Introducción a la Ciencia de Datos
Tarea I

Antonio Barragán
Omar García Ramos
Hazel Barragan

Reporte

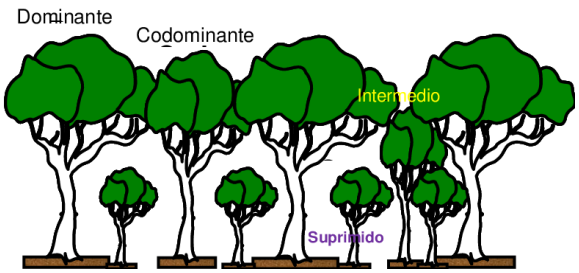
1 Exploración inicial de los datos.

Los datos analizados provienen del **proyecto ISONET** (*400 Years of Annual Reconstructions of European Climate Variability Using a Highly Resolved Isotopic Network*), cuyos objetivos específicos incluyen construir series temporales en anillos de árboles para reconstruir la variabilidad climática europea de los últimos 400 años, como se indica:

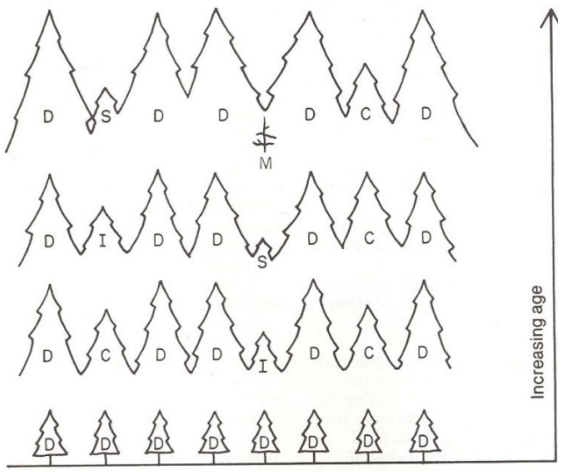
> "Data was produced within the ISONET project (400 Years of Annual Reconstructions of European Climate Variability Using a Highly Resolved Isotopic Network), to initiate an extensive spatiotemporal tree-ring stable isotope network across Europe. 24 European annually resolved stable isotope chronologies have been constructed from tree ring cellulose for the last 400 years (1600CE - 2003CE) for carbon and oxygen and for the last 100 years for hydrogen."

La selección de árboles codominantes (ver Figura 1) para el estudio representa una estrategia para minimizar sesgos en los datos. Los árboles codominantes (aquellos que ocupan una posición intermedia) presentan varias ventajas para el estudio:

- Al no estar expuestos a condiciones ambientales estresantes como los árboles emergentes, ni demasiado sombreados como los árboles suprimidos.
- Su crecimiento refleja las condiciones ambientales promedio del sitio de estudio.
- Presentan una menor influencia de factores competitivos extremos.



(a) Clasificación de árboles según sus copas



(b) Competencia de árboles en su crecimiento

Figure 1: Clasificación y competencia de árboles (Fuente: Emmingham y Elwood 1993)

Las variables estudiadas se presentan en la siguiente tabla:

Variable	Significado científico
Year CE	Referencia temporal absoluta para construcción de series temporales y fechado de eventos climáticos.
13CVPDB	Ratio isotópico $^{13}C/^{12}C$ que funciona como proxy de condiciones ambientales, especialmente estrés hídrico.
Species	Determina la especificidad de la respuesta fisiológica a condiciones ambientales.
Latitude/Longitude	Coordenadas geográficas para análisis de gradientes espaciales y patrones regionales.
Elevation	Factor altitudinal que influye en las condiciones microclimáticas.
First/Last year CE	Define el rango temporal de cada cronología para análisis de tendencias.

