El origen de los datos: ¿qué institución o proyecto los generó?

Estos datos fueron recolectados por el proyecto ISONET, financiado y coordinado por la Comisión Europea. El proyecto busca mejorar la comprensión de los sistemas climáticos europeos, proporcionando datos cuantitativos independientes para la verificación de modelos y la formulación de políticas públicas. Para ello, una red de 24 sitios ofrece cobertura dendrocronológica desde la Península Ibérica hasta Fennoscandia, incluyendo además Caledonia y el Tirol.

El propósito de la recolección: ¿qué fenómenos científicos buscan representar?

Se analizan las razones de isótopos estables (C, H, O) en series temporales resueltas anualmente y, en conjunto con modelos de procesos no lineales desarrollados dentro del proyecto, se reconstruyen los regímenes climáticos del pasado (temperatura, humedad relativa y características de precipitación) durante los últimos 400 años. La variabilidad climática se aborda en tres escalas temporales:

- Década–siglo: dominancia de masas de agua/aire.
- Interanual: cuantificación de la variabilidad de referencia, eventos extremos y tendencias recientes.
- Intraanual: exploración de alta resolución de las señales de estacionalidad en los anillos de los árboles.

De este modo, ISONET trasciende los análisis dendrocronológicos tradicionales, al ofrecer una investigación e interpretación basadas en procesos espaciales con un alcance mucho más amplio.

El significado de las variables medidas: ¿qué información aportan y cómo se relacionan con el contexto dentro del contexto en que fueron capturados?

Las variables medidas en este conjunto de datos aportan información tanto sobre las características geográficas y biológicas de los sitios de muestreo. Basándose en las propiedades isotópicas de los anillos de los árboles, fundamentales para la reconstrucción climática.

- **Código y nombre del sitio:** identifican de manera única cada localidad de muestreo, permitiendo distinguir entre regiones y facilitar el análisis comparativo.
- **Latitud y longitud:** ubican con precisión geográfica cada sitio dentro del mapa europeo, lo cual es esencial para analizar patrones espaciales de variabilidad climática.

- **Especie (nombre en latín):** especifica el tipo de árbol muestreado, relevante porque diferentes especies responden de manera distinta a las condiciones ambientales.
- **Año inicial y final:** señalan el periodo temporal cubierto por los registros de δ13C, lo que establece el horizonte histórico de análisis (desde el año más antiguo al más reciente).
- **Altitud**: refleja la ubicación vertical de los sitios sobre el nivel del mar, variable clave porque las condiciones climáticas varían con la elevación.
- **Año (CE)**: corresponde a la fecha de formación de cada anillo, proporcionando la dimensión temporal detallada de la serie.
- **δ13C VPDB:** representa la razón isotópica 13C/12C expresada en ‰ (milésimas) respecto al estándar Vienna Pee Dee Belemnite (VPDB). Esta variable es fundamental en el contexto de ISONET, ya que permite inferir condiciones climáticas pasadas —como temperatura, humedad relativa y características de precipitación— a partir de la señal isotópica en los anillos de los árboles.

En conjunto, estas variables no solo documentan la procedencia y características del material biológico recolectado, sino que también constituyen los insumos básicos para modelar y reconstruir regímenes climáticos de los últimos 400 años en Europa. De este modo, aportan un puente directo entre la información dendrocronológica y el contexto climático e histórico en que fueron capturados.

Posibles usos e interpretaciones de la base de datos: ¿qué tipo de conclusiones o hipótesis podrían extraerse a partir de su análisis?

Esta base de datos constituye un recurso valioso para la investigación interdisciplinaria, ya que ofrece series largas y consistentes de infomación ambiental con resolución anual. Además del estudio podemos concluir lo siguiente:

- **Variabilidad climática histórica:** que las series isotópicas reflejan periodos de mayor o menor disponibilidad hídrica y temperatura, lo que permite concluir que el clima europeo ha mostrado oscilaciones significativas antes de la era instrumental.
- **Eventos extremos:** que existen años o décadas con señales isotópicas anómalas ($\delta^{13}C$ o $\delta^{18}O$), lo que sustenta la hipótesis de sequías o veranos particularmente cálidos en ciertas regiones.
- **Cambio climático reciente:** que las tendencias en las últimas décadas presentan un desplazamiento respecto a la variabilidad natural reconstruida, respaldando la hipótesis de un efecto antropogénico creciente.

Decimos que el análisis permite no solo reconstruir el clima del pasado, sino también generar hipótesis sobre cómo los ecosistemas forestales responden a variaciones ambientales y cómo estos registros pueden usarse para contextualizar el cambio climático actual.

1.- Exploración inicial de los datos.

La base de datos utilizada proviene del proyecto **ISONET** "400 years of annual reconstructions of European Climate Variablility Using a Highly Resolved Isotopic Network", financiado por la Unión Europea dentro de su programa Marco para Energía, Medio Ambiente y Desarrollo Sostenible. El conjunto contiene series anuales de isótopos estables medidos en la celulosa de anillos de árboles correspondientes a 24 sitios distribuidos en Europa y regiones cercanas. Las series abarcan, en promedio, desde el año 1600 hasta 2003, para las razones isotópicas de carbono y oxígeno, y aproximadamente los últimos 100 años en el caso del hidrógeno.

Cada registro en la base corresponde a un par sitio-año con su respectiva medición isotópica. En términos generales, la estructura de la base se resume de la siguiente manera:

- **Número de sitios:** 24
- **Periodo temporal:** aproximadamente 1600-2003, con ciertas variaciones según el sitio.
- **Número de observaciones:** aproximadamente hay 10,790 datos aunque pueden ser menos debido a valores faltantes.
- **Número de variables:** 8 variables

Las variables que se presentan en la base de datos son las siguientes:

- **Site_code:** variable categórica nominal. Es un identificador abreviado (de 3 letras) que representa el sitio de muestreo. No tiene orden ni magnitud, solo distingue categorías.
- **Site_name:** variable categórica nominal. Indica el nombre completo del sitio de muestreo. Es una orden descriptiva sin orden ni jerárquica.
- **Country:** variable categórica nominal. Señala el país donde se encuentra el sitio de muestreo. Sirve como categoría geográfica general; no tiene orden natural.
- **Latitude:** variable numérica continua en escala de intervalo/razón. Indica la posición geográfica norte-sur del sitio, en grados decimales. Su rango está acotado entre -90 y +90.
- **Longitud:** variable numérica continua en escala de intervalo/razón. Indica la posición geográfica este-oeste del sitio, en grados decimales, con rango entre -180 y +180 (o 0-360 según sea conveniente).
- **Species:** variable categórica nominal. Representa la especie de árbol utilizada para la obtención de anillos, principalmente Quercus petraea, Pinus sylvestris y otras con relevancia local.
- **First year CE:** variable numérica entera en escala de intervalo. Indica el primer año del calendario común en el que se dispone de datos para un sitio específico.
- **Last year CE:** variable numérica entera en escala de intervalo. Indica el último año con datos disponibles para el sitio.
- **Elevation a.s.l. (metros sobre el nivel del mar):** variable numérica continua en escala de razón. Un valor cero tiene interpretación absoluta (nivel del mar), por lo que las comparaciones de razón son válidas.
- **Years with data:** variable numérica entera en escala de razón. Indica el número total de años con información efectiva en cada sitio. Un valor cero significa ausencia total de registros.

