaciones.

2. Detección de problemas con los datos

Los datos de cada ubicación no presentaron problema alguno, todos los datos faltantes se encuentran en las mediciones de δ^{13}C. El porcentaje de dataciones faltantes es muy elevado si consideramos el rango completo de 405 años; sin embargo, la mayor parte de datos faltantes se encuentra en los extremos del rango 1600-2003.

Por esto se decidió hacer un análisis para cada región del número de dataciones faltantes dentro del rango de años en el que se cuenta con registros. Esto es importante en la posterior imputación de los datos, ya que la visualización de gráficos de dispersión locales nos sugiere tendencias a lo largo de los años en las medidas de δ^{13}C , por lo cual sería muy arriesgado aventurarse a imputar dataciones fuera de las fechas para las cuales se tiene registro. Los valores faltantes están registrados como NA, en algunas ocasiones se registró la palabra con uno o varios espacios y se sustituyeron por NA. Los porcentajes se muestran en el Cuadro (1).

Código	% Missing Data	(Años faltantes / Total)
BRO	0.00 %	0/102
CAV	0.27 %	1/366
CAZ	0.00 %	0/403
COL	27.84 %	108/388
DRA	0.44 %	1/227
FON	29.43 %	118/401
GUT	0.25 %	1/404
ILO	0.00 %	0/403
INA	0.00 %	0/403
AHI	0.00 %	0/284
LAI	0.00 %	0/192
LIL	0.99 %	4/403
LOC	0.00 %	0/255
NIE1	0.00 %	0/404
NIE2	0.00 %	0/404
PAN	0.00 %	0/403
PED	0.74 %	3/404
POE	0.00 %	0/403
REN	5.41 %	21/388
SER	0.00 %	0/400
SUN	0.00 %	0/405
VIG	0.30 %	1/329
VIN	0.00 %	0/150
WIN	2.49 %	6/241
WOB	1.24 %	5/404

Cuadro 1: Porcentaje y número de años con datos faltantes por serie.

Para la detección de outliers se pueden usar distintos métodos, tales como el Z-score o por Rango Inter-cuartílico (RIQ). Dado que las mediciones registradas corresponden a un promedio de observaciones de 5 árboles distintos tomados de manera aleatoria anualmente en cada lugar, es de esperarse que los datos sigan una distribución aproximadamente normal, por lo que, para la detección de outliers, el método Z-score puede ser el indicado. Sin embargo, para garantizar su funcionamiento, ejecutamos una prueba de Anderson-Darling para medir la normalidad, obteniendo que no se puede rechazar la normalidad (al 5 % de significancia) únicamente para las regiones BRO, CAV, DRA, LAI, LOC y VIG. Por lo tanto, esto sugeriría que los mecanismos de detección de outliers deberían seguir métodos distintos, siendo posiblemente más apropiado el análisis por Rango Inter-cuartílico o por la Hat Matrix. Sin embargo, en la Figura (1) se muestra la comparación de la cantidad de outliers detectados por Z-score y por RIQ. Podemos notar que el método RIQ detecta una mayor cantidad de outliers, y es razonable, puesto que el método Z-score no es muy fuerte cuando los datos no siguen una distribución normal. Sin embargo, en las regiones que sí pasan la prueba de Anderson-Darling, el número de outliers detectados por ambos métodos son similares. Por otro lado, la Hat Matrix no es capaz de detectar los outliers de manera satisfactoria usando la regla práctica de corte $2p/n$, pero no se cuentan con suficientes justificaciones teóricas para cambiar dicha regla y permitir la identificación de más outliers, ya que se podría introducir un sesgo si se cambia el threshold después de haber analizado los datos.

Por último, las únicas inconsistencias o codificación ambigua encontrada fueron las relacionadas con los datos faltantes, que en ocasiones se registraban con NA con uno o varios espacios al

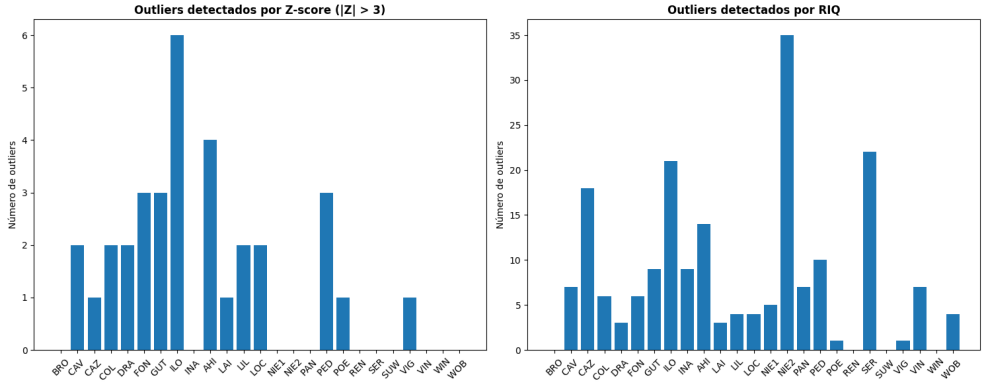


Figura 1: Comparación de métodos de detección de outliers.

final, lo cual fue corregido en el programa para evitar que esto afectara el análisis. Por otro lado, las especies no presentaban errores de codificación, ni los países, de modo que fue fácil identificar cuáles eran los distintos que había.