Table 1: Significado científico de las variables medidas en el proyecto ISONET

La variable $\delta^{13}C$ registra de manera integradora las condiciones ambientales que experimentó el árbol durante cada temporada de crecimiento. Valores más positivos de $\delta^{13}C$ generalmente indican mayor estrés hídrico y condiciones más secas, mientras que valores más negativos sugieren condiciones más húmedas y menor estrés. Los valores de $\delta^{13}C$ en celulosa de anillos de árboles constituyen una aproximación a las condiciones ambientales de la época.

2 Detección de problemas con los datos.

Es razonable asumir un mecanismo de datos faltantes **MAR** en este contexto debido a la naturaleza del diseño experimental y a las características biológicas de los árboles estudiados. En primer lugar, el rango temporal de cada cronología depende del año de inicio y del año final de los anillos disponibles en cada espécimen. Así, la ausencia de mediciones en ciertos periodos no está asociada directamente al valor de los isótopos en sí, sino a la edad del árbol o a la disponibilidad física del material. Por ejemplo, un árbol más joven no puede contener información de siglos anteriores, lo que genera faltantes determinados por la variable observada *First year*. Por otro lado el contenido isotópico (δ^{13}) depende de procesos fisiológicos de cada especie y de condiciones ambientales locales. Sin embargo, la probabilidad de que un valor esté ausente no depende del valor isotópico no observado, sino de factores externos como la dificultad de obtener muestras de ciertas especies o de árboles con problemas de preservación en los anillos más antiguos.

Table 2: Análisis de datos faltantes considerando rangos temporales. RTV = Rango Temporal Válido. FRTV= Faltantes en el Rango Total Válido, FT= Faltantes Totales.

Serie	First Year	Last Year	Total RTV	FRTV	% Faltantes RTV	FT
BRO	1901	2002	102	0	0.00	304
CAV	1637	2002	366	1	0.27	41
CAZ	1600	2002	403	0	0.00	3
COL	1600	2000	401	121	30.17	126
DRA	1776	1999	224	0	0.00	180
FON	1600	2000	401	118	29.43	123
GUT	1600	2003	404	1	0.25	3
ILO	1600	2002	403	0	0.00	3
INA	1600	2002	403	0	0.00	3
AHI	1600	1883	284	0	0.00	122
LAI	1812	2003	192	1	0.52	214
LIL	1600	2002	403	4	0.99	7
LOC	1749	2003	255	0	0.00	151
NIE1	1627	2003	377	0	0.00	29
NIE2	1627	2003	377	0	0.00	29
PAN	1816	2002	187	0	0.00	219
PED	1600	2003	404	3	0.74	5
POE	1600	2002	403	0	0.00	3
REN	1611	1998	388	21	5.41	39
SER	1604	2003	400	0	0.00	6
SUW	1600	2004	405	0	0.00	1
VIG	1675	2003	329	1	0.30	78
VIN	1850	1999	150	0	0.00	256
WIN	1763	2003	241	9	3.73	174
WOB	1604	2003	400	5	1.25	11

Para esta base de datos, la eliminación de casos no es apropiada ya que, sería altamente ineficiente: eliminaríamos cerca del 30% de los datos en series como COL o FON, aumentando sustancialmente la varianza de cualquier análisis posterior.

Las estrategias elegidas (imputación por regresión para series normales y por mediana para series no normales) son adecuadas bajo el mecanismo MAR. La **imputación por regresión** preserva las relaciones lineales entre la variable objetivo ($\delta^{13}C$) y el tiempo ($Year$), que es fundamental en el análisis de series temporales. Incorporar la incertidumbre de la predicción mediante la distribución t evita subestimar la varianza de los datos imputados. Para series no normales, la **imputación por mediana** es una estrategia robusta que no se ve afectada por outliers, proporcionando una estimación central estable para la variable en el año correspondiente. Ambas estrategias, aplicadas por serie, son más eficientes y teóricamente más sólidas para este proyecto que la simple eliminación de datos.

En cuanto a la detección de outliers, se implementaron dos enfoques basados en la distribución de los residuos de regresión lineal. Primero, mediante la prueba de Anderson-Darling se evaluó la normalidad de los residuos para cada serie temporal.