2.- Detección de problemas en los datos

Datos faltantes.

Los datos fueron segmentados en dos tablas. La información geográfica y los datos observados por región y año como se muestra continuación

site_code	site_name	Country	Latitude	Longitude	Species	First year CE	Last year CE	elevation a.s.l.	years_with_data
BRO	Bromarv	Finland	60	23.08	Quercus robur	1901	2002	5	102
CAV	Cavergno	Switzerland	46.35	8.6	Quercus petraea	1637	2002	900	366
CAZ	Cazorla	Spain	37.93	-2.97	Pinus nigra	1600	2002	1820	403
COL	Col Du Zad	Morocco	32.97	-5.07	Cedrus atlantica	1600	2000	2200	401
DRA	Dransfeld	Germany	51.51	9.78	Quercus petraea	1776	1999	320	224
FON	Fontainebleau	France	48.38	2.67	Quercus petraea	1600	2000	100	401
GUT	Gutuli	Norway	62	12.18	Pinus sylvestris	1600	2003	800	404
ILO	Sivakkovaara	Finland	62.98	31.27	Pinus sylvestris	1600	2002	200	403
INA	Inari	Finland	68.93	28.31	Pinus sylvestris	1600	2002	150	403
AHI	Perchtoldsdorf Wehrturm	Austria	48.25	16.77	Quercus petraea	1600	1883	n.s.	284
LAI	Lainzer Tiergarten	Austria	48.18	16.2	Quercus petraea	1812	2003	300	192
LIL	Pinar de Lillo	Spain	43.07	-5.25	Pinus sylvestris	1600	2002	1600	403
LOC	Lochwood	United Kingdom	55.27	-3.43	Quercus petraea	1749	2003	175	255
NIE1	Niepolomice	Poland	50.03	20.35	Quercus robur	1627	2003	190	377
NIE2	Niepolomice	Poland	50.03	20.35	Pinus sylvestris	1627	2003	190	377
PAN	Panemunes	Lithuania	54.09	23.96	Pinus sylvestris	1816	2002	45	187
PED	Pedraforca	Spain	42.23	1.7	Pinus uncinata	1600	2003	2120	404
POE	Poellau	Austria	47.31	15.81	Pinus nigra	1600	2002	500	403
REN	Renn	France	48.02	-1.83	Quercus robur	1611	1998	100	388
SER	Monte Pollino	Italy	39.93	16.21	Pinus leucodermis	1604	2003	1900	400
SUW	Suwalki	Poland	53.95	23.25	Pinus sylvestris	1600	2004	160	405
VIG	Vigera	Switzerland	46.05	8.77	Pinus sylvestris	1675	2003	1400	329
VIN	Vinuesa	Spain	42	2.75	Pinus uncinata	1850	1999	1950	150
WIN	Windsor	United Kingdom	51.43	-0.61	Pinus sylvestris	1763	2003	80	241
WOB	Woburn	United Kingdom	51.98	-0.59	Pinus sylvestris	1604	2003	10	400

Tabla con información geográfica.

site_code	Year	Value
AHI	1600	-23.9
AHI	1601	-23.9
AHI	1602	-23.6
AHI	1603	-23.8
AHI	1604	-23.9
...	...	...
WOB	2001	-25.3
WOB	2002	-25.2
WOB	2003	-25.1
WOB	2004	NaN
WOB	2005	NaN

Tabla con valores por región y año.

Como se aprecia en la primera tabla, no se identifican valores faltantes salvo en la variable de elevación para el sitio AHI, donde aparece registrado como n.s.. En este caso, no fue necesario realizar ninguna imputación.

Para validar el número de valores faltantes en las series, se consideró que todas las regiones cubren el periodo comprendido entre 1600 y 2005. A partir de la información cronológica se

generó la columna years_with_data, definida como: Last Year CE - First Year CE +1. Este valor representa la cantidad de observaciones esperadas por región. Posteriormente, se empleó el método `pandas.count()`, que excluye valores nulos, junto con un tratamiento de los caracteres `'NA'` y `'NAN'`, para cuantificar los datos efectivamente disponibles.

Los resultados obtenidos son los siguientes:

- **Cobertura completa (0 % de datos faltantes):**
  15 sitios no presentan vacíos en sus series (AHI, BRO, CAZ, ILO, INA, LAI, LOC, NIE1, NIE2, PAN, POE, SER, SUW, VIN y VIG), lo que indica que todas las observaciones anuales fueron registradas.

- **Sitios con datos faltantes mínimos (< 1 %):**
  CAV, GUT, LIL, PED y VIG presentan entre 1 y 4 años ausentes.