3. Manejo de datos faltantes

De primera impresión, y observando los datos por lugar, se puede pensar que el número de datos faltantes es superior a el número de datos observados. Sin embargo, hay que considerar que se tiene un registro de primer año de mediciones y último año de mediciones, en este sentido se puede tomar este intervalo de tiempo como el periodo real en el que se llevaron acabo las mediciones. Bajo esta forma de analizar los datos se tiene que los porcentajes de datos faltantes por lugar son los mostrados en el Cuadro (1).

Observemos que de esta manera, el porcentaje de datos faltantes de mayoría de los casos es pequeño, mientras que en solo dos casos (COL y FON) el número de datos faltantes es elevado, en estos dos casos la imputación no es recomendable, ya que se tendría un sesgo grande en los datos. Además, una exploración inicial de los datos se puede observar en la Figura (2).

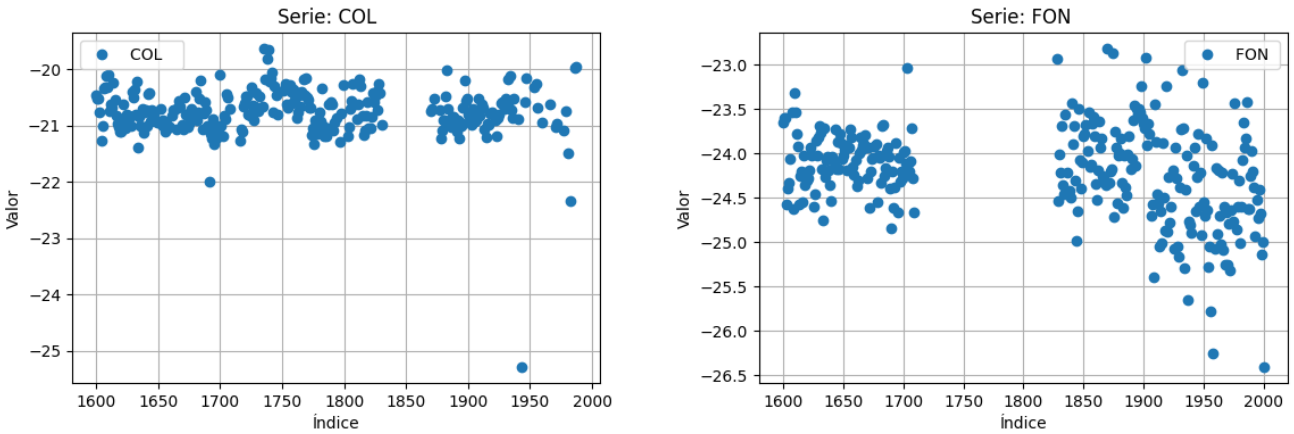


Figura 2: Datos observados de Col Du Zad y Fontainebleau.

En ambos casos se puede ver que la mayoría de los datos faltantes se encuentran agrupados en un periodo de tiempo, por lo que una opción puede ser trabajar con los datos que se encuentran después de ese periodo de datos faltantes.

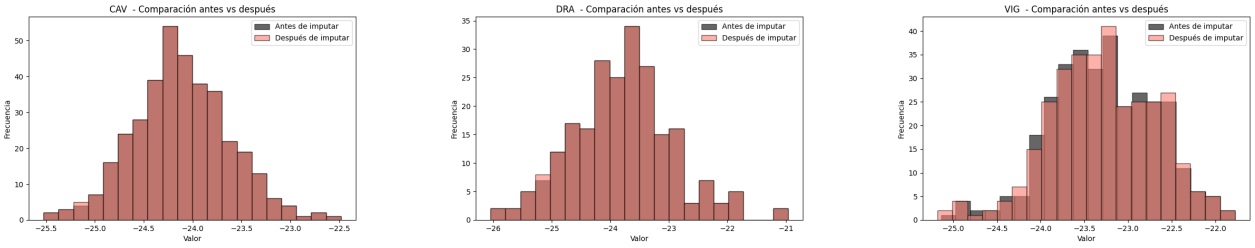
Por otro lado, notemos que los datos de concentración isótopos recolectados vienen de obtener muestras de cinco árboles en una misma región, encontrar su concentración de isótopos y registrar el valor promedio de esas cinco muestras. Buscando un poco más de información sobre el proceso para encontrar la concentración de isótopos, se pudo encontrar que dicha concentración depende de la cantidad de agua que haya en la región y del tipo de árbol que se esté midiendo. Por ejemplo, en tiempos de sequía, los anillos de los árboles adelgazan, según la especie del árbol, lo que provoca que la toma de la muestra no se pueda realizar. En este sentido, se puede pensar que se esta frente a un caso de datos faltantes del tipo MAR. En este documento se propone utilizar tres tipos de imputación, en los casos en que el número de datos faltantes es pequeño.

El primer tipo de imputación que se propone es mediante máxima verosimilitud, observando algunos histogramas de ciclos en los que faltan datos, se puede observar que estos tiene un comportamiento similar a los de una distribución normal. Por ello se realizó la prueba de normalidad de Anderson Darling de donde se obtuvieron los resultados del Cuadro (2).

Nombre	Estadístico	Valor crítico	Decisión
BRO	0.365	0.759	No se rechaza normalidad (al 5 %)
CAV	0.361	0.779	No se rechaza normalidad (al 5 %)
DRA	0.548	0.774	No se rechaza normalidad (al 5 %)
LAI	0.542	0.771	No se rechaza normalidad (al 5 %)
LOC	0.386	0.775	No se rechaza normalidad (al 5 %)
VIG	0.472	0.778	No se rechaza normalidad (al 5 %)

Cuadro 2: Resultados del test de normalidad para las series donde no se rechaza la normalidad.