Cuando los residuos siguieron una distribución normal, se aplicó el método de **distancia de Cook con umbral $4/n$**, que identifica observaciones con influencia desproporcionada en el ajuste del modelo de regresión. La distancia de Cook para cada observación i se calcula como:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{p \cdot \text{MSE}}$$

donde $\hat{y}_{j(i)}$ es la predicción para la observación j cuando la observación i es excluida del modelo, p es el número de parámetros, y MSE es el error cuadrático medio. Valores de $D_i > 4/n$ se consideraron influyentes.

Para los casos donde los residuos no mostraron normalidad, se empleó un criterio no paramétrico basado en el **rango intercuartílico aplicado a los residuos** (Q1 − 1.5×IQR, Q3 + 1.5×IQR).

Este enfoque robusto detecta valores que se desvían significativamente de la tendencia temporal.

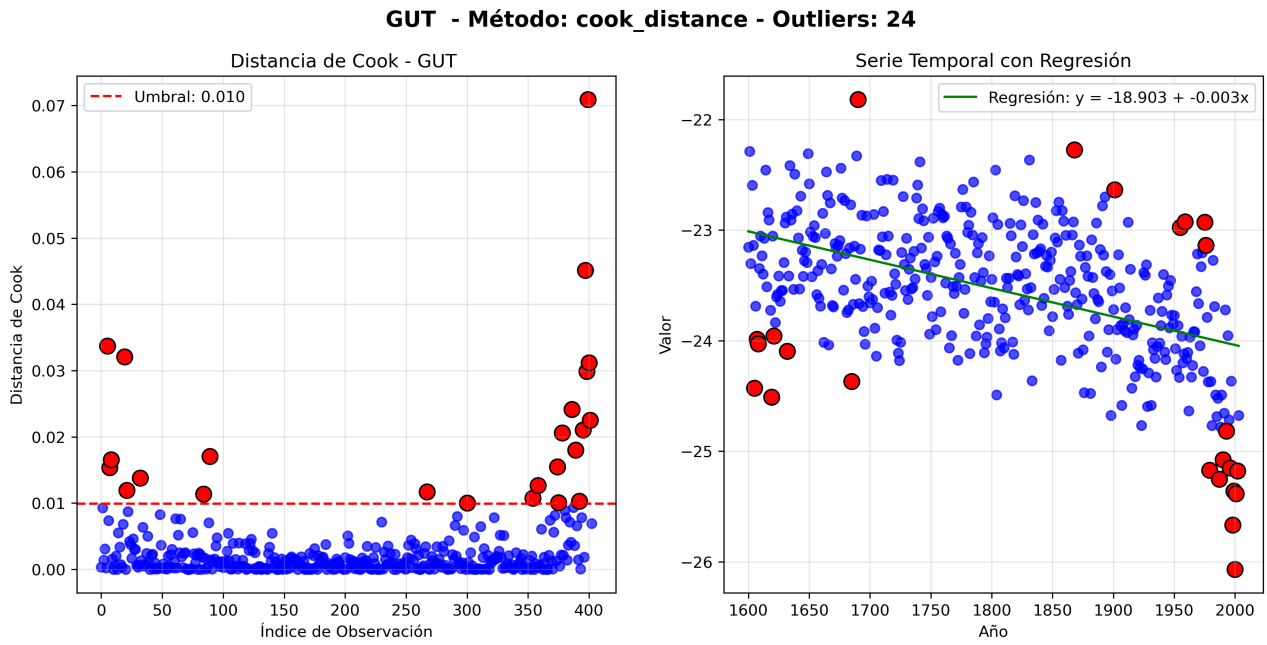


Figure 2: Obtención de outliers sobre el sitio Gutuli

Columna	N_datos	Método	Outliers_count
BRO	102	Cook	3
CAV	365	Cook	28
CAZ	403	IQR	5
COL	280	IQR	6
DRA	224	Cook	14
FON	283	IQR	9
GUT	403	Cook	24
ILO	403	IQR	5
INA	403	IQR	9
AHI	284	Cook	17
LAI	191	Cook	9
LIL	399	IQR	10
LOC	255	Cook	14
NIE1	377	IQR	11
NIE2	377	IQR	12
PAN	187	Cook	14
PED	401	Cook	25
POE	403	Cook	18
REN	367	Cook	24
SER	400	IQR	3
SUW	405	Cook	26
VIG	328	Cook	22
VIN	150	Cook	13
WIN	232	IQR	0
WOB	395	IQR	5

Table 3: Resumen de detección de outliers por método (Cook e IQR).

3 Imputación de datos