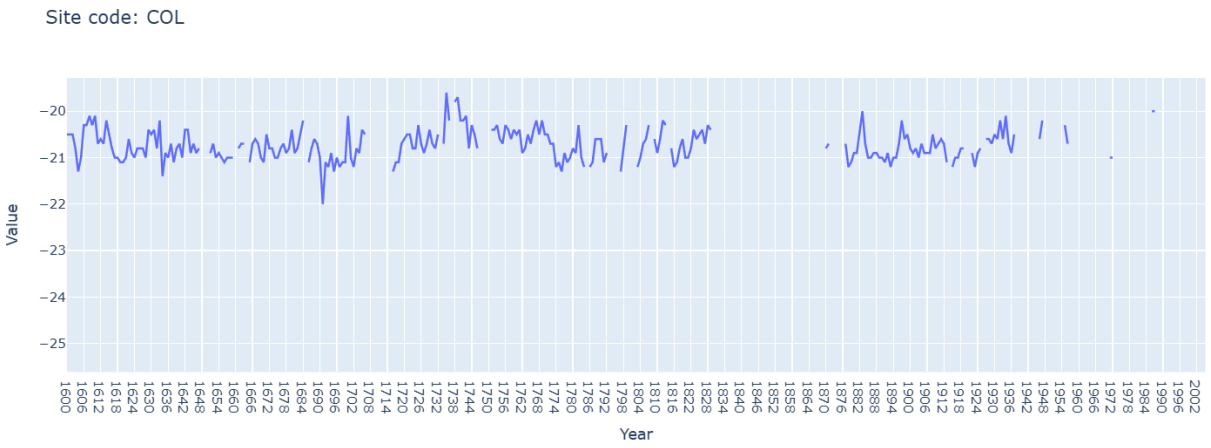
- **Sitios con datos faltantes moderados (1–5 %):**
  WOB (1.25 %) y WIN (3.73 %) muestran pérdidas pequeñas pero más notables.
  REN alcanza el 5.41 %, con 21 años de ausencia.

- **Sitios con porcentajes más altos (> 6 %):**
  COL (30.17 %) y FON (29.42 %) concentran alrededor de una tercera parte de datos faltantes.

- **Caso especial:**
  DRA presenta un valor de **-0.89 %**, lo que corresponde a un excedente de dos datos adicionales frente al número esperado. Esto sugiere un desfase en el conteo o duplicación de registros.

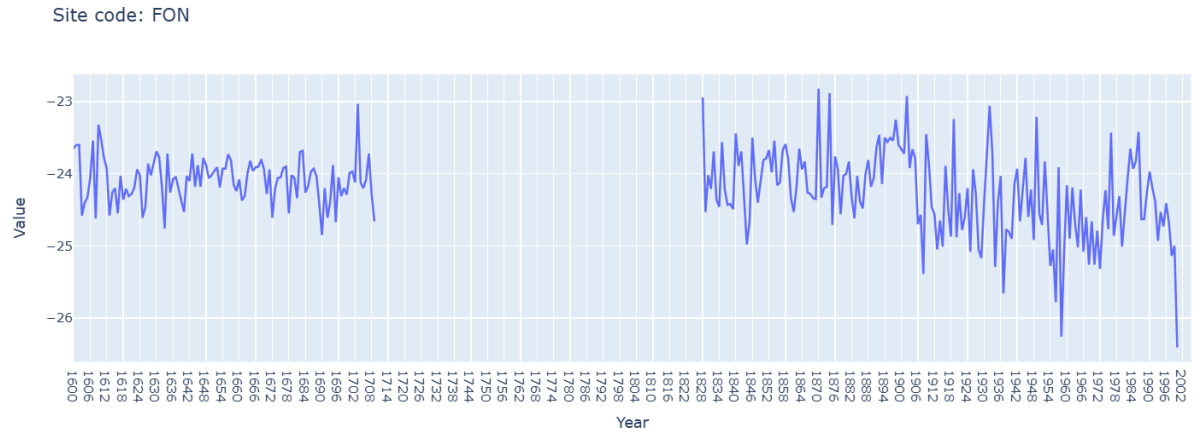
Respecto al tratamiento de los casos faltantes, se llevó a cabo una revisión visual:

- En **COL**, los datos ausentes están dispersos a lo largo de la serie, con mayor impacto entre 1820–1870 y en años recientes. Dado que este es el único caso con este patrón y la imputación sería demasiado compleja, se decidió excluirlo del análisis.

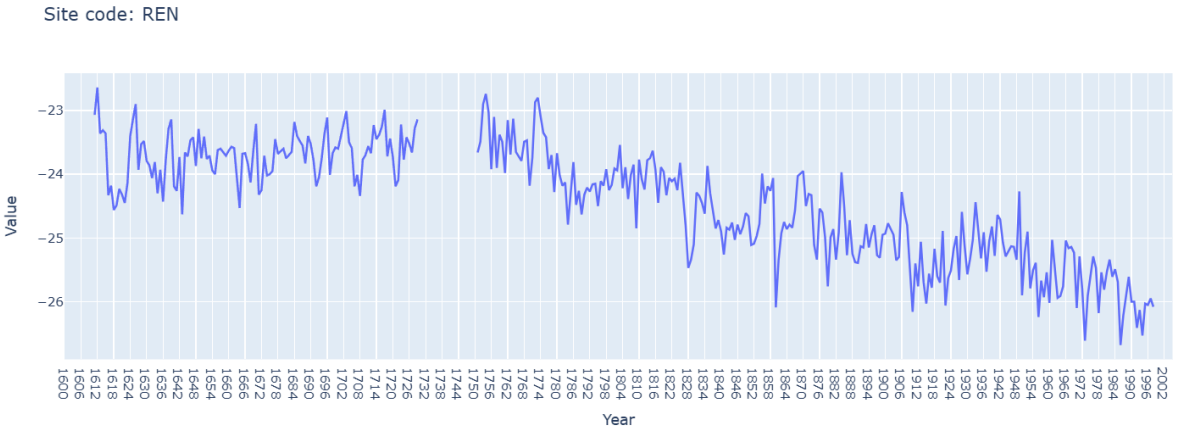


Serie de COL

- En **FON** y **REN**, las ausencias se concentran en una zona intermedia, pero con registros recientes completos. Por ello, las series se recortaron a partir de **1829** (FON) y **1751** (REN).



Serie de FON



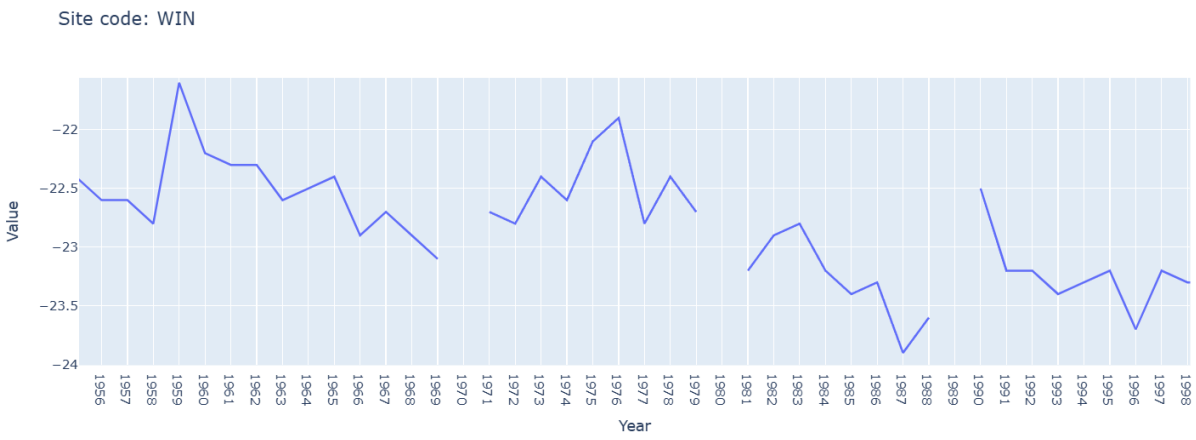
Serie de REN

- En los demás sitios con pérdidas marginales, las series se conservaron tal como están. En caso de requerir imputación futura, podrían aplicarse métodos sencillos de interpolación o sustitución por media.