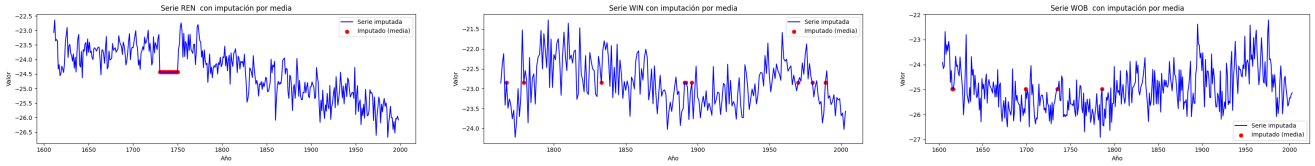
Se puede ver que para los únicos casos en los que no se rechaza la prueba y tienen datos faltantes son en CAV, DRA y VIG. Por ello, para estos casos, se estimó la media muestral y la varianza muestral para después simular el número de datos faltantes en cada caso como provenientes de una normal con estos parámetros. Una vez que se hace esto, se imputan estos datos y se grafican los histogramas de la Figura (3).



(a) CAV (b) DRA (c) VIG

Figura 3: Imputación con normalidad para diferentes series.

Otra forma en la que se pueden imputar los datos, es utilizando la media muestral, en cuyo caso se tienen resultados como en la Figura (4).

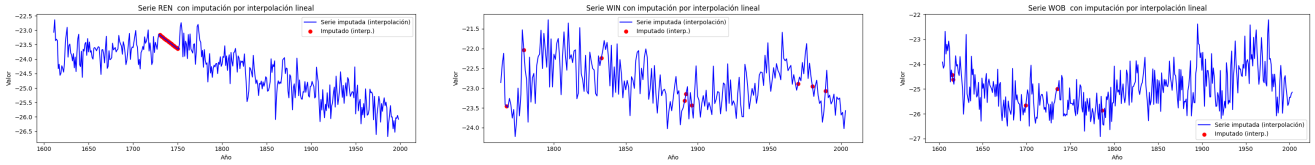


(a) REN (b) WIN (c) WOB

Figura 4: Imputación con media para diferentes series.

Observemos que este tipo de imputación tiene buenos resultados cuando el número de datos faltantes es pequeño y cuando no se tiene un número grande de datos faltantes seguidos. En el caso de REN, se puede observar que imputar mediante la media no parece ser la mejor opción, ya que no se sigue ningún patrón como los que se ve en los datos originales.

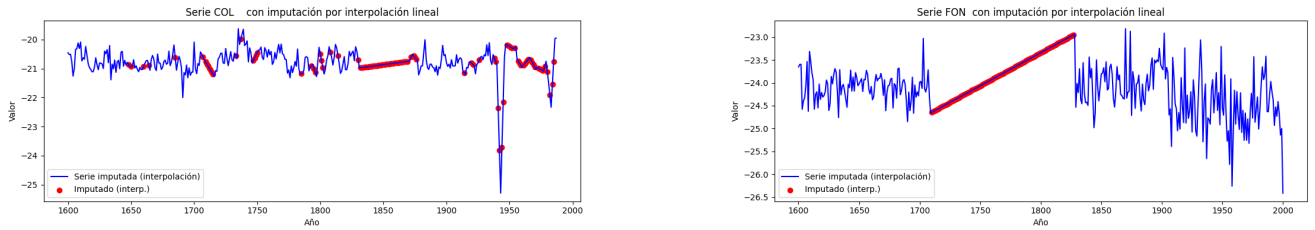
Una última propuesta es realizar imputación mediante interpolación, y en este caso para datos faltantes dispersos sigue dando una buena respuesta. Sin embargo, para intervalos de tiempo en los que hay muchos datos faltantes se sigue observando la falta de naturalidad de los datos. Los resultados se muestran en la Figura (5).



(a) REN (b) WIN (c) WOB

Figura 5: Imputación con interpolación para diferentes series.

Sin embargo, observemos que para series en las que el número de datos faltantes es grande los métodos anteriores no proporcionan buenos resultados, como lo muestra la Figura (6). Como



(a) COL (b) FON

Figura 6: Comparación de imputación por interpolación en diferentes sitios.

podemos observar, los diferentes tipos de imputación muchas veces dependen del objetivo del estudio que se tenga en mente. Formalmente, si se estuviera bajo MCAR, la eliminación de datos completos es insesgada, lo cual se sigue de los ejercicios teóricos 4 y 5 de la tarea, en donde se probó que $\mathbb{E}\left[\overline{Y}_{obs}\right] = \mu$, y que $L_{obs}(\theta) \propto \mathbb{P}\left[Y_{obs} \,|\, \theta\right]$, por lo que la eliminación de casos completos produce un estimador insesgado, y es posible ignorar el mecanismo de faltantes para la inferencia de los parámetros. Sin embargo, en el ejercicio 5 también se probó que la varianza tiende a ser más grande, por lo que estos estimadores son menos eficientes.