Se implementó un esquema de imputación diferenciado basado en la normalidad distribucional de cada serie temporal:

3.1 Series con Distribución Normal

Sea el modelo:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

con estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ y varianza residual estimada:

$$MS_{\text{Res}} = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Para una nueva observación en x_0, el valor predicho es:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

La varianza de la predicción tiene dos componentes:

$$\begin{aligned} \mathbb{V}(y_0 - \hat{y}_0) &= \mathbb{V}(\hat{y}_0) + \mathbb{V}(\varepsilon_0) \\ &= MS_{\text{Res}} \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) + MS_{\text{Res}} \end{aligned}$$

donde $S_{xx} = \sum (x_i - \bar{x})^2$.

El estadístico estandarizado sigue una distribución t:

$$\frac{y_0 - \hat{y}_0}{\sqrt{MS_{\text{Res}} \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$$

Mecanismo de Imputación

Para imputar y_0 incorporando incertidumbre:

$$y_0^{\text{imp}} = \hat{y}_0 + t^* \cdot \sqrt{MS_{\text{Res}} \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

donde t^* es una realización de t_{n-2}.

3.2 Series con Distribución No Normal

Para distribuciones no normales, los métodos paramétricos basados en regresión pierden eficacia, usaremos la Mediana. Como medida de dispersión:

$$\text{IQR} = Q_3 - Q_1$$

donde Q_1 y Q_3 son los percentiles 25 y 75 respectivamente.