Imputación de datos.

En la mayoría de los sitios, los valores faltantes se presentan de manera muy aislada y dispersa, con porcentajes bajos (≤ 6 %) y sin intervalos prolongados de ausencia. Esto se aprecia claramente en la serie de WIN, donde los huecos corresponden a pocos años y se encuentran intercalados entre datos completos, lo que permite una reconstrucción confiable mediante interpolación lineal.

Dado que el comportamiento de la serie es estable y las pérdidas son puntuales, la interpolación lineal se establece como el método principal para el tratamiento de los datos faltantes, asegurando continuidad temporal y evitando la introducción de oscilaciones artificiales.

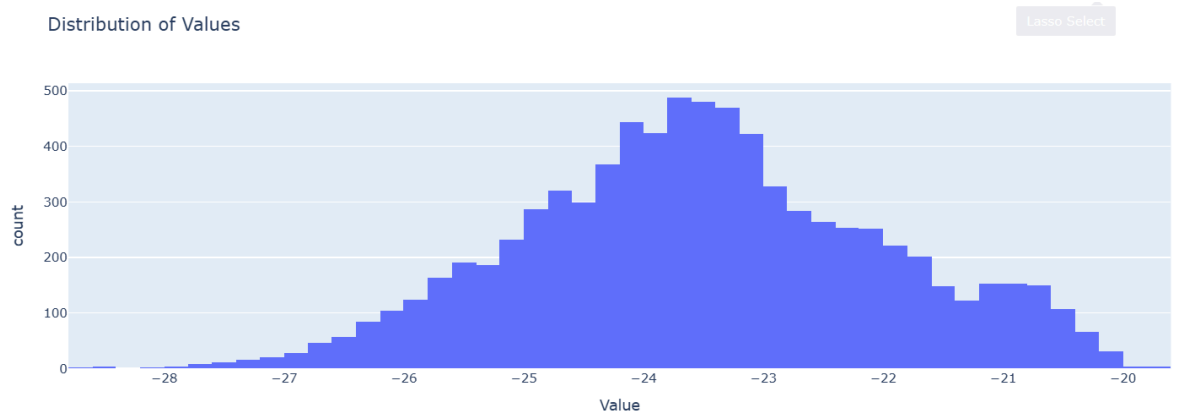


Ejemplificación de datos faltantes en WIN 1959-1998

De manera complementaria, en los casos en que la interpolación no pueda aplicarse (por ejemplo, cuando el dato faltante se encuentra al inicio o al final de la serie), se considera como alternativa la imputación por media regional, aprovechando la similitud climática y de crecimiento entre sitios cercanos. Sin embargo, esta opción es secundaria y no sustituye el criterio general de interpolación lineal.

Detección de valores atípicos

Para evaluar la presencia de outliers se analizó primero la distribución conjunta de todos los datos provenientes de las distintas regiones. El histograma global mostró un comportamiento cercano a una distribución normal, lo cual fue corroborado mediante la prueba de Anderson–Darling.



Histograma de datos de manera global.

El resultado de la prueba fue: A^2=9.2699 y parámetros de ajuste: μ=−23.55, σ=1.50.

La prueba indicó un ajuste exitoso de los datos a una distribución normal, lo que valida el uso de técnicas basadas en estandarización.

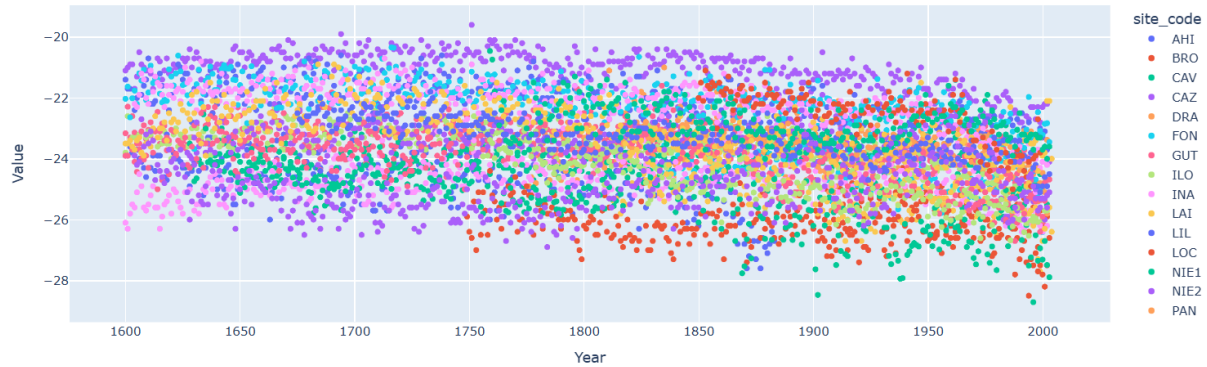
Posteriormente, se aplicó una transformación z-score a todas las observaciones, identificando como valores atípicos aquellos con un valor absoluto mayor a 3. Bajo este criterio, se detectaron los siguientes casos:

Índice	Sitio	Año	Valor original	z-score
5266	LOC	1994	-28.50	-3.309
5273	LOC	2001	-28.20	-3.109
5580	NIE1	1902	-28.47	-3.289
5674	NIE1	1996	-28.71	-3.450

De un total de 10,150 observaciones, únicamente 4 valores (0.039 %) fueron clasificados como outliers, lo que confirma que el conjunto de datos es en general consistente y sin anomalías graves.

Adicionalmente, se aplicó la metodología del hat matrix, utilizando como variable dependiente el año (Y) y como predictores las observaciones de los sitios. Este análisis no detectó valores atípicos adicionales, reforzando la conclusión de que la base de datos es robusta y que las pocas anomalías identificadas corresponden a casos muy puntuales.

