CIMAT

Análisis de datos: Stable carbon 'ISONET'

Introducción a Ciencia de Datos.

Autores: Luz María Salazar M.† Rodrigo Gonzaga S.† María Alejandra Borrego L.†
Profesor: Dr. Marco Antonio Aquino López†

Centro de Investigación en Matemáticas A. C.†

RESUMEN:

Este reporte presenta un análisis estadístico de la base de datos ISONET, que contiene mediciones isotópicas de $\delta^{13}C$ en anillos de árboles europeos. Se identificaron patrones de datos faltantes predominantemente del tipo MCAR, se detectaron *outliers* mediante métodos IQR y Z-score, y se compararon estrategias de manejo de datos: eliminación por período común e imputación temporal con un modelo de espacio de estados y suavizado de Kalman. Además, se argumenta que la elección de la estrategia depende del objetivo de investigación, con la imputación preservando mejor la información pero introduciendo supuestos del modelo.

Índice

1. Introducción

El análisis de los patrones de anillos de los árboles, o dendrocronología, es una ciencia muy exacta y una importante técnica de datación. La información contenida en los anillos anuales de los árboles es un recurso valioso para el estudio del cambio ambiental. El clima pasado puede reconstruirse a partir de los cambios interanuales en el ancho y la densidad de los anillos anuales.

Los datos considerados fueron producidos dentro del proyecto ISONET (400 Years of Annual Reconstructions of European Climate Variability Using a Highly Resolved Isotopic Network), y comprenden los registros ISONET de δ13C.

Con el fin de iniciar una amplia red espaciotemporal de isótopos estables en anillos de árboles a lo largo de Europa, financiada como parte del Quinto Programa Marco de la CE "Energía, Medio Ambiente y Desarrollo Sostenible", se consideraron 24 cronologías europeas de isótopos estables con resolución anual a partir de celulosa de anillos de árbol para los últimos 400 años (1600 – 2003) en Carbono y Oxígeno, y para los últimos 100 años en Hidrógeno. El proyecto ISONET se ha esforzado por mejorar en gran medida nuestra comprensión de los sistemas climáticos europeos, proporcionando datos cuantitativos independientes para la verificación de modelos y la elaboración de políticas. La red de 24 sitios mencionada ofrece cobertura dendrocronológica desde Iberia hasta Fennoscandia, Caledonia y el Tirol. Este proyecto analiza las razones isotópicas estables (C, H, O) de estas series temporales resueltas anualmente, esto para reconstruir los regímenes climáticos pasados (temperatura, humedad relativa y características de la precipitación) de los últimos 400 años.

En la base de datos mencionada encontramos la clasificación de los datos categóricos por el sitio en el que fue tomada la medición. La primera fila corresponde al código del Sitio, el cual es un código de 3 letras correspondiente a cada sitio. Las 9 filas siguientes muestran la información de la siguiente manera: **Site name:** corresponde al bosque o ciudad más cercana donde se hizo la medición; **Country:** nombre del país donde se encuentra el sitio; **Latitude:** las coordenadas geográficas de los sitios, medida en grados decimales; **Longitude:** las coordenadas geográficas de los sitios, medida en grados decimales; **Species:** nombre científico de la especie de árbol al que se le realizó la medición; **First year CE:** primer año registrado del registro δ13C del sitio, (años d.C.); **Last Year CE:** primer año registrado del registro δ13C del siti, (años d.C.); **Elevation:** elevación promedio del sitio en metros sobre el nivel del mar; **Year CE:** año del anillo de crecimiento, expresado en el calendario común (años d.C.).

Es importante notar que sobre la fila de Year CE lo que aparece es la relación isotópica $^{13}C/^{12}C$, mientras que el año del anillo podemos apreciarlo sobre la primera columna, como un tipo index. Así, en cada columna (para cada sitio) se presentan 405 filas correspondientes a las mediciones. Sin embargo, encontramos una gran cantidad de datos faltantes en muchos de los sitios estudiados.