En el caso de la base de datos trabajada, no existen observaciones con absolutamente todos los datos, así que la eliminación de casos incompletos nos dejaría en una posición donde no hay ninguna información para hacer inferencia estadística, además de que es muy poco probable

que el mecanismo de faltantes sea del tipo MCAR. Por otro lado, la imputación muchas veces depende del objetivo del estudio: si se requiere analizar todos los datos durante el mismo periodo, probablemente haya que hacer una imputación completa, y en este caso las mejores estrategias serían el muestreo normal (si la prueba de Anderson-Darling no rechaza dicha hipótesis), o la imputación por media, para tratar de conservar las "tendencias". Sin embargo, si se requiere analizar de manera individual, entonces se puede reducir el enfoque a solo los periodos en donde hay casos tomados de manera continua, y como vimos, en este caso el porcentaje de faltantes es poco, por lo que cualquiera de los métodos anteriormente mencionados debería funcionar de una manera adecuada.

4. Codificación y escalamiento

Observemos que en este caso, las variables categóricas que se tienen son, "Site name", "Country" y "Species". En este caso, se sabe que las opciones que se tienen para realizar las transformaciones son *one-hot* o codificación ordinaria, sin embargo no es recomendable utilizar una codificación ordinaria ya que esto agrega cierto orden el cual no necesariamente se tiene. Por otro lado, recordemos que debido a la naturaleza del proceso que se lleva a cabo para recolectar datos, las variables categóricas que interesa transformar son "Site name" y "Species", esto debído a que los datos faltantes pueden depender de estas variables. Realizando estas transformaciones se tiene una base de datos de la forma del Cuadro (3).

	Year	13CVPDB	BRO	CAV	CAZ	COL	DRA	FON	GUT	...	POE	REN	SER	SUN	VIG	VIN	WIN	WOB
360	1960	-26.1914	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
361	1961	-25.99718	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
362	1962	-26.441466	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0

Cuadro 3: Tabla con variables dummies.

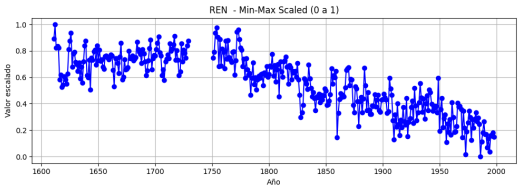
Otra de las variables que se pueden codificar es "Country", y al hacerlo para cada país se obtiene un código en binario en lugar de los nombres. Esto permite que no haya un orden implícito en los datos pero a su vez transformarlos a variables numéricas para que los distintos modelos estadísticos que así lo requieran puedan interpretarlo más fácilmente. También se puede observar que una especie de "codificación" ya se encuentra realizada para los nombres de los sitios, con el identificador de 3 letras que tienen asignado.

Se sabe que lo escalamientos de variables sirven principalmente para homogenizar unidades, evitar sesgos y visualmente ayuda a tener una perspectiva más clara de las cosas. Existen distintas maneras de escalar datos, según sean las necesidades se pueden escoger estas técnicas. Sin embargo, ya que en este caso no se tiene un tarea en específico, nos centraremos en poder realizar un análisis visual del comportamiento de los datos, esperando tener una visión más clara del comportamiento de los datos.

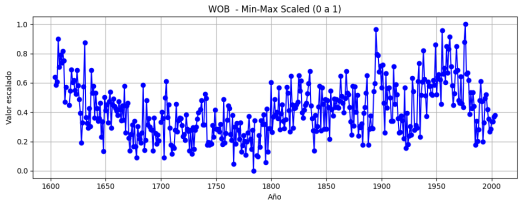
Entre los datos que podemos escalar se encuentran las variables geográficas, tales como altura sobre el nivel del mar, latitud y longitud; y los datos propios de los isótopos de carbono. Para las variables geográficas anteriormente mencionadas el reescalamiento por Z-score es el más adecuado, puesto que hacer una transformación tipo Min-Máx implícitamente podría implicar que tiene sentido hablar de porcentajes o razones entre las ubicaciones geográficas, lo cual no es cierto. Por otro lado, el método Z-score permite identificar qué tan alejados se encuentran los datos del "punto central", así que su interpretación geográfica es más clara. Ambos métodos se encuentran programados en el código adjunto.

Por otro lado, una primera técnica de escalamiento para los isótopos es el método Min-Máx, el cual reescala a un intervalo $[0, 1]$ y es sensible a outliers. Este es útil cuando se quieren encontrar

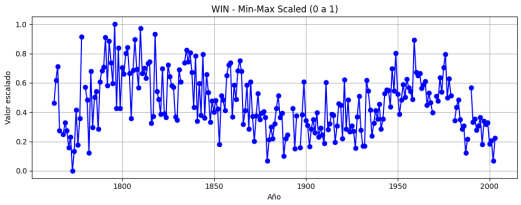
razones entre los datos, pero funciona mejor si no hay valores tan extremos, y permite ver qué tan "cercanos" son entre sí. En la Figura (7) se muestran los resultados para las localizaciones REN, WIN y WOB.



(a) REN Min-Max



(b) WOB Min-Max



(c) WIN Min-Max

Figura 7: Comparación de las columnas escaladas con Min-Max.

En este caso se realizó el escalamiento con los datos de las columnas REN, WIN y WOB, en donde se puede apreciar que el comportamiento de los datos no muestra grandes alteraciones a esta estandarización. Por otro lado, para el caso de la columna NIE2, en la cual ya había sospecha de outliers se tiene lo que muestra la Figura (8).