Consideraciones sobre homogeneidad entre regiones y especies



Aunque el análisis global muestra un comportamiento general cercano a la normalidad y permite detectar valores extremos de forma conjunta, al desagregar la información por sitio, especie o región geográfica, se observan diferencias notables en los niveles y en la dinámica temporal de las series.

Esto implica que no existe un patrón claro y consistente que pueda ser explicado únicamente por especie o por localización geográfica. Dado que las variaciones parecen responder a múltiples factores simultáneos (condiciones locales, ambientales y de sitio), un análisis más detallado requeriría el uso de técnicas de agrupamiento (clusterización) que permitan identificar grupos de comportamiento similares sin imponer supuestos previos de homogeneidad.

Sin embargo, dicho enfoque queda fuera del alcance del presente proyecto, donde el objetivo principal es la exploración y limpieza de datos.

3.- Manejo de datos faltantes.

Como lo mencionamos anteriormente, a partir del análisis realizado a la variable years_with_data se clasificaron los siguientes valores faltantes:

- **Faltantes mínimos:** sitios con menos del 1% de los años sin datos.
- **Faltantes moderados:** sitios con entre 0% y 6% de valores ausentes, generalmente distribuidos de manera aislada.
- **Faltantes significativos:** sitios con más del 6% de los años sin datos, lo cual compromete la continuidad de las series.
- **Inconsistencias:** casos donde el número de registros supera el valor esperado, lo que sugiere errores de codificación o diferencias en el conteo.

Análisis de los datos faltantes.

Desde el punto de vista teórico, los mecanismos de ausencia se clasifican en tres categorías:

- MCAR (Missing Completely At Random)
- MAR (Missing At Random)
- MNAR (Missing Not At Random)

En nuestro caso, la evidencia sugiere que los datos faltantes se comportan como MCAR o MAR, ya que muchos se deben a limitaciones de predicción (por ejemplo, años sin muestra disponible) lo que difícilmente está relacionado con el valor real del isótopo. Sin embargo, no se descarta que algunos huecos prolongados puedan tener componentes MNAR (por ejemplo, pérdida selectiva de material en años extremos).

Discusión teórica

Bajo el supuesto de MCAR, sabemos que la eliminación de casos completos produce estimadores insesgados de la media poblacional, pero con mayor varianza, esto es,

$$\mathrm{Var}(\bar{Y}_{\mathrm{obs}}) = \frac{\sigma^2}{n_{\mathrm{obs}}} > \frac{\sigma^2}{n},$$

lo que implica pérdida de eficiencia. En el contexto de la base de datos, eliminar casos completos supondría reducir de forma drástica la longitud de las series, perdiendo información climática histórica.

Por lo tanto, aunque la eliminación es válida en teoría, se justifica emplear técnicas de imputación que aprovechen la estructura temporal y espacial de los datos.

Justificación de estrategia.

Del punto anterior, mencionamos que el manejo de datos faltantes en este proyecto se basa principalmente en la interpolación lineal, complementando por la imputación por media en algunos casos específicos. Esta estrategia nos asegura un tratamiento consistente con la naturaleza temporal de los anillos de árbol, preserva la continuidad de las series y minimiza el sesgo introducido, garantizando así mayor confiabilidad en las etapas posteriores de análisis y modelado.

4. Escalamiento de datos.

Codificación

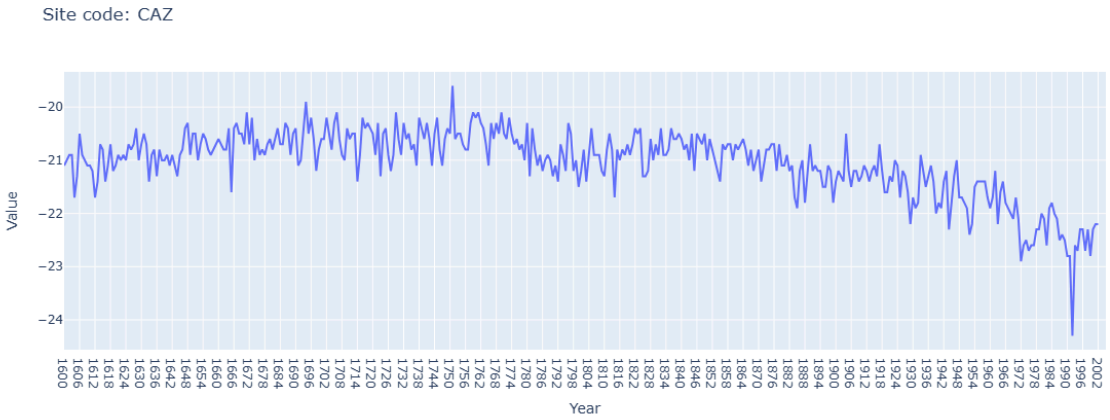
En este conjunto de datos hay dos variables categóricas para las cuáles podría ser útil codificarlas (de acuerdo al objetivo), las cuales son el sitio en el que se realizan las mediciones y la especie de la cual vienen los datos. Como vimos antes, los datos parecen comportarse de manera normal (globalmente), y en el gráfico de dispersión podríamos incluso pensar en una regresión lineal, pues los datos exhiben un comportamiento que asimila linealidad e incluso varianza constante. El caso en el que quisiéramos realizar una regresión lineal para el conjunto de datos global, una pregunta natural es cuánto peso se le adjudica a cada especie, y para ello podemos llevar a cabo una codificación one-hot para las especies y entonces realizar la regresión sobre estas variables y sus respectivas mediciones. La codificación one-hot para las especies se vería como sigue.

