Tarea 1 Reporte Ciencia de Datos
Parte Práctica

Rodolfo Ramírez
Guillermo Aguilar
Jesús Alonso

September 13, 2025

1 Introducción

Los siguientes datos fueron recabados dentro del proyecto ISONET cuyo objetivo era crear una red amplia de datos isotópicos en árboles de toda Europa. El proyecto fue financiado dentro del quinto EC Framework Programme "Energy, Environment and Sustainable Development". Estos datos ayudan a comprender de una manera más clara el sistema climático europeo.

2 Objetivo

Los datos provienen del proyecto europeo ISONET (400 Years of Annual Reconstructions of European Climate Variability Using a Highly Resolved Isotopic Network), desarrollado en el marco del Quinto Programa Marco de la Comisión Europea ("Energy, Environment and Sustainable Development")

El objetivo central fue reconstruir la variabilidad climática europea de los últimos 400 años mediante indicadores altamente resolutivos en el tiempo: los isótopos estables en la celulosa de anillos de árboles.

La variable principal contenida en la base de datos corresponde a la razón isotópica del carbono $\delta^{13}C$ en la celulosa de anillos de crecimiento anual de distintas especies arbóreas. Este valor se expresa en ‰ con respecto al estándar internacional Vienna Pee Dee Belemnite (VPDB). El registro se organiza en series temporales anuales (1600–2003 CE, dependiendo del sitio), y se acompaña de metadatos específicos: código y nombre del sitio, coordenadas geográficas (latitud y longitud), especie muestreada, elevación promedio, y los años inicial y final de cobertura de la cronología.

El conjunto de cronologías de $\delta^{13}C$ ofrece múltiples aplicaciones en el ámbito de las ciencias del clima y de la ecología forestal. Entre las más relevantes se encuentran:

- **Reconstrucciones climáticas históricas**: permiten inferir series de sequías, humedad estival y disponibilidad hídrica a lo largo de los últimos cuatro siglos en distintas regiones de Europa.
- **Estudio de la eficiencia en el uso del agua (WUE)**: el análisis de tendencias de $\delta^{13}C$ en relación con el aumento de las concentraciones de CO_2 atmosférico durante la era industrial posibilita evaluar cambios fisiológicos en los bosques europeos.
- **Comparaciones espaciales**: los registros de diferentes regiones (boreales, atlánticas, mediterráneas y alpinas) permiten identificar patrones contrastantes en la respuesta isotópica a la variabilidad climática.
- **Validación de modelos climáticos y ecofisiológicos**: los datos empíricos de $\delta^{13}C$ constituyen una referencia independiente para contrastar simulaciones de evapotranspiración y fotosíntesis a escala continental.
- **Análisis de eventos extremos**: la detección de valores anómalos facilita la identificación de sequías severas, ondas de calor u otras anomalías hidroclimáticas pasadas, contribuyendo al estudio de su frecuencia e intensidad.

3 Exploración inicial de los datos

La tabla proporcionada cuenta con 24 series de tiempo de diferentes localidades en Europa que registran los isótopos de carbono que se encuentran en la celulosa de los anillos de los árboles de estas localidades. Estos anillos proporcionan información acerca de las condiciones climáticas de la localidad del árbol de años atrás. El significado de cada una de las variables del archivo de datos es el siguiente:

- Site Code: Código del sitio dónde se tomó la muestra.
- Site Name: El sitio dónde se tomó la muestra.
- Country: País del sitio.
- Latitude: Latitud del sitio.
- Longitude: Longitud del sitio.
- Species: Especie del Arbol del que fue tomada la muestra.
- First year CE: Primer año del que se tienen datos.
- Last year CE : Último año del que se tienen datos.
- elevation a.s.l. : Elevación sobre el nivel del mar de los arboles.
- Year CE: Fecha del anillo del arbol.
- 13CVPDB: 13C/12C ratio en por mil contra Vienna PDB (VPDB). El ratio se compara con un estandar internacional y es el número que se reporta.

Los siguientes datos nos pueden dar una idea del nivel de carbono por año en cada una de las localidades registradas. Esto puede ser últil para conocer periodos de sequía. Cuando una planta sufre de falta de agua (sequía), cierra sus estomas (poros de las hojas) para conservarla. Esto limita la entrada de CO_2, la planta fotosintetiza con el carbono interno disponible y, como resultado, se enriquece en C^{13} . Una serie temporal de $\delta^{13}C$ puede revelar así períodos de sequía severa a lo largo de siglos.

4 Detección de problemas en los datos

Se analizó el porcentaje de datos faltantes por cada una de las variables, estos datos se encuentran en la tabla. Es posible saber el porcentaje de faltantes de la variable $13CVPDB$ segmentado por el Código del Sitio, esto lo visualizamos en la tabla (1). Las demás variables no poseen datos faltantes a excepción de la elevación sobre el nivel del mar que corresponde al registro del Código de Sitio **AHI**.

Código	Porcentaje	Código	Porcentaje
BRO	74.88%	ILO	0.74%
CAV	10.10%	INA	0.74%
CAZ	0.74%	AHI	30.05%
COL	31.03%	LAI	52.71%
DRA	44.34%	LIL	1.72%
FON	30.30%	LOC	37.19%
GUT	0.74%	NIE1	7.14%
		NIE2	7.14%
PED	1.23%	PAN	53.94%
POE	0.74%	REN	9.61%
SER	1.48%	SUW	0.25%
VIG	19.21%	VIN	63.05%
WIN	42.86%	WOB	2.71%

Table 1: Porcentaje de Datos Faltantes

Dado que los datos llevan una cronología, no podemos descartar que existan patrones temporales como la tendencia en los datos. Analizando algunas series de tiempo como la de Pinar de Lillo (LIL) Figura (1) y Monte Pollino (SER) Figura (2) se ve una clara tendencia de los datos. Así, bajo la sospecha de la existencia de tendencia en nuestros datos, calculamos las diferencias a un periodo de cada una de las series de tiempo y analizamos a detalle las subidas y bajadas muy grandes (atípicas dentro de la distirbución normal de subidas y bajadas), esto usando el z-score o el IQR después, marcamos los datos "atípicos" para investigarlos y ver si estos se tratan de Outliers.

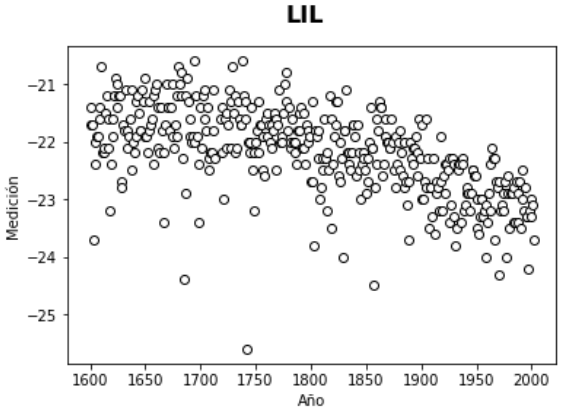


Figure 1: Gráfico de Puntos de las mediciones tomadas en Pinar de Lillo (LIL).