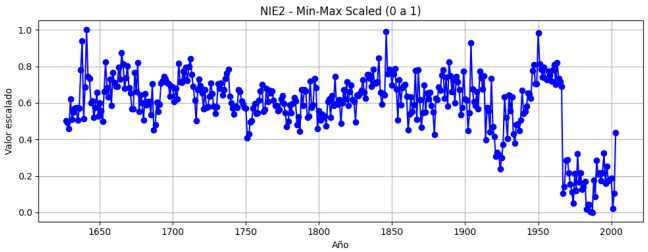
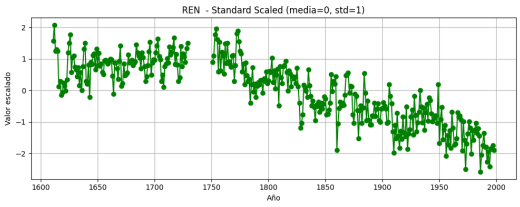


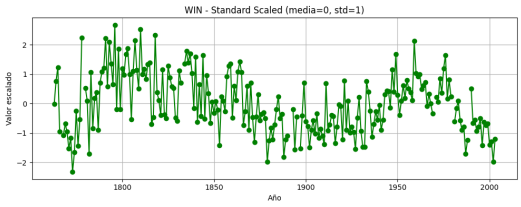
Figura 8: NIE2 Min-Max

En este caso, se puede notar que hay dos "grupos de datos" con tendencias o comportamientos distintos y que se pueden distinguir fácilmente por inspección visual. Esto alienta a pensar que el segundo grupo son posibles outliers, o que pueden provenir de algún mecanismo diferente, y como veremos más adelante, la identificación de outliers con IQR para dicha región también sugiere esta división de datos. Por consiguiente, dependiendo del objetivo del estudio, puede ser útil hacer un análisis de clusters en este caso.

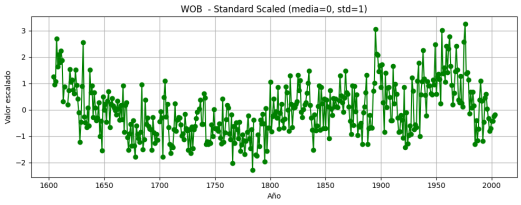
Otra forma de realizar escalamiento en los datos es mediante el método Z-score, en el cual a cada dato se le resta la media muestra y se escala por la varianza muestral. Este método es útil cuando se utilizan algoritmos que dependen de las distancias, como PCA o K-means. Del mismo modo se aplicó esta técnica de escalamiento a las mismas columnas que en el escalamiento anterior, de esta manera se tienen los resultados de la Figura (9).



(a) REN Standard



(b) WIN Standard



(c) WOB Standard

Figura 9: Comparación de las columnas escaladas con Standard Scaling.

Observemos que el comportamiento con esta estandarización tampoco se ve muy alterado, algunos de los cambios que se pueden observar es la escala, sin embargo no se ve resaltado algún conjunto de puntos ni algún patrón en específico. Por otro lado, observemos en la Figura (10) que, para la columna NIE2, el patrón no cambia mucho en comparación de el método de escalamiento Min-Max.

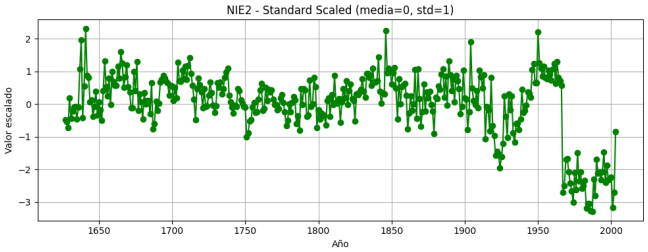


Figura 10: NIE2 Standard

Al igual que en los casos anteriores, la diferencia más evidente es el cambio de escala entre ambos métodos de estandarización, puesto que los datos no se encuentran tan dispersos, así que la elección óptima del método de escalamiento debería depender de los objetivos que se quieran lograr con los análisis.

5. Visualización exploratoria

Puesto que los datos contienen mediciones de especies repetidas pero en distintos lugares, una de las primeras visualizaciones exploratorias que se pueden hacer consiste en un diagrama de dispersión para cada especie, que nos permita observar si se comportan de manera diferente dependiendo del lugar en donde se encuentran. En la Figura (11) se presentan ejemplos de dichos diagramas combinados para las especies *Quercus robur* y *Pinus nigra*.