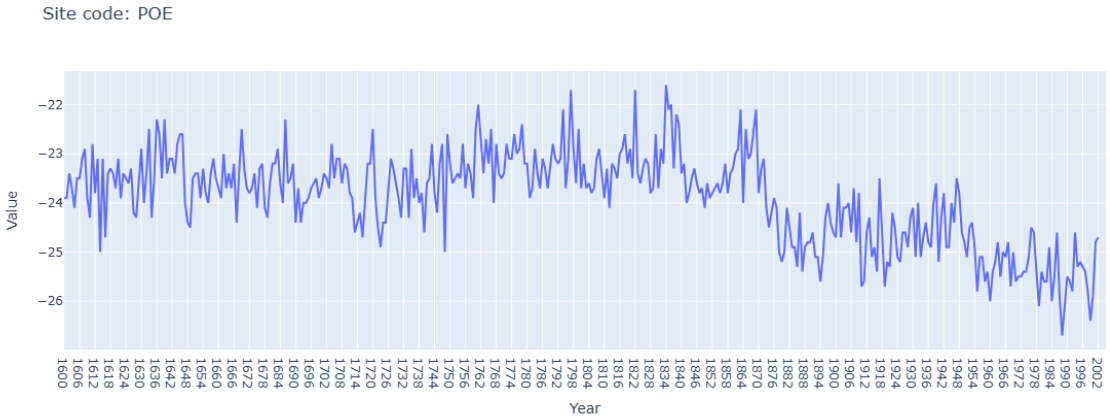
	site_code	Year	Value	standarized_value	Species	Species_Pinus leucodermis	Species_Pinus nigra	Species_Pinus sylvestris	Species_Pinus uncinata	Species_Quercus petraea	Species_Quercus robur
0	AHI	1600	-23.9	-0.178591	Quercus petraea	0	0	0	0	1	0
1	AHI	1601	-23.9	-0.178591	Quercus petraea	0	0	0	0	1	0
2	AHI	1602	-23.6	0.029841	Quercus petraea	0	0	0	0	1	0
3	AHI	1603	-23.8	-0.109113	Quercus petraea	0	0	0	0	1	0
4	AHI	1604	-23.9	-0.178591	Quercus petraea	0	0	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...
9739	WOB	2001	-25.3	-1.151272	Pinus sylvestris	0	0	1	0	0	0
9740	WOB	2002	-25.2	-1.081795	Pinus sylvestris	0	0	1	0	0	0
9741	WOB	2003	-25.1	-1.012317	Pinus sylvestris	0	0	1	0	0	0
9742	WOB	2004	NaN	NaN	Pinus sylvestris	0	0	1	0	0	0
9743	WOB	2005	NaN	NaN	Pinus sylvestris	0	0	1	0	0	0

Por supuesto, hay otro tipo de preguntas que surgen con respecto al sitio donde se toman las mediciones y podría realizarse el mismo tipo de codificación con estas variables.

Escalamiento

Para el escalamiento de datos consideramos los sitios CAZ y POE, en los cuales se realizaron mediciones a la especie Pinus Nigra, donde se cuenta con todas las observaciones entre los años 1600 y 2002. Estos conjuntos de datos se ven gráficamente como sigue.





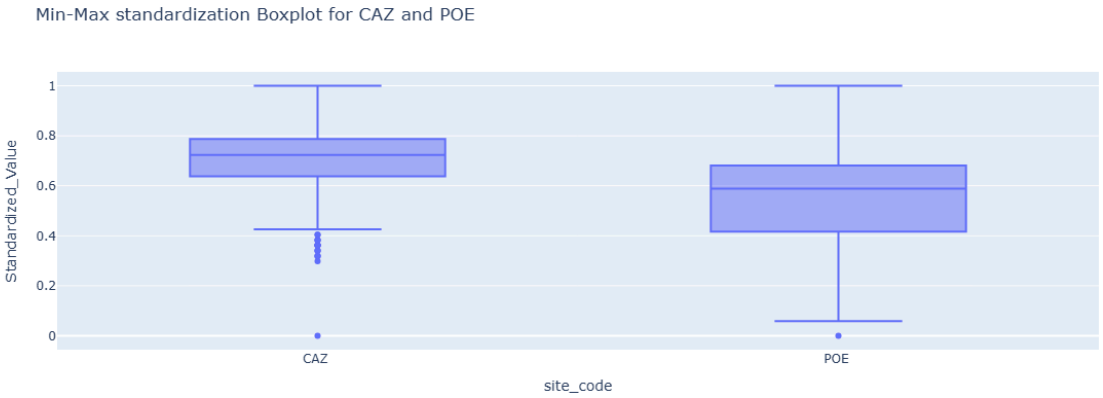
Una vez transformados nuestros datos obtenemos las tablas,

	site_code	Year	Value	min_max	z_score
0	CAZ	1600	-21.1	0.680851	-0.081777
1	CAZ	1601	-21.0	0.702128	0.080568
2	CAZ	1602	-20.9	0.723404	0.242913
3	CAZ	1603	-20.9	0.723404	0.242913
4	CAZ	1604	-21.7	0.553191	-1.055845
...	...	...	...	...	...
398	CAZ	1998	-22.3	0.425532	-2.029913
399	CAZ	1999	-22.8	0.319149	-2.841637
400	CAZ	2000	-22.3	0.425532	-2.029913
401	CAZ	2001	-22.2	0.446809	-1.867569
402	CAZ	2002	-22.2	0.446809	-1.867569

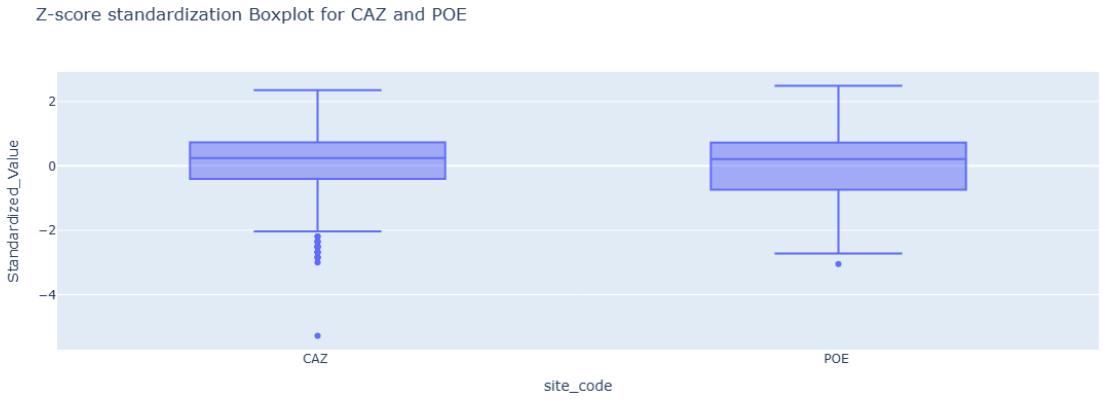
	site_code	Year	Value	min_max	z_score
0	POE	1600	-23.9	0.549020	-0.008077
1	POE	1601	-23.9	0.549020	-0.008077
2	POE	1602	-23.4	0.647059	0.534449
3	POE	1603	-23.7	0.588235	0.208933
4	POE	1604	-24.1	0.509804	-0.225088
...	...	...	...	...	...
398	POE	1998	-25.8	0.176471	-2.069676
399	POE	1999	-26.4	0.058824	-2.720707
400	POE	2000	-25.9	0.156863	-2.178181
401	POE	2001	-24.8	0.372549	-0.984624
402	POE	2002	-24.7	0.392157	-0.876119

en las cuales se encuentran tanto los datos originales como los transformados.