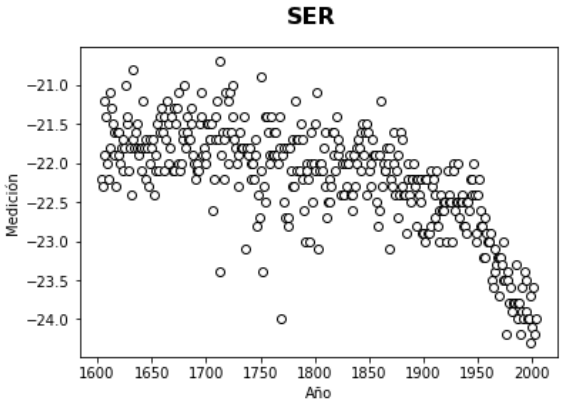


Figure 2: Gráfico de Puntos de las mediciones tomadas en Monte Pollino (SER).

Para ilustrar la detección de Outliers, usamos los dos ejemplos mencionados anteriormente. En el lado derecho de la Figura (3) vemos la detección de Outliers sobre la serie diferenciada para la serie de tiempo de LIL, del lado derecho se marcan los datos que tuvieron este incremento o decremento fuera de lo común en la serie original. Los misma información pero usando el IQR para detectar los outliers en SER se encuentran en la Figura 4. La tabla con el número de outliers para cada sitio por el método de Z-Score se encuentran en la Tabla (3) y por el método de IQR se encuentran en la Tabla (2). Observando las gráficas vemos que dentro de los datos atípicos que encontramos se encuentran posibles outliers, cambios de tendencia o simplemente datos atípicos en los incrementos o decrementos que no necesariamente son outliers en la serie original (puntos rojos dentro de la nube de puntos).

Otra forma de detección de Outliers es haciendo una regresión sobre los datos dónde la variable dependiente es el tiempo y la variable de interés las mediciones. De seguir los residuos una distirbución Normal, es posible considerar como datos outliers a puntos cuyo residuo en valor absoluto esté alejado más de 3 desviaciones estandar de la media. Esto pues una distribución normal acumula tan sólo 3% de probabiliadad en esta región. Se ajustó una Regresión lineal como se mencionó y se realizó una prueba de Shapiro Wilks para evaluar la Normalidad de los datos, los p-valores los encontramos en la Tabla (4), se procedió a la identificación de datos outliers con el método mencionado. El número de outliers en los sitios cuyos errores pasaron la prueba de normalidad se observan en la Tabla (5). El caso particular de AHI se puede ver en la Figura (5) y el de CAV en la Figura (6). El número de errores para los sitios cuyos errores no eran Normales pero aún así se ajustó una regresión los encontramos en la Tabla (6). Ejemplos particulares los vemos para NIE1 y LIL en las Figuras (7) y (8) respectivamente, cómo podemos ver, a pesar de no satisface el supuesto de normalidad en los errores, los datos marcados como outliers parecen efectivamente serlo.

Se encontraron inconsistencias en la información proporcionada en la tabla pues en el caso de las mediciones de Dransfel (DRA) se indica que la última medición de la que se tiene registro es la de 1999 sin embargo la tabla aún tiene datos para ese sitio hasta el 2002, caso similar ocurre con LAI.

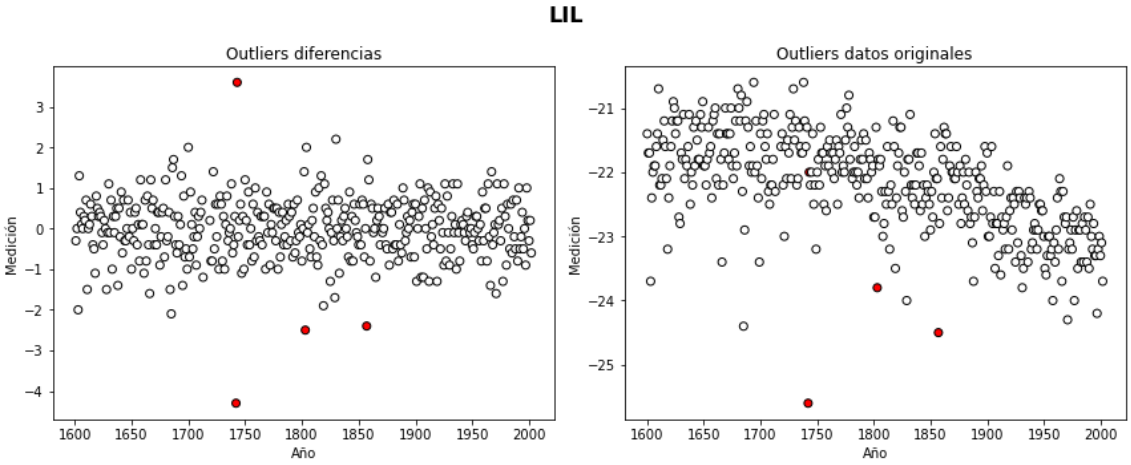


Figure 3: Identificación de Outliers de la serie de tiempo de LIL, mediante el z-score y el método de Diferencias.

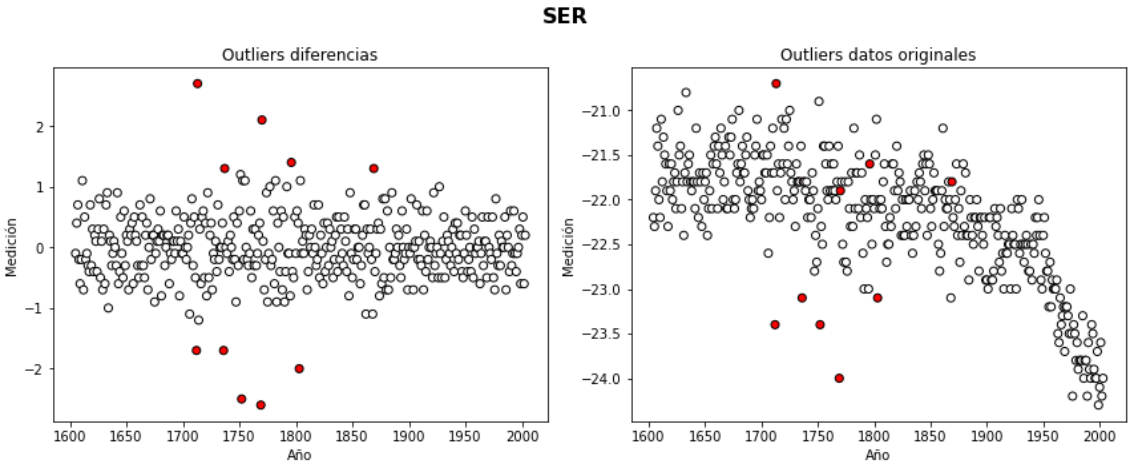


Figure 4: Identificación de Outliers de la serie de tiempo de SER, mediante el IQR y el método de Diferencias.