Clasificamos todas estas variables según su escala de medición y presentamos dicha clasificación en la siguiente tabla:

Variable	Escala de medición	Variable	Escala de medición
Site name	**Nominal**	Species	**Nominal**
Country	**Nominal**	First year CE	**Intervalo**
Latitude	**Intervalo**	Last year CE	**Intervalo**
Longitude	**Intervalo**	Elevation	**Racional**

A continuación presentamos la gráfica de dispersión de los datos sin hacer distinción del sitio en el que se encuentra

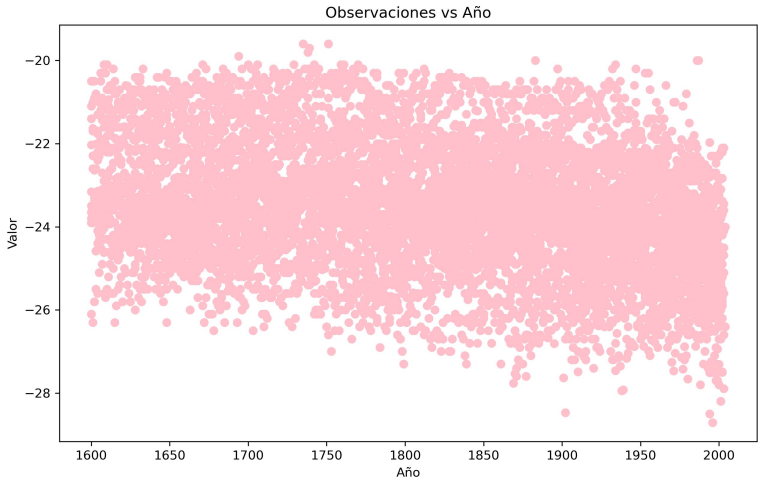


Figura 1: Gráfica de dispersión sin distinción de sitios.

Lo que podemos ver en esta gráfica es que los puntos se ven relativamente bien distribuidos, no encontramos grandes huecos a lo largo del tiempo y no se observa una tendencia clara, pues la nube parece bastante homogénea en todos los años observados.

La variación isotópica que vemos en esta gráfica es una muestra del rango general que nos permite ver la variabilidad temporal genérica pero no estamos observando la información local, es decir, las diferencias entre sitios. Para ello, presentamos la siguiente gráfica en donde solo mostramos la dispersión de las observaciones en los sitios con mayor cobertura de las mediciones, es decir, los sitios en los que se tienen más datos, y esta nos permite observar del rango de las mediciones isotópicas que se encuentran en cada sitio. En ella también podemos notar cómo cada sitio tiende a quedarse en un rango no muy grande de variación isotópica hasta el año 1900, aproximadamente, pero después muestran una caída en las mediciones que podría ser interesante estudiar.

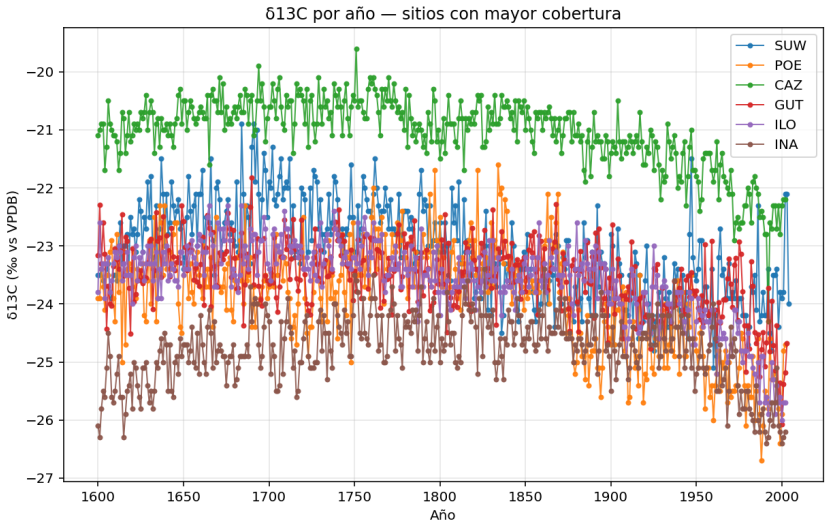


Figura 2: Gráfico de dispersión con identificación.

2. Detección de problemas en los datos

2.1. ¿Qué tipo de faltantes hay en la base de datos?

La Figura 3 muestra un mapa de calor con los años en el eje vertical y los sitios en el eje horizontal, donde el color claro representa valores faltantes. Se observan varios patrones relevantes, algunos sitios (`BRO`, `LAI`, `PAN`, `VIN`) carecen de información durante casi todos los años, reflejando que empezaron a medirse en épocas recientes al siglo XX.Hay un caso particular en `AHI` en el que se tiene mucha información pasada, pero dejó de medirse en epocas cercanas al siglo XX. Otros sitios, como `CAZ` o `ILO`, presentan series muy largas desde 1600 hasta 2000, con relativamente pocos huecos intermedios. Sitios como `COL` tienen datos dispersos con huecos frecuentes a lo largo de la serie.