A continuación se presenta el boxplot de los datos bajo el escalamiento min-max, en el cual podemos observar que el sitio CAZ pareciera tener un valor atípicamente bajo, de manera que los datos se comprimen ligeramente alrededor del 1. Por otro lado, observamos una buena simetría en el boxplot del sitio POE.



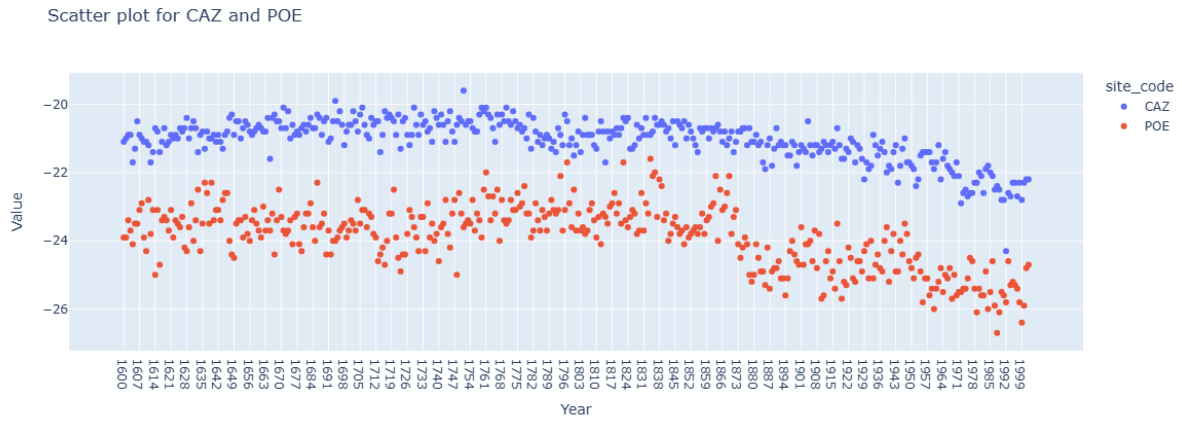
Bajo el escalamiento z-score observamos un comportamiento similar, en el cual puede verse que el sitio POE se mantiene simétrico alrededor del 0, mientras que el sitio CAZ se mantiene cargado hacia arriba. Esto se visualiza en el siguiente boxplot.



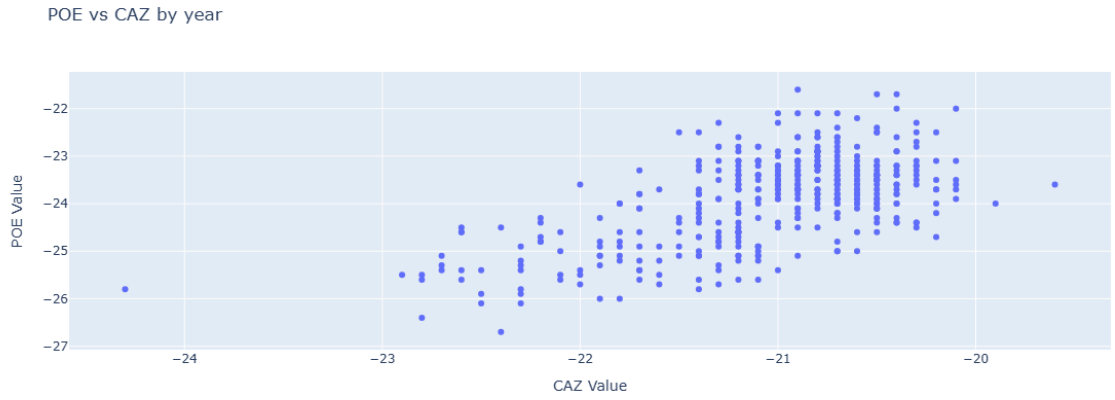
Lo anterior nos permite observar que a pesar de que los sitios CAZ y POE analizaron la misma especie (Pinus Nigra), sus datos se comportan un poco distintos. En los siguientes dos apartados profundizaremos un poco más en la comparación de los datos de CAZ y POE.

5. Visualización exploratoria

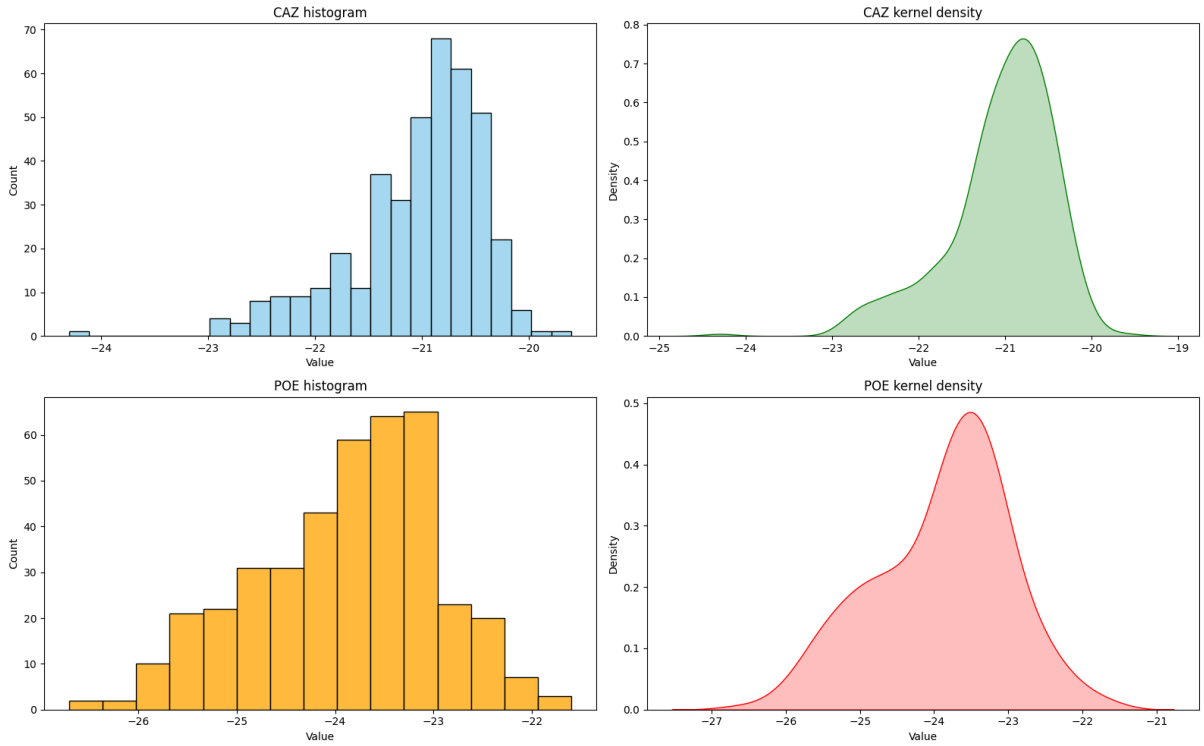
Primeramente presentamos un gráfico de dispersión para los datos de CAZ y POE con respecto al año de las mediciones. En esta imagen podemos ver aún más pronunciadamente el comportamiento distinto de estos datos, donde los datos de POE toman valores más pequeños que los de CAZ.