5 Manejo de Datos Faltantes

Por lo leído en los documentos, entendemos que los datos faltantes se deben a que el árbol no contenía información referente a ese año esto posiblemente debido a que era difícil la extracción de la muestra para un año en particular, esto ocurre usualmente porque los anillos de celulosa correspondientes a cierto año son muy delgados para poder extraerlos sin contaminarlos . Se realizó una investigación y se sabe que los tamaños de estos anillos están asociados al estrés hídrico de la región, variable que no es observable, sin embargo, observamos el año de la muestra que esta directamente asociado con el estrés hídrico por ende consideramos que el tipo de datos faltantes es MAR.

Sabemos que en la regresión lineal, una predicción menos su estimación entre su varianza, usando la varianza estimada se distribuye como una t-student con n-2 grados de libertad i.e.

$$\frac{y - \hat{y}}{\sqrt{\mathbb{V}(\hat{y}) + MS_{\text{Res}}}} \sim t_{n-2}$$

dónde $\mathbb{V}(\hat{y})$ se calcula cómo:

$$\mathbb{V}(\hat{y}) := MS_{\text{Res}} \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

así, para poder imputar datos, podemos tomarnos una simulación de la predicción. Consideramos que no es posible imputar los missings de los valores fuera del rango de fechas de las que se encontraron mediciones esto dado que , por lo que sólo imputaremos aquellos datos faltantes que sí se encuentran dentro del rango de fechas de las que se encontraron mediciones. En la Tabla (7) podemos encontrar el número de datos faltantes que requieren de ser imputados por Sitio.

Site Code	Número de Outliers
BRO	0
CAV	9
CAZ	4
COL	6
DRA	2
FON	13
GUT	2
ILO	3
INA	3
AHI	3
LAI	0
LIL	10
LOC	5
NIE1	13
NIE2	8
PAN	7
PED	4
POE	8
REN	2
SER	10
SUW	14
VIG	2
VIN	2
WIN	4
WOB	6

Table 2: Número de Outliers de las diferencias bajo el IQR.

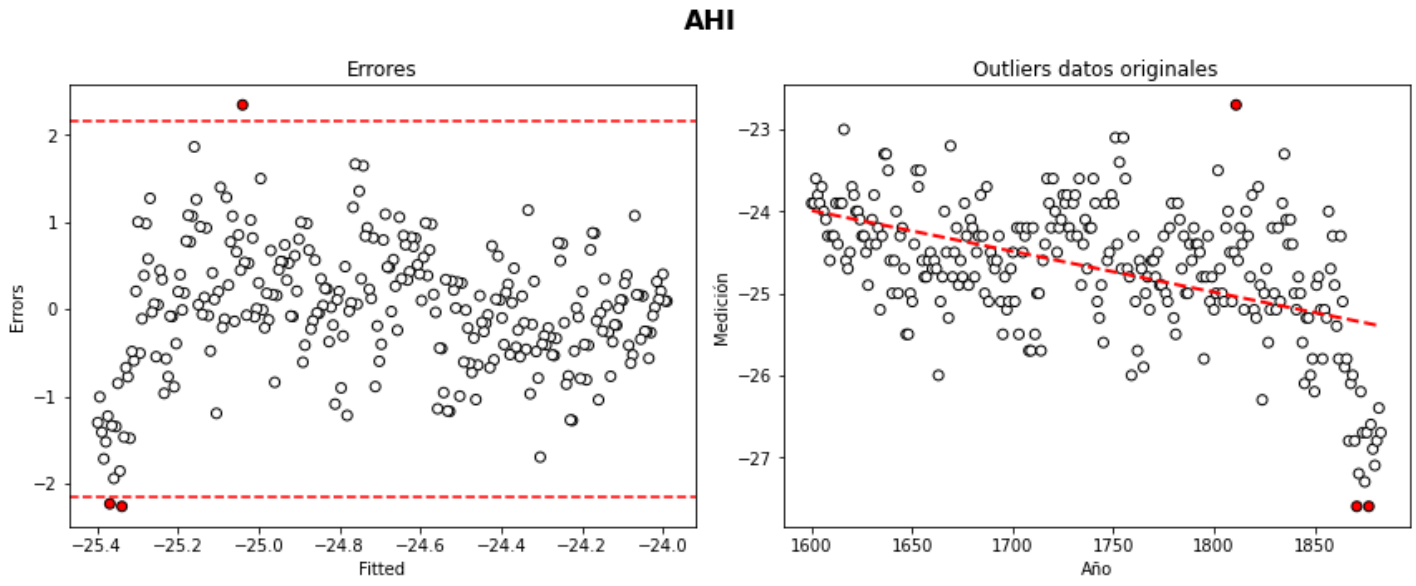


Figure 5: Identificación de Outliers de la serie de tiempo de AHI, mediante una regresión lineal.

Site Code	Número de Outliers
BRO	0
CAV	1
CAZ	2
COL	4
DRA	1
FON	3
GUT	1
ILO	1
INA	1
AHI	0
LAI	0
LIL	4
LOC	2
NIE1	6
NIE2	5
PAN	0
PED	4
POE	2
REN	2
SER	7
SUW	1
VIG	2
VIN	1
WIN	0
WOB	2

Table 3: Número de Outliers de las diferencias bajo el Z-Score.

Código	p-Valor	Código	p-Valor
BRO	0.68	ILO	0.00
CAV	0.42	INA	0.00
CAZ	0.00	AHI	0.17
COL	0.00	LAI	0.18
DRA	0.26	LIL	0.00
FON	0.00	LOC	0.91
GUT	0.19	NIE1	0.00
		NIE2	0.00
PED	0.01	PAN	0.79
POE	0.01	REN	0.55
SER	0.00	SUW	0.51
VIG	0.56	VIN	0.09
WIN	0.22	WOB	0.00

Table 4: Resultados de la prueba de Shapiro-Wilk (p-value) por sitio.

Dado que encontramos datos atípicos y estos pueden jalar la recta de regresión, estos serán retirados de los datos para poder imputar los missings. Después de seguir la metodología mencionada en las siguientes gráficas se muestran ejemplos de los datos imputados: Para los missings de LIL Figura(9), para FON la Figura (10) y para PED la Figura (11). Otra técnica que se puede usar para imputar es simplemente usando la predicción de la regresión. Consideramos esta técnica la más adecuada dada la pequeña cantidad de series de tiempo con outliers y, además, la practicidad para calcularlos.

Código	Número de Outliers	Código	Número de Outliers
BRO	0	AHI	3
CAV	2	LAI	0
DRA	1	LOC	1
GUT	2	PAN	0
REN	0	SUW	1
VIG	1	VIN	1
WIN	0		

Table 5: Número de Outliers Ajustando una Regresión a los Sitios cuyos errores pasaron la prueba de Normalidad.

Site Code	Número de outliers	Site Code	Número de outliers
CAZ	1	LIL	4
COL	2	NIE1	5
FON	2	NIE2	0
ILO	3	PED	2
INA	0	POE	2
SER	2	WOB	2

Table 6: Número de Outliers Ajustando una Regresión a los Sitios cuyos errores no pasaron la prueba de Normalidad.