En general, el patrón de faltantes aparece en bloques por sitio y época, no alrededor de valores atípicos. Esto refuerza la idea de que se trata de un mecanismo de diseño y no de valores faltantes dependientes de δ^{13}C.

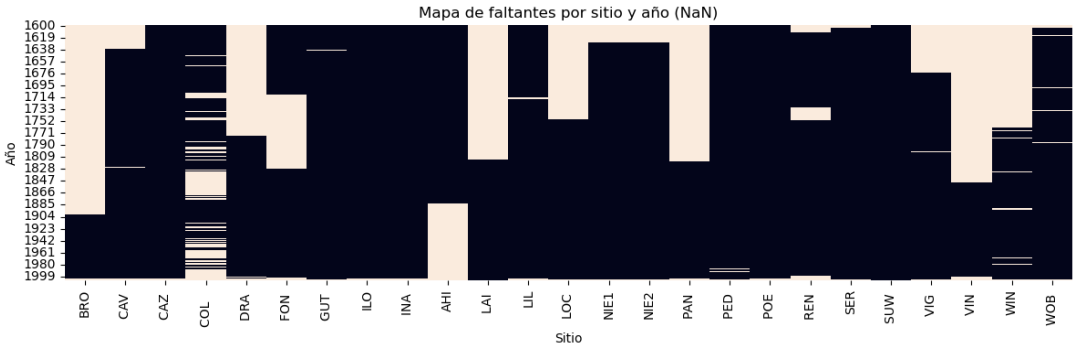


Figura 3: Mapa de datos faltantes por sitio y año en la base ISONET. Color claro = (`NaN`).

La figura 4, muestra con detalle las variables con valores faltantes, de mayor a menor.

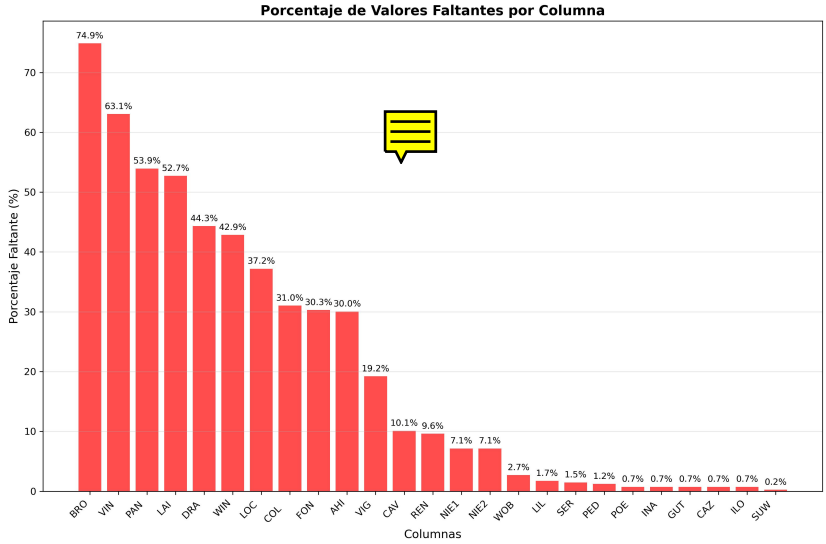


Figura 4: Histograma de valores faltantes.

Observe que los sitios con el 50 % de datos faltantes son: BRO, VIN, PAN, LAI. Entre (30-50 % faltantes) son: DRA, WIN, LOC, COL, FON, AHI. Entre 10-30 % faltantes son: VIG, CAV, REN. Entre 5-10 % son: NIE1, NIE2. Finalmente, los que tiene menos datos faltantes son: WOB, LIL, SER, PED, GUT, CAZ, INA, ILO, POE, SUW.

2.2. Detección de Outliers e inconsistencias.

De acuerdo con ([Wickham et al., 2017]), en el análisis estadístico, la detección de valores atípicos (outliers) es crucial para garantizar la calidad de los datos y la validez de los resultados. Dos métodos ampliamente utilizados son el **Z-score** y el **Rango Intercuartílico (IQR)**. Cada método tiene sus ventajas, limitaciones y aplicaciones específicas. El primero, permite comparar datos de diferentes distribuciones e indica cuán extremo es un valor en términos de desviaciones estándar. Por otra parte, el segundo no se ve afectado por valores extremos y no asume distribución normal de los datos.