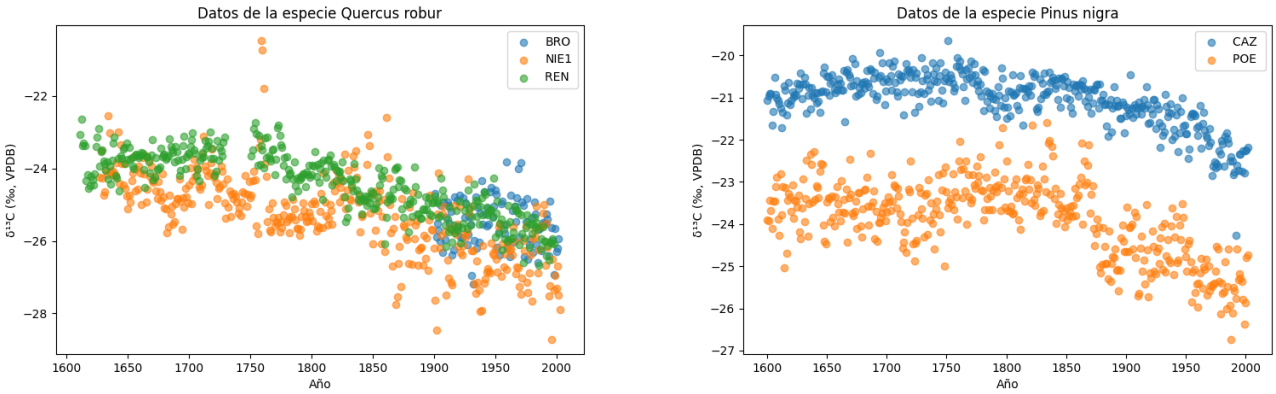


Figura 11: Diagramas de dispersión de las especies *Quercus robur* y *Pinus nigra*.

Este tipo de visualizaciones permite descubrir si las mismas especies se comportan o no de manera diferente dependiendo de su localización; y como se puede observar, mientras que para la especie *Quercus robur* es difícil hacer una clasificación o separación de los datos, puesto que todos parecen comportarse de la misma manera sin importar la localización, para la especie *Pinus nigra* es más clara la separación de los datos provenientes de las distintas zonas. Por lo tanto, dependiendo del propósito del estudio que se quiera realizar, en general podría ser un poco más conveniente hacer un análisis por zonas y no por especies, y posteriormente conjuntar las conclusiones para verificar si la ubicación era relevante para el estudio.

Con lo anterior en mente, procederemos a realizar diagramas dependientes de la zona de recolección de los datos. Por ejemplo, para diagnosticar la distribución de la razón de isótopos de carbono δ^{13}C (en partes por milésimas) con respecto al estándar VPDB (Vienna Pee Dee Belemnite), podemos hacer histogramas de las distintas zonas como se muestra en la Figura (12).

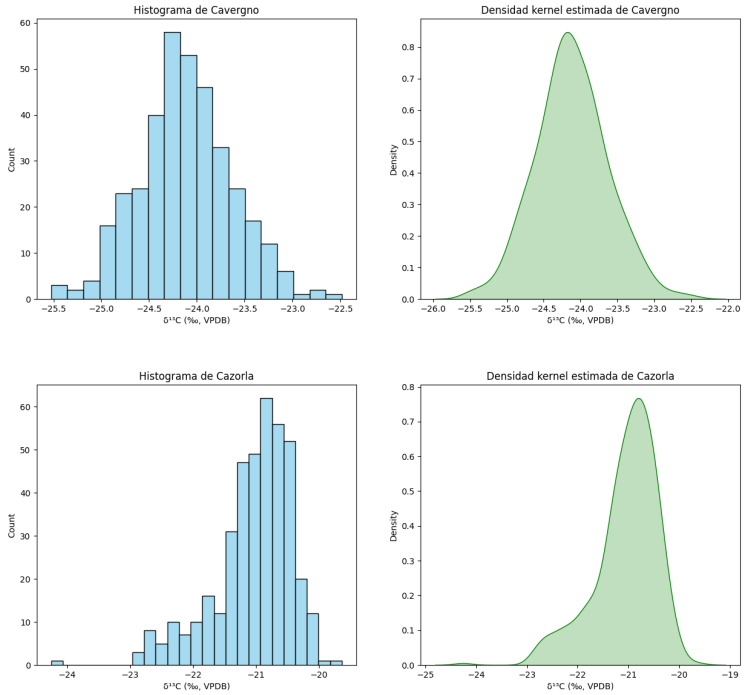


Figura 12: Histogramas de δ^{13}C‰VPDB en Cavergno y Cazorla.

Este tipo de visualizaciones ayuda a comprender el comportamiento de la distribución de dicha

variable aleatoria. Por ejemplo, para Cavergno se observa una distribución aproximadamente simétrica (si acaso con colas un poco más pesadas a la derecha), pero que podría sugerir un comportamiento aproximado a una normal. Sabiendo esto, se puede realizar una prueba de Shapiro Wilks para comprobar dicha normalidad, obteniendo que no se puede rechazar la hipótesis nula de normalidad.

Por otra parte, al observar el histograma de la zona Cazorla se ve claramente un comportamiento asimétrico con colas pesadas a la izquierda; y, en efecto, la prueba de Shapiro Wilks rechaza la hipótesis de normalidad con un p-valor de 2.94×10^{-13}. Otra ilustración útil puede ser con el histograma y encima su densidad estimada, como en la Figura (13), que facilita la visualización de ambos componentes.

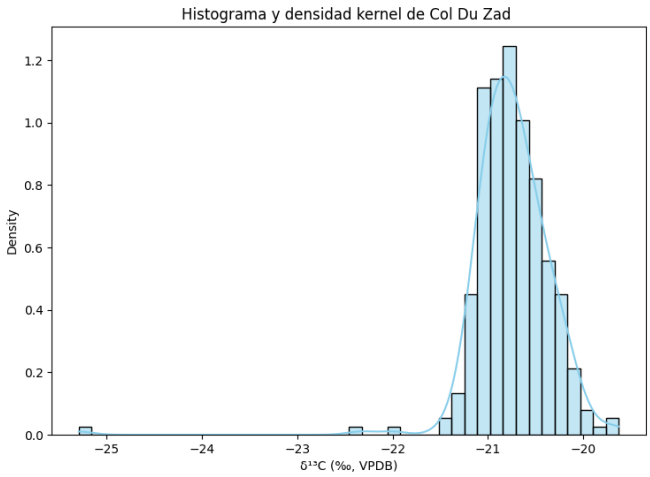


Figura 13: Histograma y densidad kernel estimada de δ^{13}C‰VPDB en Col Du Zad.