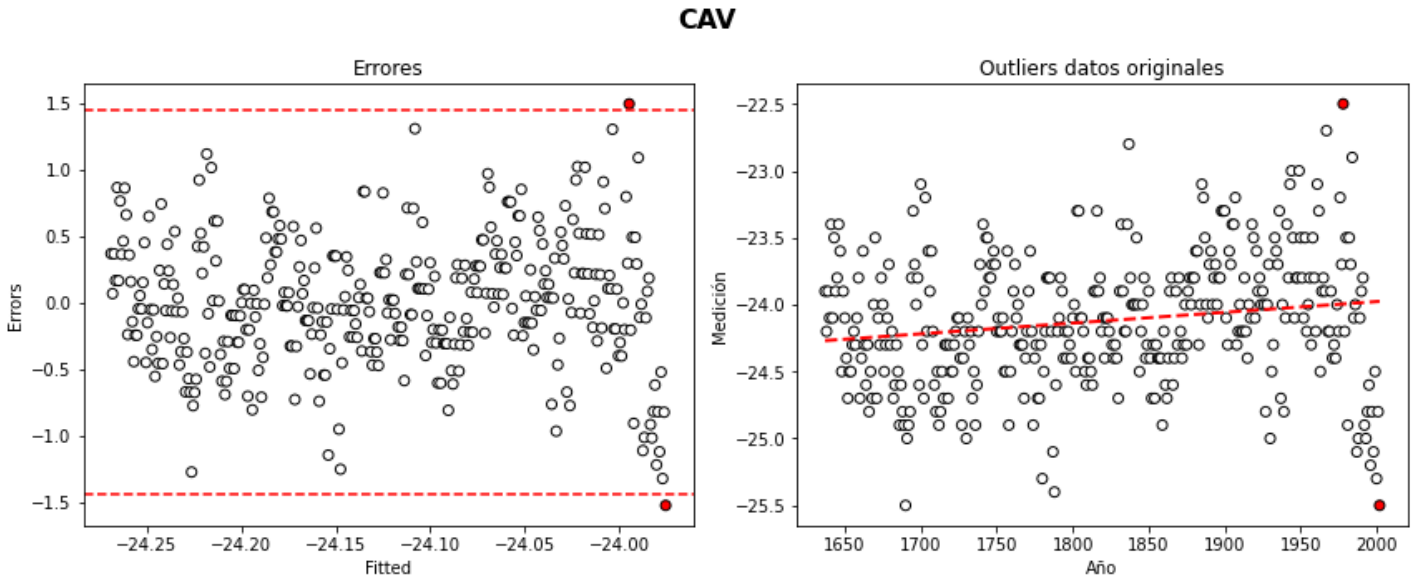


Figure 6: Identificación de Outliers de la serie de tiempo de CAV, mediante una regresión lineal.

6 Codificación y escalamiento.

Las variables categóricas a trabajar que requerirían transformación serian "Site name", "Country" y "Species". Dicho lo anterior, podemos considerar a la one-hot encoding como la transformacion mas acertada del modelo. Tomando el caso de la variable "Species" el modelo one-hot enconding nos evita introducir un "orden falso" entre las especies, ya que Cada especie se representa de forma independiente. De echo en los datos obtenidos el valor tiende a variar con forme el año de muestra. Ademas funciona muy bien en clasificadores lineales o de distancias. Similar mente podemos dar el mismo argumento para las variables categóricas "Site name" y "Country". Ya que estos representan sitios de las zonas de Europa, no seria conveniente darles un orden, a menos que los datos obtenidos sugieran lo contrario. Ya que los datos obtenido sugieren que las variables categóricas mas relevantes para un modelo serían probablemente Country y Species, ya que podrían tener relación con las diferencias en los valores de isótopos de carbono. Una muestra de la codificación One-Hot encoding se puede visualizar en la Tabla 8.

En esta parte se escalaron 4 variables numéricas, las cuales corresponden a "Year CE", los datos de isotopos de carbono

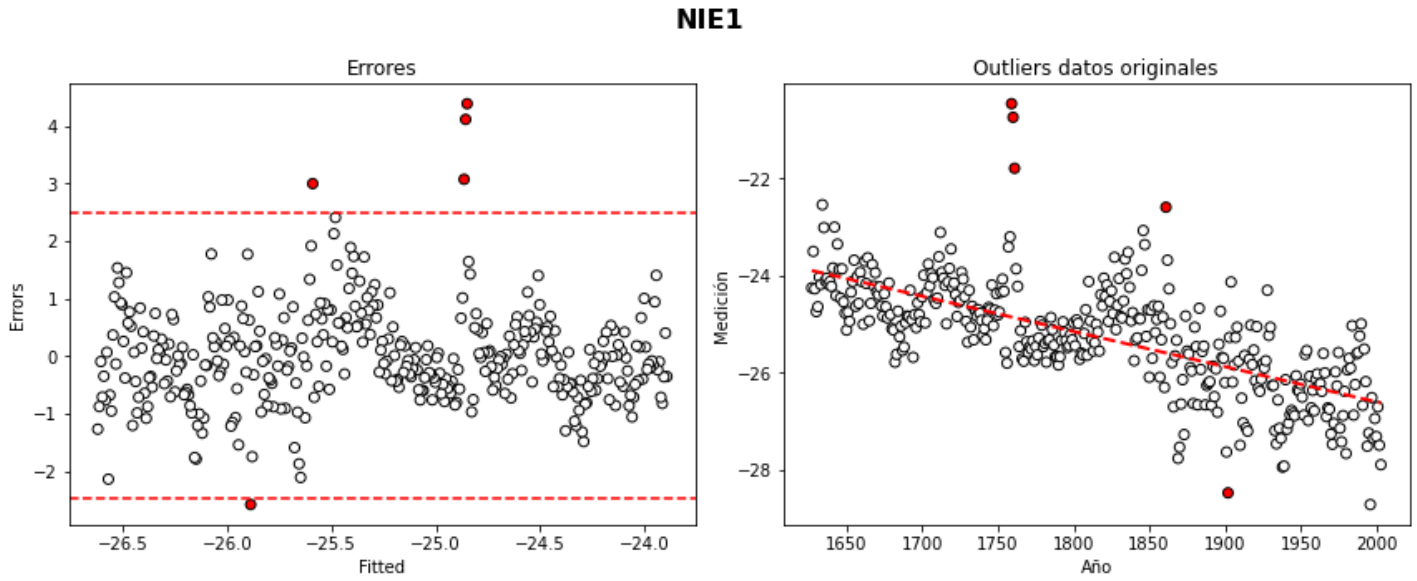


Figure 7: Identificación de Outliers de la serie de tiempo de NIE1, mediante una regresión lineal.