En el caso de los datos de ISONET (δ^{13}C), el uso se debe a la naturaleza de los datos, pues los valores de δ^{13}C pueden tener distribuciones no normales. Los diferentes sitios pueden tener diferentes patrones de distribución. Pero hay que tener cuidado, algunos "outliers"pueden ser valores biológicamente válidos pero extremos. La siguiente tabla, muestra la cantidad de Outliers obtenidos, por cada método:

Sitio	**Outliers IQR**	**% IQR**	**Outliers Z-Score**	**% Z-Score**
BRO	0	0.0 %	0	0.0 %
CAV	9	2.5 %	1	0.3 %
CAZ	18	4.5 %	2	0.5 %
COL	5	1.8 %	2	0.7 %
DRA	3	1.3 %	2	0.9 %
FON	6	2.1 %	3	1.1 %
GUT	8	2.0 %	3	0.7 %
ILO	17	4.2 %	9	2.2 %
INA	11	2.7 %	0	0.0 %
AHI	14	4.9 %	4	1.4 %
LAI	3	1.6 %	1	0.5 %
LIL	4	1.0 %	2	0.5 %
LOC	3	1.2 %	2	0.8 %
NIE1	5	1.3 %	4	1.1 %
NIE2	34	9.0 %	7	1.9 %
PAN	7	3.7 %	1	0.5 %
PED	10	2.5 %	3	0.7 %
POE	1	0.2 %	1	0.2 %
SER	20	5.0 %	1	0.2 %
VIG	1	0.3 %	1	0.3 %
VIN	8	5.3 %	0	0.0 %
WOB	3	0.8 %	3	0.8 %
Total	**172**	**2.8 %**	**48**	**0.8 %**

El método IQR detecta **3.6 veces más outliers** que Z-score (172 vs 48). Eso se debe a que IQR es más sensible a la dispersión de los datos y el Z-score es más conservador y detecta solo valores extremos. Observe que el sitio **NIE2** tiene un alto porcentaje en ambos métodos (9.0 % IQR, 1.9 % Z-score). También, **ILO** tiene un alto porcentaje consistente (4.2 % IQR, 2.2 % Z-score).

Sobre las inconsistencias o codificación ambigua, por ejemplo que aparezcan las palabras New York, Newyork, newyork, NY; para referirse a lo mismo.

3. Manejo y manetenimiento de los datos

Clasificación del mecanismo.

MCAR (Missing Completely At Random): si interpretamos que los faltantes dependen exclusivamente de la decisión de cuándo iniciar o terminar la medición, entonces los datos ausentes son independientes de los valores reales de δ^{13}C. Esto permite considerar la eliminación de casos completos como insesgada (aunque menos eficiente).
MAR (Missing At Random): otra interpretación es que la ausencia depende de covariables observadas (por ejemplo, especie, localización, protocolo). Así, la probabilidad de tener datos sí varía según el sitio, pero no según el valor isotópico dentro de cada sitio.
MNAR (Missing Not At Random) no parece plausible aquí, pues no hay evidencia de que los faltantes se concentren en años con valores extremos.

3.1. Estrategias elegidas

Estrategia 1. Eliminación por período común

De acuerdo con ([Durbin and Koopman, 2012]), esta estrategia restringe el análisis únicamente al intervalo temporal donde todos los sitios seleccionados tienen datos disponibles. En nuestro caso, se fijó el inicio en 1900 y se excluyó el sitio `AHI`, ya que no aporta mediciones posteriores a esa fecha.

El objetivo de esta estrategia busca garantizar que todas las series sean comparables entre sí en las mismas fechas. La ventaja de esto es que asegura homogeneidad temporal y evita supuestos adicionales, la desventaja sin embargo, es que produce una reducción drástica del tamaño muestral.

Estrategia 2. Imputación temporal con modelo en espacio de estados + suavizado de Kalman

Siguiendo a ([Durbin and Koopman, 2012]), para cada sitio se ajusta un modelo estructural

$$y_t = \mu_t + \varepsilon_t, \qquad \mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t, \qquad \beta_t = \beta_{t-1} + \zeta_t, \qquad \varepsilon_t \sim \mathrm{AR}(1),$$

(tendencia local lineal + componente AR(1)). Con el filtro/suavizador de Kalman se obtienen medias suavizadas para los años faltantes dentro del soporte observado (no se extrapolan bloques largos).