$$\hat{y_{\text{imp}}} = \text{Mediana}(y_{\text{obs}}) + \epsilon$$

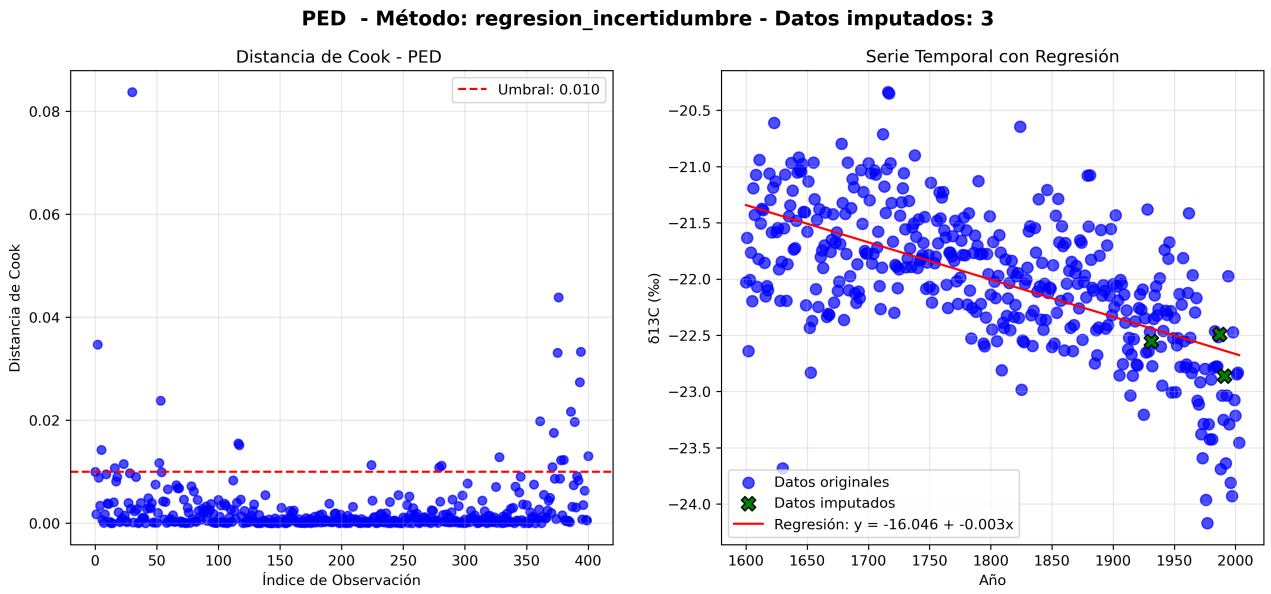


Figure 3: Imputación de datos sobre el sitio Pedraforca mediante el método de regresión lineal

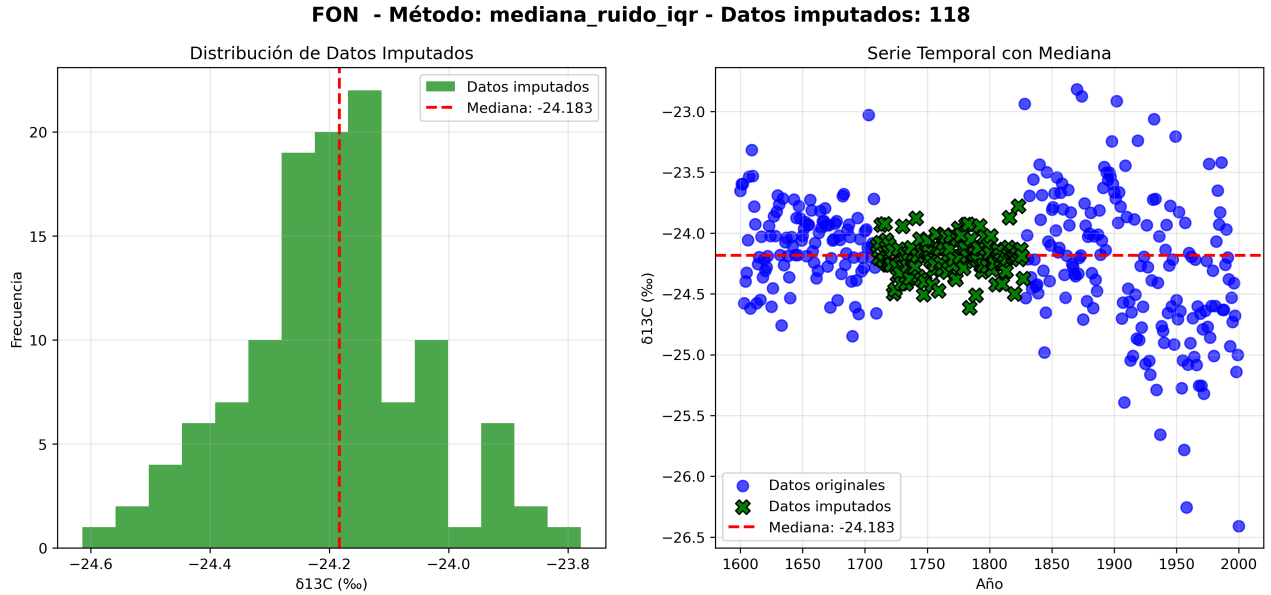


Figure 4: Imputación de datos sobre el sitio Fontainebleau mediante el método de IVQ

4 Codificación y Escalamiento

La variable categórica `Species` requiere transformación para ser utilizada en modelos. Si estuviéramos interesados en clasificar si algunas especies son más eficientes en el uso de recursos, y por ende crecen más por año, podríamos utilizar técnicas de codificación como one-hot encoding o label encoding, dependiendo del tipo de modelo a implementar.