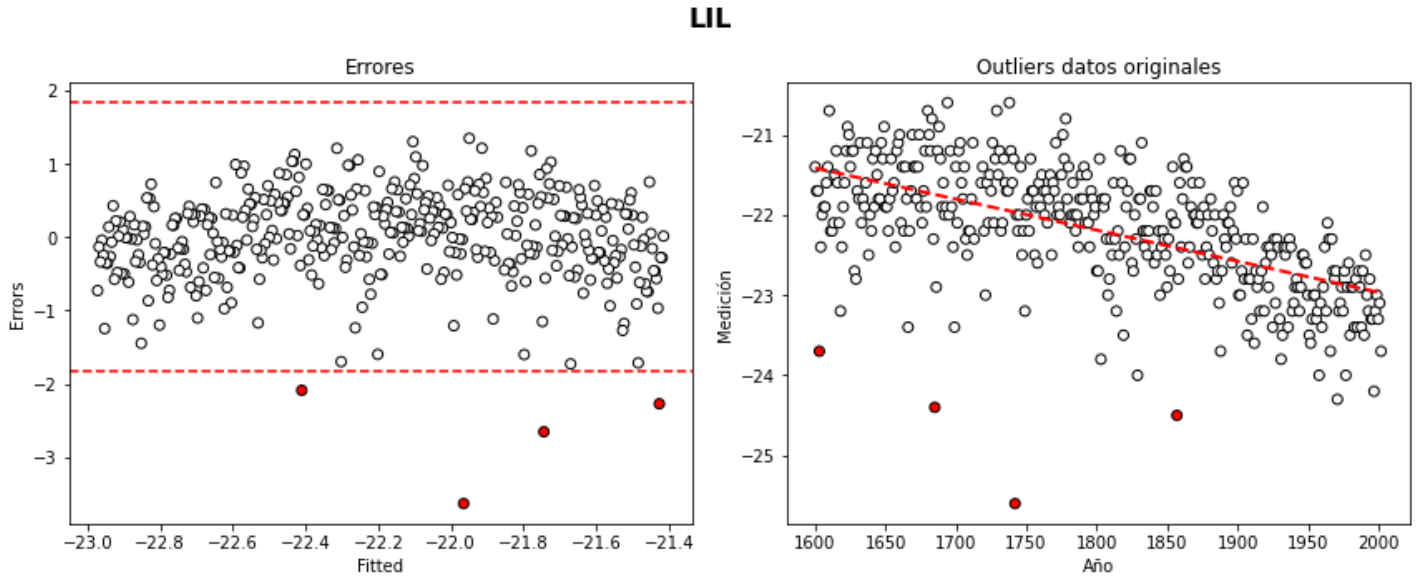


Figure 8: Identificación de Outliers de la serie de tiempo de LIL, mediante una regresión lineal.

de las especies de arboles

1. "Pinus uncinata" cuya muestra fue tomada en Pedraforca Spain.
2. "Pinus sylvestris" cuya muestra fue tomada en Suwalki Poland.
3. "Pinus sylvestris" cuya muestra fue tomada en Windsor United Kingdom.

Dado la falta de datos, se trataron utilizando la mediana de los datos respectivos, para evitar posibles errores con el escalamiento min-max y z-score. Dicho lo anterior se obtubieron las siguientes gráficas.

Estas ultimas imágenes pueden indicarnos que si queremos graficar y comparar tendencias de diferentes sitios/especies en el tiempo, Min-Max es nuestra mejor elección. Aunque si queremos analizar patrones estadísticos o entrenar un modelo que dependa de dispersión, tendríamos que elegir Z-score.

Site Code	Número de Missings	Site Code	Número de Missings
BRO	0	ILO	0
CAV	1	INA	0
CAZ	0	AHI	0
COL	108	LAI	0
DRA	1	LIL	4
FON	118	LOC	0
GUT	1	NIE1	0
PAN	0	NIE2	0
PED	3	POE	0
REN	21	SER	0
SUW	0	VIG	1
VIN	0	WIN	9
WOB	5		

Table 7: Número de Missings por Site Code

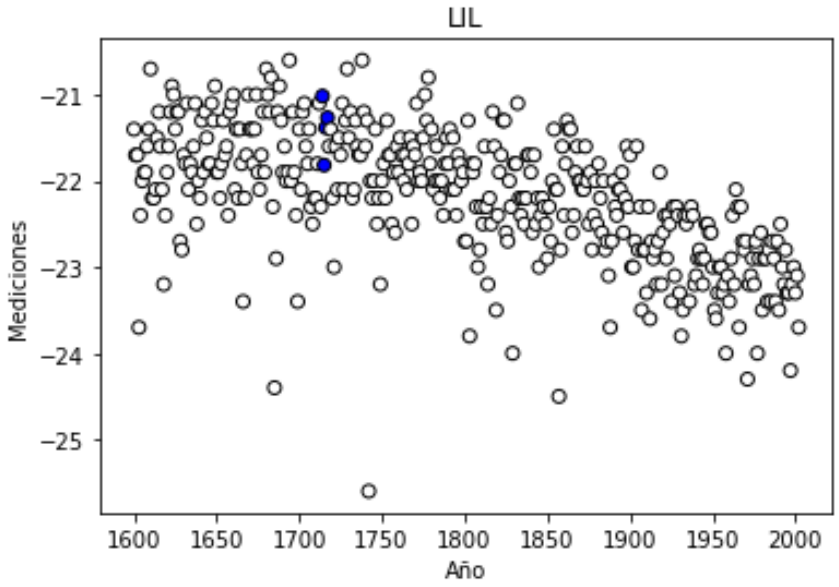


Figure 9: Imputación de Datos de LIL, puntos azules son los datos imputados.

Site	Species	P. leucodermis	P. nigra	P. sylvestris	P. uncinata	Q. petraea	Q. robur
BRO	Q. robur	0	0	0	0	0	1
CAV	Q. petraea	0	0	0	0	1	0
CAZ	P. nigra	0	1	0	0	0	0
COL	C. atlantica	0	0	0	0	0	0
DRA	Q. petraea	0	0	0	0	1	0

Table 8: Distribución de especies por Site Code

7 Visualización Exploratoria

Trabajando los histogramas utilizando 9 columnas, las especies y localidades propuestas en el ejemplo anterior, obtenemos los siguientes histogramas

Estos parecen indicar que los datos distribuyen de manera normal, con una media aproximada de -23.07 y una desviación estándar de 0.62163348. Pero esto ultimo no lo podemos asegurar, ya que en la base de datos hay datos faltantes.