Las ventaja de esta estrategia es que respeta la autocorrelación interanual y la suavidad de las cronologías anuales de δ^{13}C, evitando el aplanamiento excesivo de métodos simples (media/interpolación lineal). La desventaja es que introduce supuestos de modelo y en huecos muy largos puede subestimar la variabilidad.

3.2. Bajo MCAR la eliminación es insesgada pero menos eficiente

Bajo MCAR, la eliminación de casos completos produce un estimador insesgado, pero menos eficiente debido a la pérdida de tamaño muestral efectivo y, en consecuencia, mayor varianza y esto podemos verlo demostrado formalmente en el ejercicio 5 de la parte teórica de la tarea.

3.3. Comparación de resultados del manejo de datos faltantes

La Figura 5 muestra histogramas comparativos entre los valores originales (azul) y los resultados de aplicar dos estrategias: (i) la eliminación por período común (naranja, arriba) y (ii) la imputación temporal con un modelo en espacio de estados y filtro de Kalman (naranja, abajo).

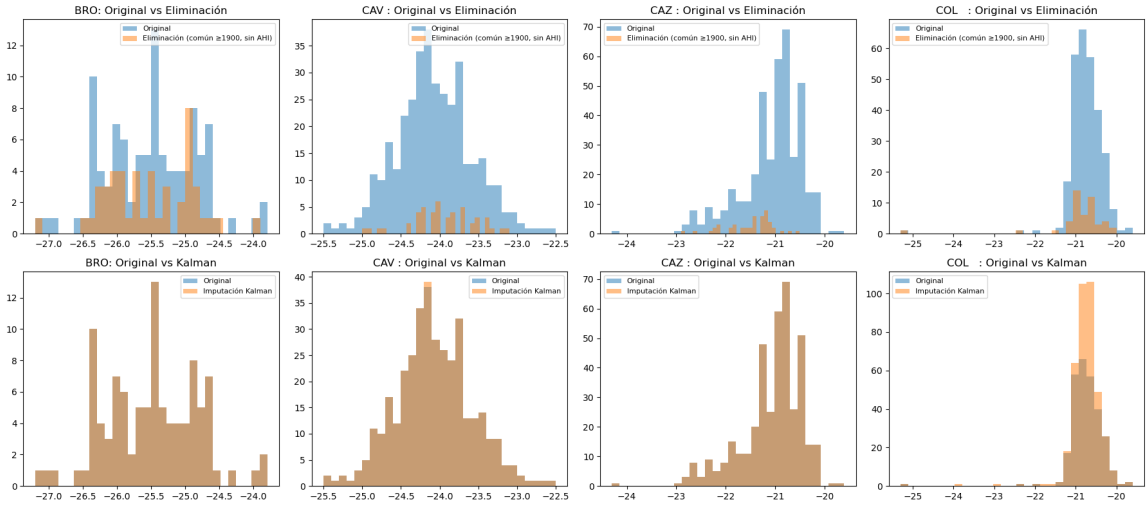


Figura 5: Comparación de histogramas: original (azul), eliminación por período común (naranja, arriba) e imputación Kalman (naranja, abajo).

Vemos que la estrategia de eliminación reduce de manera notable el número de observaciones: los histogramas naranjas queda mucho más pequeños que los azules, lo que refleja la pérdida de eficiencia. Sin embargo, la ubicación de los valores se mantiene en el mismo rango, por lo que no se introduce sesgo.

Con la imputación Kalman, el histograma naranja prácticamente se superpone con el azul, lo que indica que la distribución de la serie se conserva. En el caso de COL, Kalman rellena la mayoría de los huecos, de modo que el histograma naranja domina en tamaño. Si bien la forma de la distribución se mantiene, la cantidad de valores imputados puede generar una falsa sensación de mayor certidumbre de la que realmente existe.

Además de los histogramas, también es útil observar las series temporales imputadas en comparación con los valores originales. Las Figuras 6 y 7 muestran tres ejemplos representativos.