5 Visualización Exploratoria

Considerando los temas vistos en clase, podemos trabajar la visualización mediante los siguientes recursos:

1. Graficar las especies de árboles para identificar tendencias.
2. Histogramas de distribución.
3. Estimación de densidad mediante kernels.

Sobre la parte 1, obtenemos los siguientes resultados:

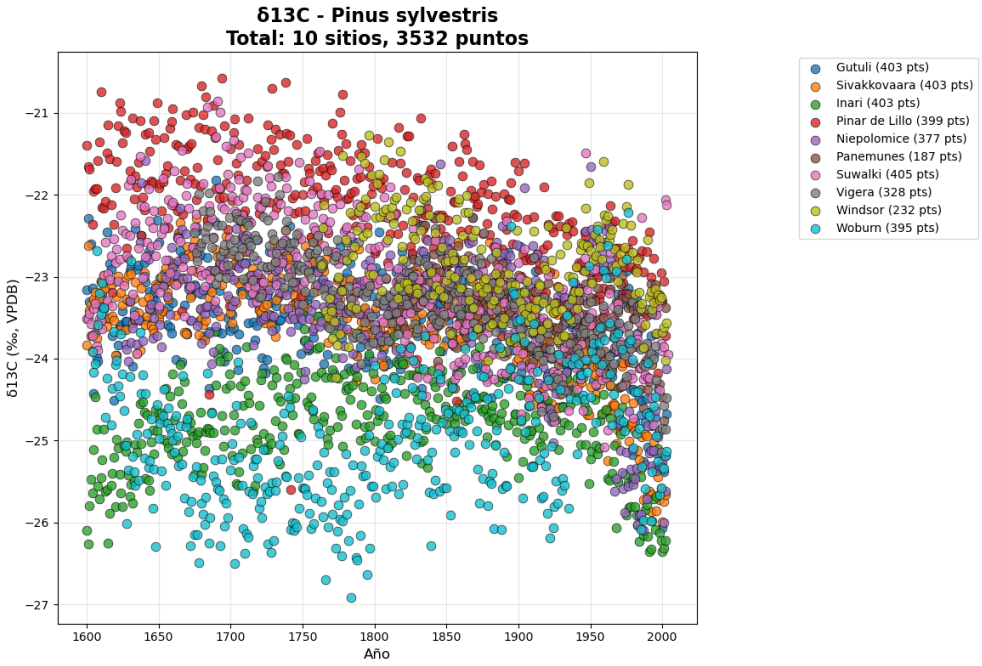


Figure 5: Especie de árbol *Pinus sylvestris* en todos los sitios de Europa donde fue recolectada