Aunque existen datos faltantes, sabemos que los datos faltantes son MC[illegible] entonces el dato faltande no depende ni de los valores observados ni de los no observados. Por lo que usamos la imputación sobre la media. A continuación veremos una comparación entre los datos imputados y sin imputar

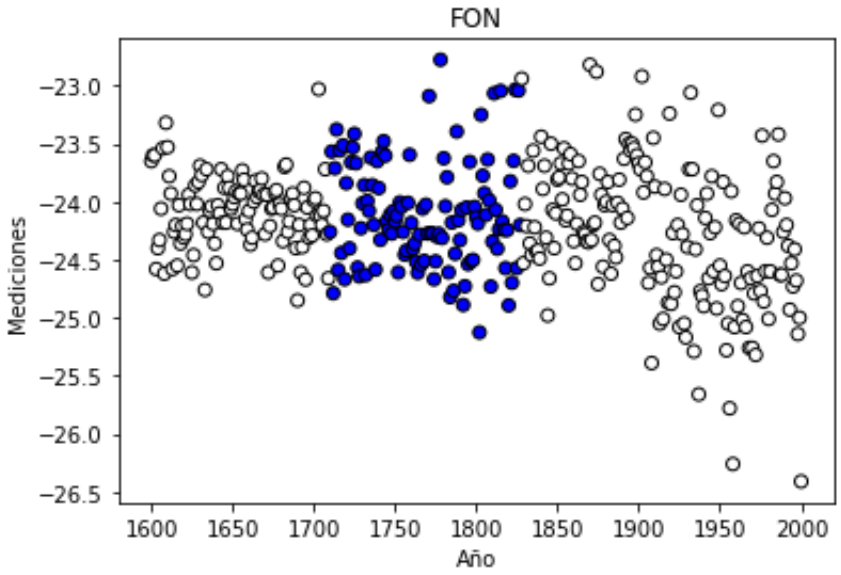


Figure 10: Imputación de Datos de FON, puntos azules son los datos imputados.

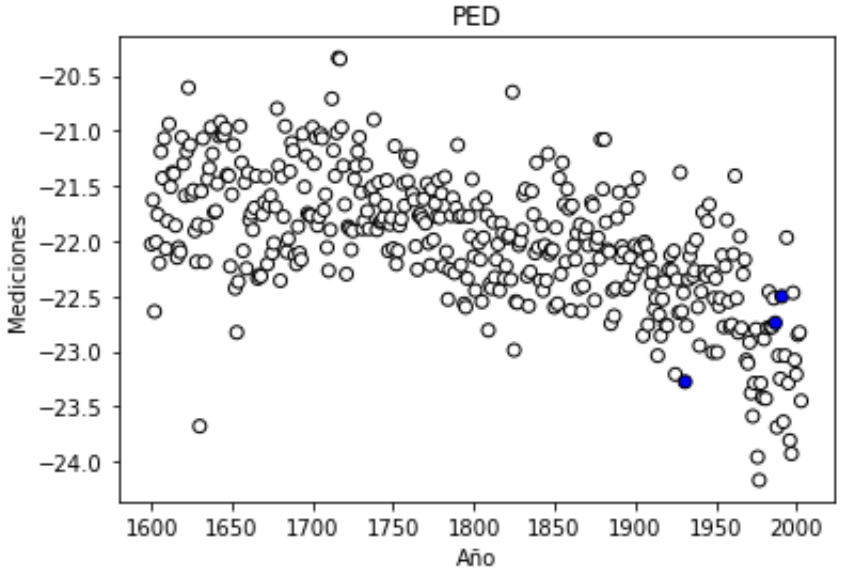


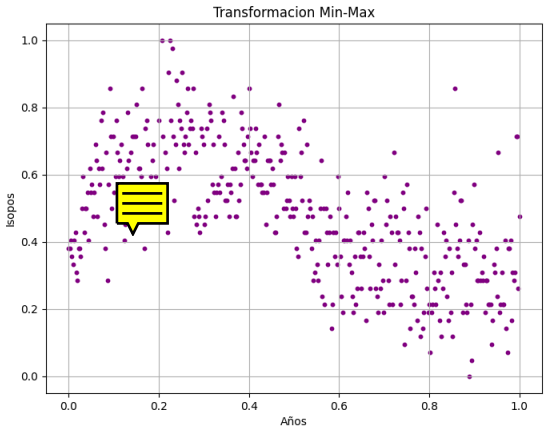
Figure 11: Imputación de Datos de PED, puntos azules son los datos imputados.

A juzgar por los los datos obtenidos en Pinus sylvestris Polan de Polan y Spain tenemos una tendencia decreciente entre los años venideros y el valor del isopo de carbono, es decir que hubo más disponibilidad de CO_2 en dicho año. Finalmente para podemos analizar la precencia de Outliers que puedan llevarnos a conclusiones erróneas en el estudio, podemos usar IQR, por lo que obtenemos

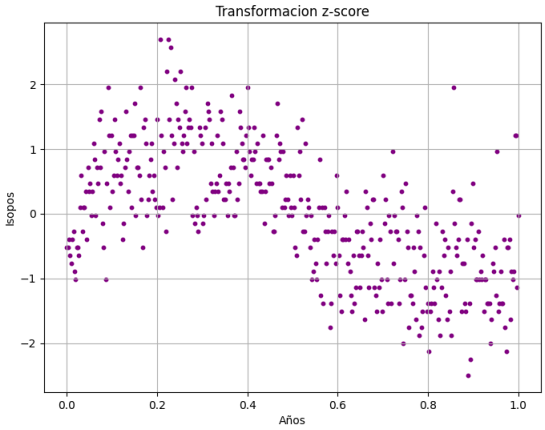
Hay mas probabilidad de tener outliers en los datos de Pinus sylvestris de Polan, pero notemos que son tan pocos y acumulados que la tendencia decreciente de los datos se mantiene, además de mantener el histograma con un comportamiento similar a la normal.

8 Reflexión Crítica

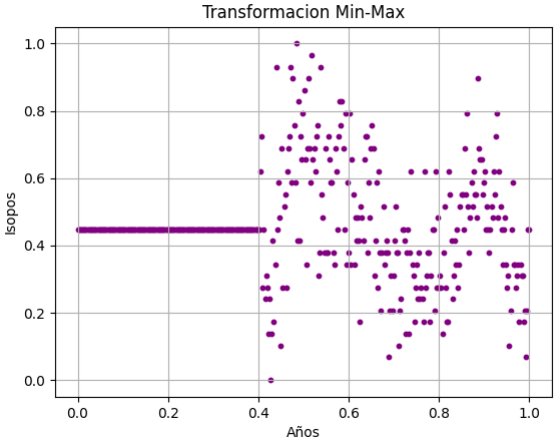
Antes de empezar a modelar cualquier conjunto de datos, es necesario comprender lo más posible su origen y el funcionamiento del mecanismo que los genera, así como el objetvio de estudio y las preguntas que buscamos responder. Esto

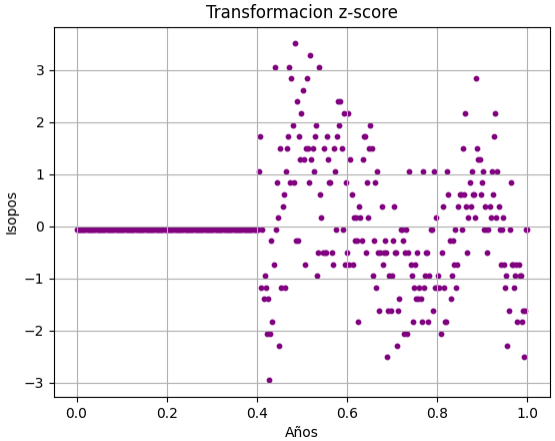


(a) Pinus sylvestris Polan, escalado con Min-Max



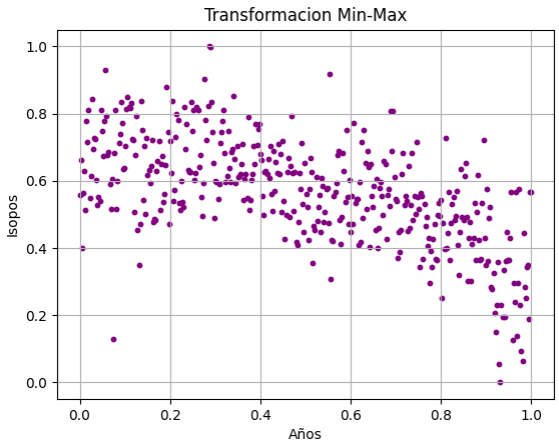
(b) Pinus sylvestris Polan, escalado con z-score