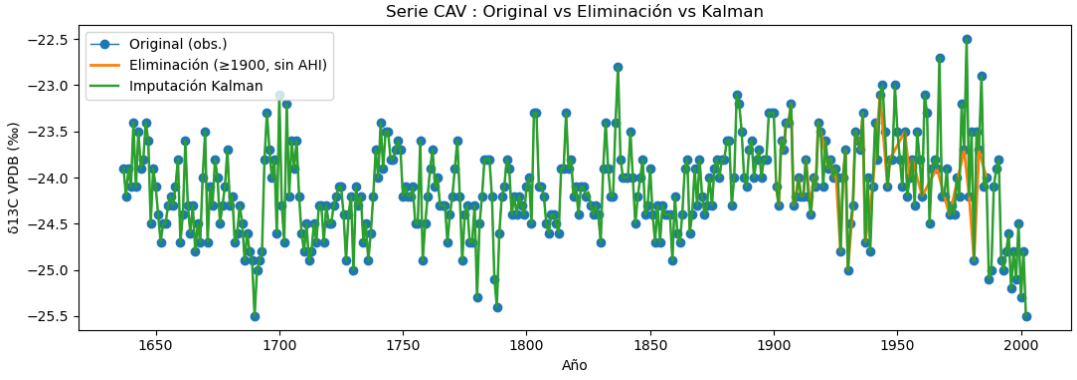


Figura 6: Serie CAV: original (azul), eliminación por período común (naranja) e imputación Kalman (verde).

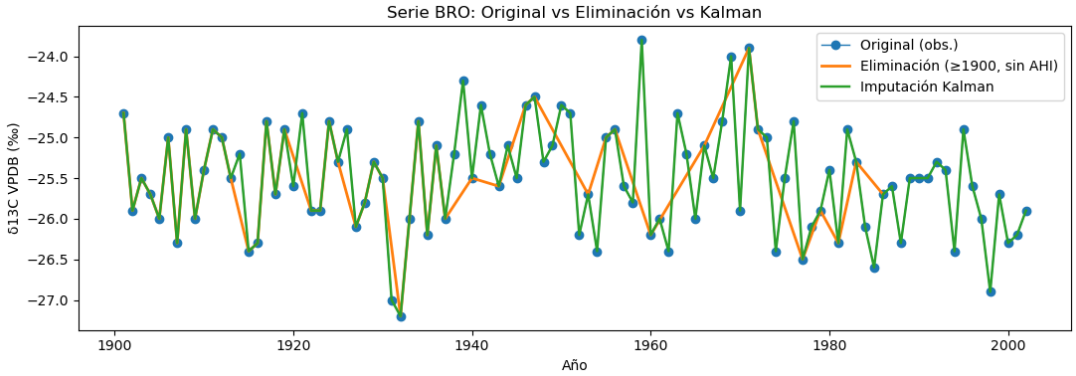


Figura 7: Serie BRO: original (azul), eliminación por período común (naranja) e imputación Kalman (verde).

En general, en series largas y completas, la eliminación descarta una proporción enorme de datos y reduce la capacidad de analizar la variabilidad histórica. La imputación Kalman, en cambio, conserva prácticamente toda la información.

En series con inicio tardío pero buena cobertura (BRO), ambas estrategias producen resultados similares en el período reciente, aunque Kalman permite aprovechar incluso los años con faltantes dispersos.

¿Qué estrategia es más apropiada?

Ambas estrategias responden a objetivos distintos y, por lo tanto, su conveniencia depende de la pregunta de investigación:

Eliminación por período común. Bajo el supuesto MCAR, esta estrategia es insesgada y garantiza homogeneidad temporal entre sitios. Es útil si el objetivo es comparar directamente varios sitios en el mismo intervalo de años, por ejemplo para estudiar variaciones regionales en el siglo XX. Sin embargo, en la base ISONET su costo en eficiencia es muy alto.

Imputación temporal (Kalman). Este método modela cada serie individualmente, respetando su estructura de tendencia y autocorrelación. Rellena los huecos dispersos de manera coherente y conserva prácticamente toda la información disponible. Su limitación es que introduce supuestos de modelo y, en sitios con demasiados faltantes, las imputaciones pueden dominar sobre los datos observados

4. Codificación y escalamiento

Si queremos trabajar con variables categóricas en la implementación de modelos o algoritmos necesaitamos transformarlas para que tenga sentido tomarlas en cuenta en algún modelo. Para ello, contamos con la codificación one-hot enconding, la cual transforma nuestra variable categórica en variables binarias.

En el caso de los datos analizados, si nos interesara estudiar las mediciones de isotópos por especie de árbol, considerando que estas sí tienen un efecto importante en los resultados, entonces la variable "Species" necesitaría una transformación que podemos hacer codificándola con el método de one-hot enconding y de esta manera, en el modelo con el que estemos trabajando podemos incluirla para capturar las diferencias entre especies sin imponerles un orden artificial, de manera que podemos interpretar cada categoría como una característica independiente que contribuye al comportamiento de la variable de interés.