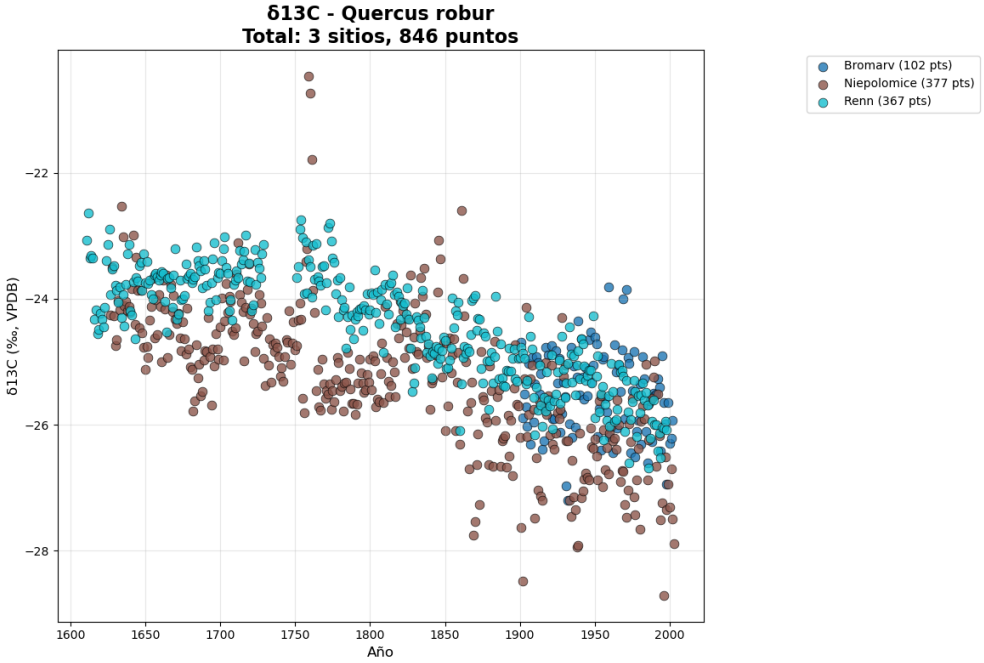


Figure 6: Especie de árbol *Quercus robur* en todos los sitios de Europa donde fue recolectada

Dado que la nube de puntos no presenta un patrón claro para los datos de la Figura 5, mientras que para la Figura 6 observamos que los lugares donde se encuentra el árbol presentan condiciones climatológicas similares (lo cual es distinto a la otra especie), se debe realizar un análisis considerando estas variables ambientales adicionales.

Para los puntos 2 y 3, al analizar el histograma de cada una de las especies en su respectivo lugar de origen:

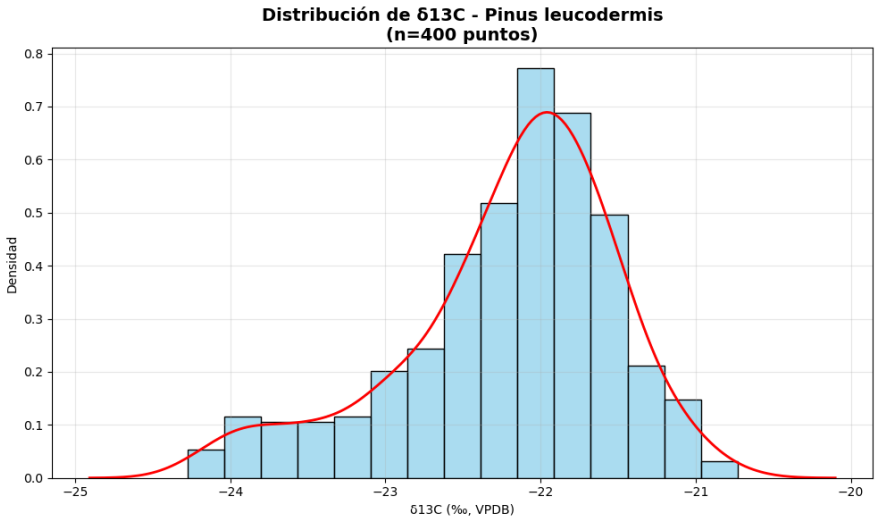


Figure 7: Estimación de densidad kernel aplicada a los datos de *Pinus leucodermis*

Para esta especie notamos una distribución asimétrica, por lo que podemos inferir que la especie no es muy efectiva en capturar recursos, pues sin importar el lugar de donde provenga, presenta un valor pequeño de δ^{13}C.