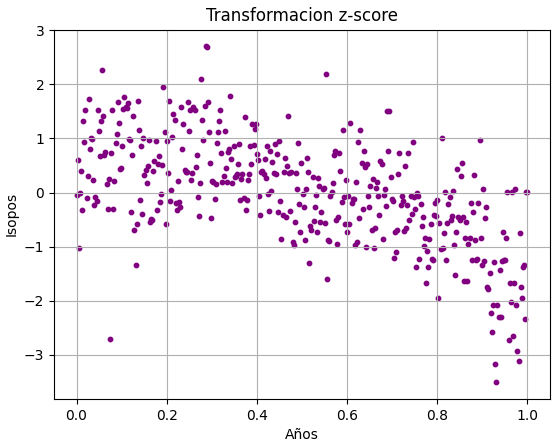


(a) Pinus sylvestris United Kingdom, escalado con Min-Max (b) Pinus sylvestris United Kingdom, escalado con z-score

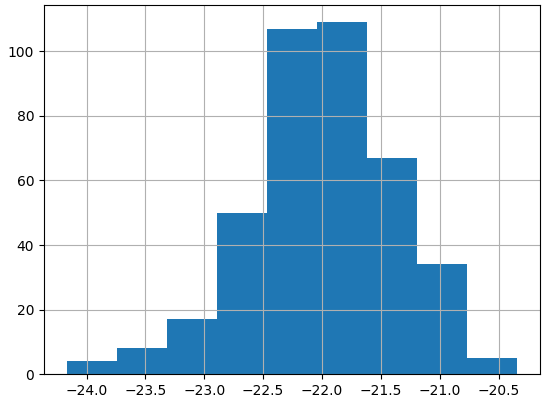
permite dar una mejor interpretabilidad a los resultados e inferencias que se realicen sobre los datos. Posteriormente, se debe llevar a cabo un proceso de familiarización con los datos; para esto, es necesaria la limpieza de los mismos con el fin de estudiarlos. En esta etapa se suelen encontrar inconsistencias, errores, duplicados, que de no ser tratados podrían sesgar o provocar análisis erróneos. En el caso particular de esta práctica encontramos inconsistencias en el dato que proporcionaba



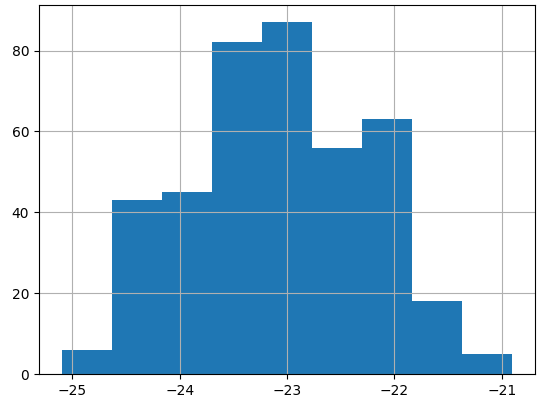
(a) Pinus uncinata Spain, escalado con Min-Max



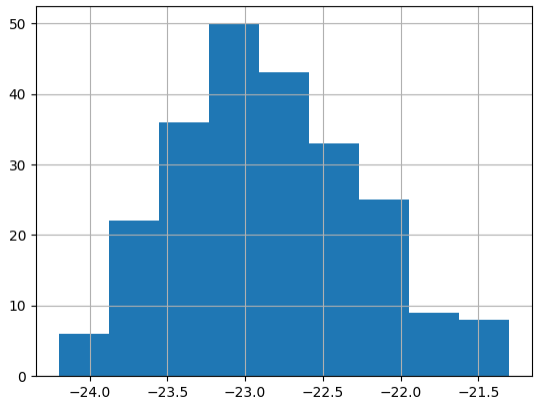
(b) Pinus uncinata Spai, escalado con z-score



(a) Pinus sylvestris Spain



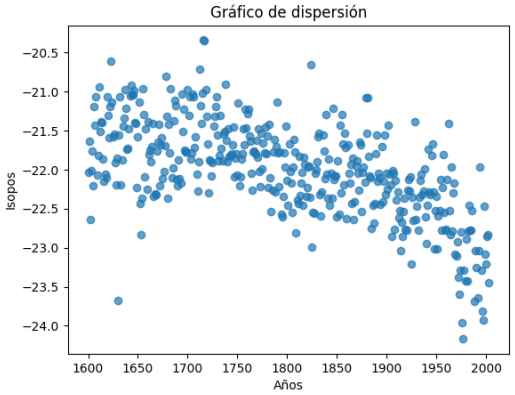
(b) Pinus sylvestris Poland



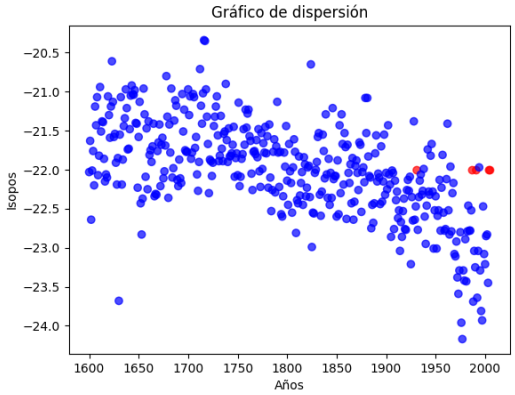
(c) Pinus sylvestris United Kingdom

la última fecha con dato no nulo de algunos Sitios, esta inconsistencia provocaba que el porcentaje de valores no nulos fuera mayor al 100%. lo que es una conclusión errónea.