Otra manera de analizar variables, en este caso numéricas, es reescalándolas para que sea posible compararlas. Dos tipos de escalamiento que aquí realizamos son el de normalización mín-máx y el de

estandarización Z-score.

Consideramos dos de los 24 sitios estudiados: Suwalki (SUW) y Woburn (WOB) en donde las mediciones se realizaron a la misma especie de árbol *pinus sylvestris*. Realizamos el escalamiento con normalización mín-máx y mostramos la gráfica box-plot para su comparación:

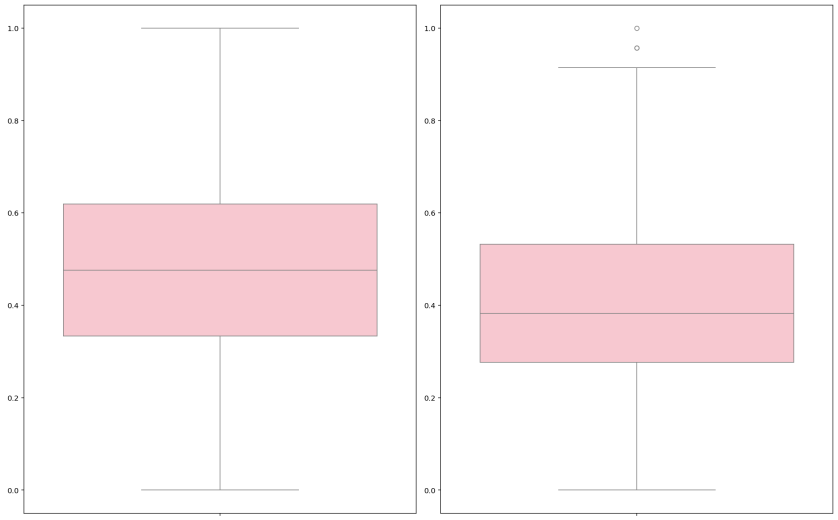


Figura 8: Box-plot comparativo con normalización mín-máx.

Podemos observar que en ambos sitios se presentan distribuciones relativamente simétricas, (un poco menos para WOB) y que en Suwalki (izquierda) la mediana está un poco más alta que en Woburn, lo cual sugiere que, en general, las mediciones en Suwalki tienden a ser mayores. También en Woburn se muestran un par de outliers altos, lo que indica que aunque la mayoría de valores son más bajos y estables, existen casos puntuales con mediciones mucho más altas. Podemos decir que, aunque se esté midiendo la misma especie de árbol, los resultados son algo diferentes en los dos sitios, y esto sugiere que el ambiente del sitio, ya sea el suelo, clima, altitud, etc., influye en las mediciones isotópicas aun dentro de la misma especie de árbol.

Realizamos además el escalamiento con estandarización Z-score y mostramos la gráfica de histograma para su comparación:

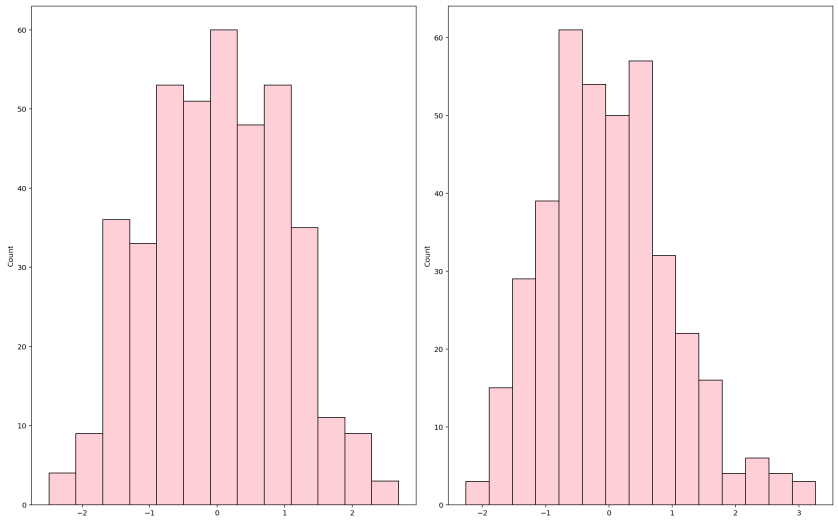


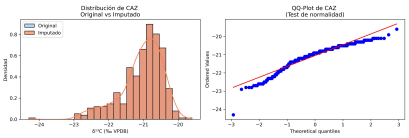
Figura 9: Box-plot comparativo con normalización mín-máx.