Asimismo, hacemos la gráfica de dispersión de los datos de CAZ contra los de POE, los cuales están pareados con respecto al año de las mediciones. Este gráfico nos muestra que ambos conjuntos de datos parecieran tomar valores mayores cuando el otro lo hace. Es decir, observamos que en años en los cuales el sitio CAZ tomó valores grandes, el sitio POE en general también, y recíprocamente para valores pequeños.



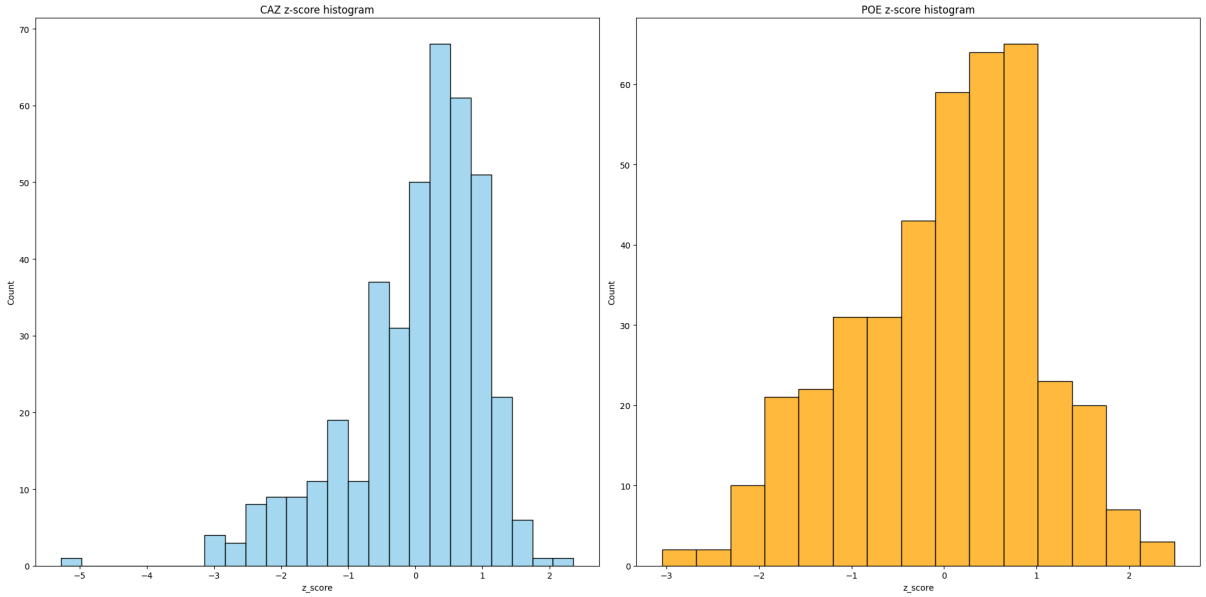
Finalmente, hacemos un diagnóstico de sus distribuciones a través de herramientas como el histograma y la densidad kernel (donde se utilizó kernel gaussiano).



Lo anterior nos confirma nuevamente que los sitios CAZ y POE se comportan de manera distinta; por ejemplo, podemos observar que los histogramas se encuentran localizados en distintas zonas del eje horizontal (o sea, tienen medias distintas). Además, podemos observar una mayor varianza en los datos POE.

El análisis anterior acerca de los datos de los sitios CAZ y POE resulta interesante pues contrasta el análisis que realizamos sobre los datos conjuntos. Anteriormente vimos que, conjuntamente, los datos recopilados se comportan de manera similar a una distribución normal. Sin embargo, un análisis por separado nos muestra que los datos de CAZ y POE no se comportan similar, y ni siquiera parecieran comportarse de manera gaussiana.

Finalmente, analizamos los datos escalados a través del z-score en busca de valores atípicos (outliers), donde identificamos aquellos datos que se encuentren a una distancia mayor a 3 del origen. En el siguiente gráfico podemos observar que todos los datos se comportan ciertamente bien, como para no ser considerados atípicos, a excepción de un valor de CAZ, el cual se encuentra alrededor de -5. Dicho dato es muy probablemente un valor atípico (con respecto a los datos de CAZ), y podría resultar importante revisar a qué se debe esa gran desviación de la media.



En este reporte nos concentramos en estudiar los datos de todos los sitios conjuntamente, pero es importante observar que dependiendo el objetivo del análisis, podría ser fundamental estudiar los datos de acuerdo a su sitio, o al menos en algún tipo de agrupación de ellos, ya que muy posiblemente las conclusiones cambiarían.

6.- Reflexión Crítica

Las decisiones de limpieza, imputación, codificación y escalamiento son críticas porque agregan supuestos al conjunto de datos. Según el método elegido, podemos introducir sesgos, alterar la varianza o modificar la representación de los predictores, lo cual influye directamente en la interpretación, robustez y generalización del modelo. Un ejemplo específico de nuestra información es la interpolación lineal que a pesar de que el histograma no cambia significativamente al realizar la prueba de Anderson Darling los parámetros estimados ahora son $\mu=-23.64$, $\sigma=1.44$. Esto ejemplifica la alteración de la varianza. Por otro lado, es claro con el análisis realizado que detectar outliers a nivel global usando una normal va a ser completamente distintos a las series mostradas a nivel sitio para la estimación de kernel.