No obstante, en la especie *Cedrus atlantica*, tenemos que la distribución normal es razonable asumirla, por lo que podemos considerar:

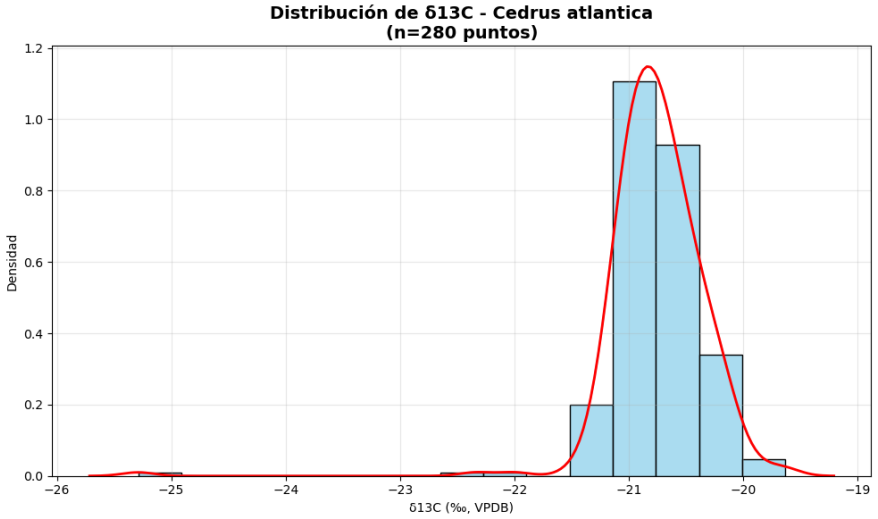


Figure 8: Estimación de densidad kernel para *Cedrus atlantica* mostrando distribución normal

La diferencia en los patrones distribucionales entre *Pinus leucodermis* (Figura 7) y *Cedrus atlantica* (Figura 8) sugiere diferencias fisiológicas significativas en la eficiencia del uso de recursos hídricos. Mientras que *Cedrus atlantica* muestra una distribución normal típica de especies adaptadas a condiciones ambientales estables, la distribución asimétrica de *Pinus leucodermis* indica una menor eficiencia en la captura de recursos, posiblemente relacionada con estrategias ecofisiológicas diferentes o limitaciones adaptativas. Este hallazgo justifica un análisis más profundo considerando variables ambientales específicas de cada sitio de muestreo.

6 Conclusiones

El proceso de preparación de la base de datos presentó desafíos iniciales significativos. La carga de los datos reveló inconsistencias en la codificación de valores faltantes, los cuales aparecían representados de múltiples formas, incluyendo 'NA', 'NaN'. Una limpieza exhaustiva fue esencial para homogeneizar estas entradas como NaN y así permitir su correcto procesamiento como datos ausentes, sentando las bases para un análisis.

La estrategia de análisis se centró en los rangos temporales válidos para cada serie individual, descartando los años anteriores al First Year y posteriores al Last Year de cada árbol, que no contienen información real. Dentro de estos intervalos, la detección y manejo de valores atípicos se realizó de forma específica: aplicando la distancia de Cook para series con residuos normales y el rango intercuartílico (IQR) para las que no seguían una distribución normal. Este enfoque dual nos permitió identificar y manejar observaciones influyentes o erróneas de manera estadísticamente rigurosa, sin aplicar un criterio único a todas las series.

El análisis de las series, tanto de forma individual como agrupada por especie, arrojó resultados reveladores. Para especies como Cedrus atlantica y Quercus robur, se observaron patrones distribucionales claros y consistentes (normales o unimodales), lo que sugiere una respuesta fisiológica homogénea al clima dentro de su área de distribución. Por el contrario, en especies como Pinus sylvestris y Pinus leucodermis, la distribución de los valores de $\delta^{13}C$ fue mucho más amplia y asimétrica, sin un patrón claro.

Esta marcada diferencia nos lleva a que la fuerte variabilidad observada en algunas especies no es atribuible únicamente a su fisiología, sino que es un reflejo directo de la amplia diversidad de condiciones climáticas y geográficas de los sitios donde crecen. Un árbol de la misma especie puede experimentar condiciones de estrés hídrico radicalmente diferentes en el norte de Europa comparado con el sur, y esta señal ambiental queda impresa en los isótopos de sus anillos.

References

[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; Krapiec, 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-Krapiec, 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