Por otro lado, con la imputación de datos realizada, por ejemplo usando la media, podemos ver la relación entre el año y los isótopos de carbono en una gráfica como la que presenta la Figura (14). En dicha figura se grafican los datos imputados y se hace una regresión lineal, cuyo ajuste se muestra con la recta roja.

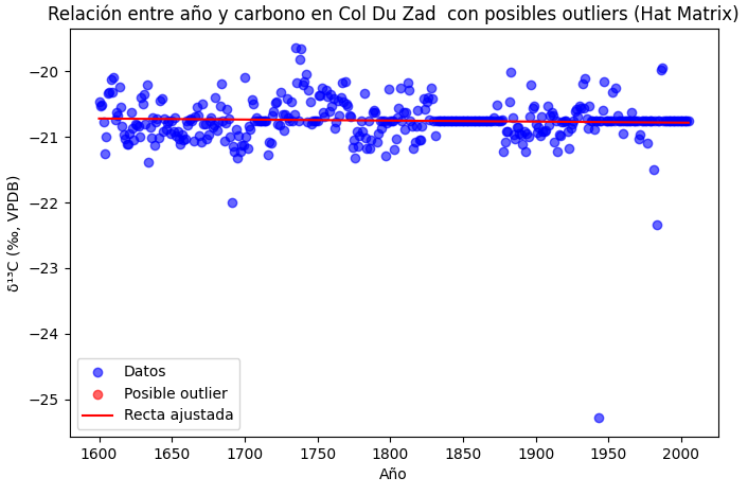


Figura 14: Relación entre δ^{13}C‰VPDB y el año en Col Du Zad.

En el caso de Col Du Zad, la regresión lineal parece explicar de manera adecuada a los datos,

y se observa que los datos imputados caen precisamente encima de la recta ajustada, lo que puede sugerir que este tipo de imputación es adecuada, puesto que conserva la tendencia de los datos originales.

Por otra parte, aunque a simple vista podríamos sospechar de la presencia de outliers, la Hat Matrix en este caso no los distingue. Por consiguiente, podrían ser necesarios otros métodos que ayuden a evaluar si hay datos extremos o outliers. Ya que en este caso la prueba de Shapiro-Wilks también descarta la hipótesis de normalidad para los datos (no imputados), la prueba Z-score no es necesariamente la mejor para detectarlos, puesto que no se satisface la hipótesis de normalidad, y esto sugiere que se pueda hacer una identificación de outliers por medio del Rango Inter-cuartílico. La Figura (15) muestra ambos resultados para su comparación.

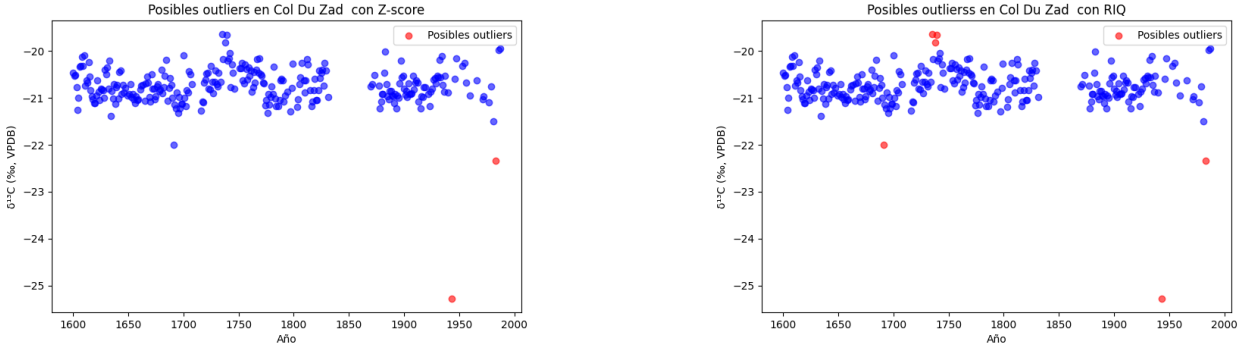


Figura 15: Detección de outliers en Col Du Zad con dos métodos.

Para Col Du Zad, podemos observar que el método Z-score proporciona buenos resultados, con la identificación de dos posibles outliers, que visualmente tienen sentido. Por otra parte, el Rango Inter-cuartílico (RIQ) es un poco más sensible, identificando unos tres datos más, y por la robustez de dicho estimador, se tienen más fundamentos teóricos para sospechar de que dichos datos son, en efecto, outliers. Por último, un comportamiento interesante se puede notar al hacer la detección de outliers en Niepolomice (NIE2), que se muestra en la Figura (16), en donde el método Z-score no detecta ninguno, pero el IQR detecta una gran cantidad, los cuales se encuentran agrupados en lo que parecen clusters. Dependiendo del objetivo en mente de la investigación, eso sugiere que podría ser sensato un análisis más sofisticado usando métodos de clasificación, entre otros.