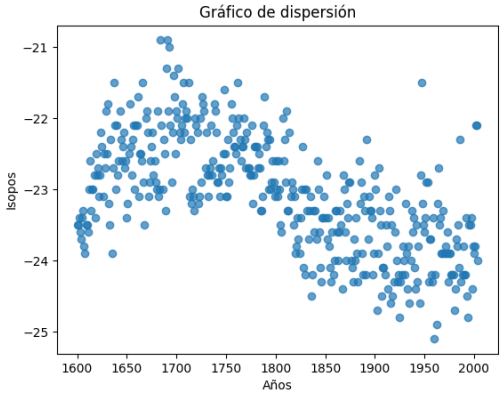
Es importante justificar en todo momento el uso de los métodos estadísticos que se implementan, de otra forma la intuición nos puede llevar a conclusiones que no son del todo correctas. En particular podemos hablar sobre el intento de



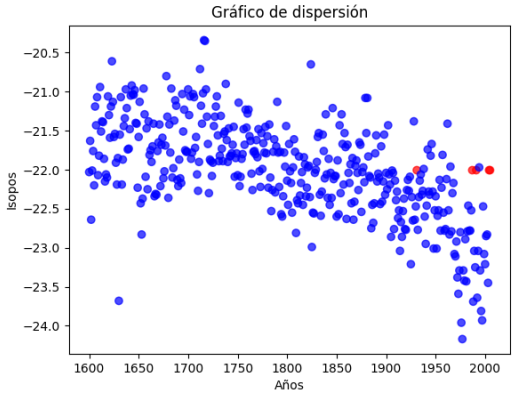
(a) Pinus sylvestris Polan



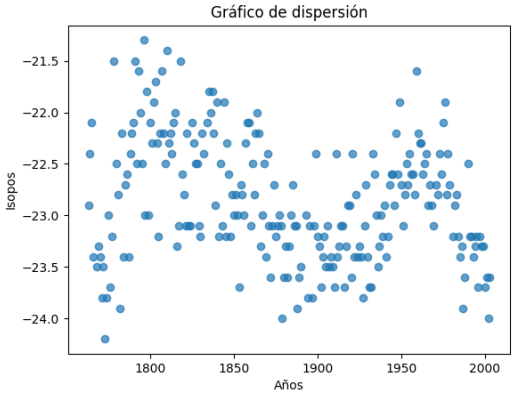
(b) Pinus sylvestris Spain



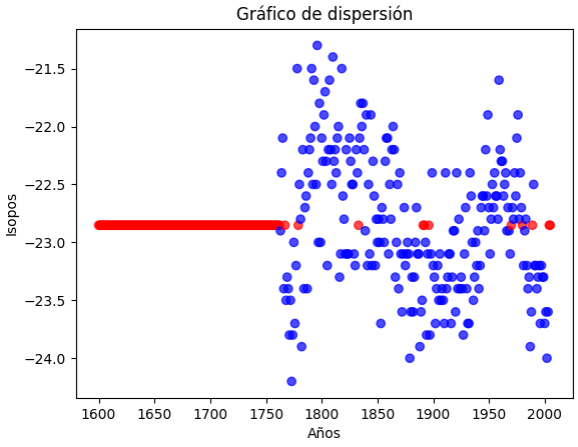
(a) Pinus sylvestris United Kingdom



(b) Pinus sylvestris Spain



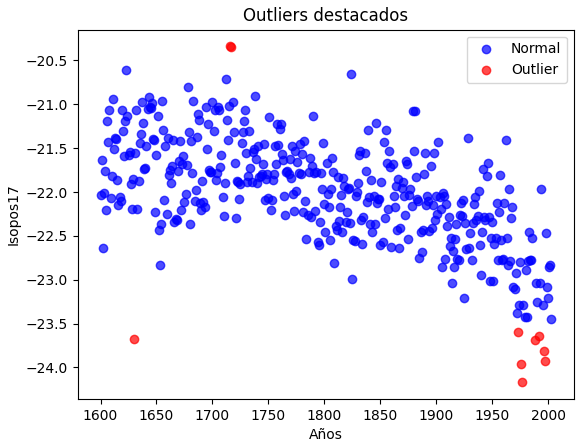
(a) Pinus sylvestris Poland



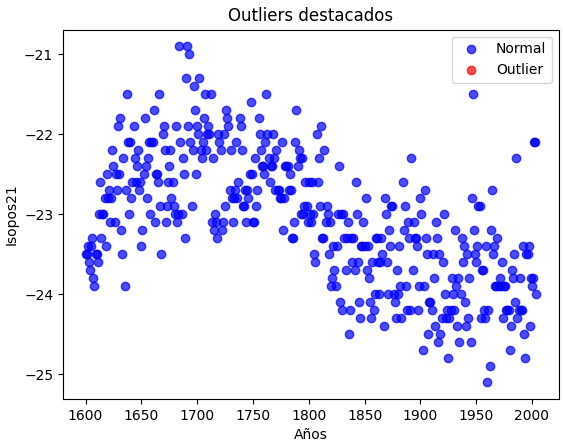
(b) Pinus sylvestris Poland

detectar valores outliers apartir de la diferenciación de la serie. Cómo vimos en el ejemplo de la serie SER, algunos de los valores que salían como puntos outliers bajo esta metodología, realmente no lo eran pues se encontraban dentro de la nube de puntos o estos también podían representar un cambio en la tendencia de la serie. Al usar la regresión lineal para la detección de outliers, esta nos llevó a conclusiones más coherentes sobre los datos que podían ser atípicos.

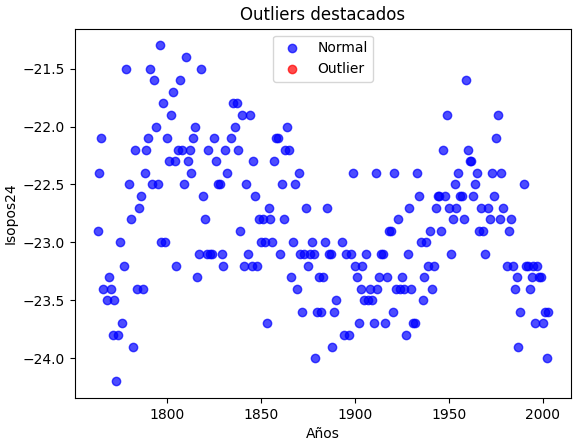
La detección del tipo de dato faltante es sumamente importante antes de tratar de imputarlos o quitar los registros, esto dado que la eliminación de estos podría inducir un sesgo en nuestro análisis, es el ejemplo de los missings tipo MNAR. Por otro lado, la imputación es una herramienta que permite mantener el tamaño de la muestra intacta así como la



(a) Pinus sylvestris Polan



(b) Pinus sylvestris United Kingdom



(c) Pinus sylvestris United Kingdom

información en esta, esto dado que si el registro posee otras variables, quitar los datos podría hacernos perder información valiosa, una estrategia de imputación bien justificada estadísticamente, nos permite evitar este problema. El ajuste de la recta de regresión a los datos y el cumplimiento de los supuestos para algunas series, nos permitió imputar los datos satisfactoriamente, de forma que las gráficas con los datos imputados se ven naturales.

El reescalamiento de una variable puede ser una metodología útil dependiendo del fin que se tenga con el estudio. En el caso de hacer inferencia sobre las variables más importantes en una regresión, el reescalamiento de una variable nos permite hacer comparables las betas de esta. En caso de algunos métodos de machine learning, en particular los que dependen de distancias, estos pueden sesgar sus resultados hacia las variables con mayor rango numérico.

En conjunto, estos procesos de preparación de datos garantizan que el modelo estadístico parta de insumos consistentes, comparables y completos.

References

[1] L. Leticia Ramírez Ramírez (2024). *Modelos Estadísticos I. Modelos Lineales y Lineales Generalizados.*