En estas gráficas podemos observar que histogramas se ven relativamente simétricos y con forma cercana a una campana, lo cual indica que la transformación Z-score funcionó adecuadamente y los

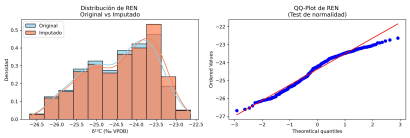
valores quedaron centrados en torno a 0. También que Suwalki presenta una dispersión más concentrada alrededor de la media, mientras que el Woburn muestra una ligera asimetría hacia valores altos.

5. Visualización para datos imputados

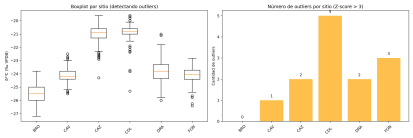
Por espacio del reporte, solo se realizó la visualización con dos variables. Se realizó una prueba de normalidad para los sitios CAZ y REN, en las figuras 10a y 10b puede observar que REN se ajustó bien a una distribución normal. La figura 10c muestra 6 gráficas de caja y bigote para la detección de Outliers, la gráfica muestra la cantidad de Outliers de las primeras 6 variables. Note que BRO tenía demasiados datos faltantes y ahora ya no.



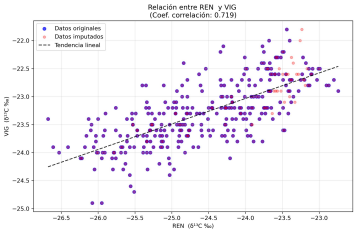
(a) QQ-Plot de CAZ



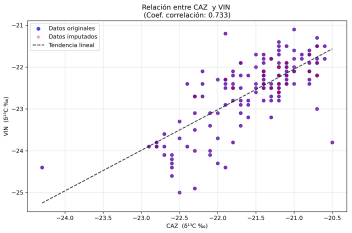
(b) QQ-Plot de REN



(c) Outliers



(d) Relación REN-VIG



(e) Relación CAZ-VIN

Figura 10: Análisis exploratorio de datos ISONET con datos imputados.

Finalmente, las figuras 10d y 10e muestran la relación de REN y CAZ, con los sitios que mayor correlación tuvieron, es decir con VIG y VIN, respectivamente. Claramente, REN-VIG parece tener una relación lineal.

6. Conclusiones

Las decisiones tomadas en la etapa de preparación de los datos tienen un efecto directo sobre la calidad de los modelos estadísticos posteriores. Por ejemplo, en sitios como `NIE2` se identificó un alto porcentaje de outliers; mantenerlos sin justificación puede sesgar parámetros de un modelo, mientras que eliminarlos sin criterio puede ocultar señales climáticas reales.

En el caso de la estrategia elegida para tratar datos faltantes, la eliminación por período común, al restringir a ≥ 1900 sólo sobrevivió un 11.3 % de las observaciones, lo que reduce poder estadístico y aumenta la varianza. En cambio, con imputación Kalman se conserva la continuidad de series largas, permitiendo que un modelo de predicción tenga más información para ajustar dinámicas temporales. Sin embargo, la imputación en sitios con muchos huecos puede dar una falsa sensación de certidumbre.

El uso de variables como la especie de árbol o el país del sitio requiere transformarlas para que los algoritmos las interpreten. La codificación one-hot permite incluir estas categorías, además el reescalamiento facilita la comparación entre sitios y evita que las magnitudes dominen sobre otras en algoritmos sensibles a la escala. En el ejemplo de los sitios `SUW` y `WOB`, la normalización min–max permitió comparar distribuciones en el mismo rango [0,1], mientras que Z-score centró las mediciones en torno a 0 con varianza unitaria.

La base ISONET, abre la puerta a diversas líneas de análisis: reconstrucción climática de largo plazo, comparación entre regiones y especies, validación de modelos climáticos, etc.

Referencias

[Durbin and Koopman, 2012] Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford university press.

[Schleser et al., 2023] Schleser, G. H., Andreu-Hayles, L., Bednarz, Z., Berninger, F., Boettger, T., Dorado-Liñán, I., Esper, J., Grabner, M., Gutiérrez, E., Helle, G., et al. (2023). Stable carbon isotope ratios of tree-ring cellulose from the site network of the eu-project 'isonet'.

[Wickham et al., 2017] Wickham, H., Grolemund, G., et al. (2017). *R for data science*, volume 2. O'Reilly Sebastopol.