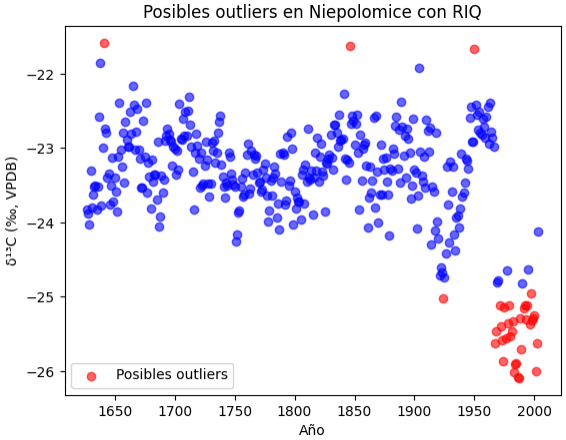


Figura 16: Detección de outliers en Niepolomice (NIE2) con IQR.

6. Reflexión crítica

Los datos analizados en este proyecto muestran, en general, pocas inconsistencias o estructuras codificadas de manera incorrecta. Sin embargo, el preprocesamiento siempre es importante para hacer un correcto análisis estadístico. Con las técnicas implementadas, este tipo de datos pueden ayudar a modelar el comportamiento climático en Europa a lo largo de los años, en una escala de tiempo de más de 400 años, e incluso pueden servir para probar hipótesis relacionadas con el cambio climático, analizar si la industrialización tuvo un fuerte impacto en el clima de la época, o modelar periodos de estrés de los árboles y el comportamiento de las especies analizadas.

Las decisiones que se toman en el preprocesamiento de los datos son muy importantes, puesto que, si no se hacen con cuidado, pueden introducir sesgos o tendencias a los datos que no se tenían originalmente, afectando la validez de los modelos, y posiblemente su interpretabilidad y robustez.

La etapa de limpieza de una base de datos es fundamental antes de empezar un análisis estadístico, puesto que los errores de registro podrían influir significativamente en el análisis. Por ejemplo, en nuestro análisis de la base de datos trabajada identificamos que había 7 especies distintas de árboles registradas, y al hacer una exploración visual fue fácil comprobar que estas eran, en efecto, especies distintas, y no se habían detectado como diferentes por errores tipográficos. Esto es importante de ver, pues de otro modo podríamos hacer un análisis pensando que los datos corresponden a fenómenos diferentes, cuando en realidad no es así, afectando nuestras interpretaciones.

Por otra parte, como parte de la limpieza fue necesario reescribir los datos faltantes "NA " (que aparecían con espacio) a tipo NaN, puesto que el programa lo reconocía como una cadena y no como un NaN. Esta clase de limpieza es necesaria puesto que de otro modo pueden surgir problemas en la manipulación de datos, o se pueden identificar como categóricos cuando en realidad no lo son.

Por otro lado, si se detectan muchos datos faltantes, es probable que las inferencias realizadas no sean tan robustas, por lo que vale la pena considerar la imputación. Sin embargo, esta es una parte delicada, puesto que si no se hace con cuidado puede introducir sesgo o modificar e incluso eliminar las tendencias que podrían presentar los datos. Por consiguiente, antes de realizar cualquier tipo de imputación, es necesario analizar el mecanismo que pudo haber ocasionado la presencia de los datos faltantes, para ver si tiene sentido, teóricamente, imputar datos, y cuáles serían las mejores opciones.

En la base de datos trabajada pudimos observar que la interpolación lineal no siempre daba los mejores resultados, puesto que introducía "tendencias" en los datos que no existían previamente. Sin embargo, pudimos resolver este problema para aquellas localizaciones en las que los datos seguían una distribución normal aproximada, al imputar con datos muestreados de dicha distribución. Esto permite conservar la estructura general de los datos, y mantener la generalidad al hacer análisis estadísticos posteriores, aunque la decisión final debería depender del objetivo de la investigación en curso.

La codificación también es una etapa que puede influir en los análisis estadísticos posteriores con la que se debe tener cuidado. Cuando transformamos variables categóricas a numéricas, en general no se recomienda aplicar *label encoding* si es que las categorías no tenían un orden intrínseco, puesto que en el análisis posterior, los modelos podrían pensar que la categoría 1 es menor a la categoría 2, por ejemplo, cuando los datos originales no tenían esa relación de orden. Esto sucede por ejemplo al querer codificar los países o las especies de árboles, en donde no existe un orden entre las categorías, y por consiguiente es mejor aplicar otro tipo de codificación, como *one-hot*.

Por último, ciertos modelos estadísticos son muy sensibles a las magnitudes de los valores, como

PCA, k-means, entre otros, en donde las distancias entre los datos es importante. Por consiguiente, si hay datos medidos en distintas unidades, el escalamiento es la opción ideal, puesto que esto evita que ciertos datos tengan mayor influencia por el simple hecho de ser más grandes, y viceversa. Por ejemplo, si en la base de datos trabajada se quieren analizar los factores geográficos relacionados con la presencia de isótopos, la latitud y longitud pueden resultar muy pequeños en comparación con la altura sobre el nivel del mar, lo que puede provocar que resulten insignificantes, cuando en realidad sí lo sean. Por consiguiente, es importante detectar estas situaciones antes de proceder con análisis estadísticos más complejos.

En conclusión, el preprocesamiento de datos es una tarea estadística muy importante para el análisis del fenómeno de interés, pero a su vez siempre debe estar sustentada por fundamentos estadísticos que permitan evitar la introducción de sesgos u otras situaciones que disminuyan la calidad de nuestro análisis.

7. Referencias

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