Detección de anomalías en datos meteorológicos mediante métodos de análisis avanzados

Yamuna Devi Mena Ramirez

Proyecto Aplicado para optar al título de Magister en Ciencia de Datos

Director:
Antal Alexander Buss Molina

Asesores:
Ruby Viviana Ortiz Martínez
Juan Leonardo Moreno Rincón

Facultad de Ingeniería y Ciencias
Maestría en Ciencia de Datos
Pontificia Universidad Javeriana Cali
Enero 2025

Resumen

Dada la creciente incidencia de fenómenos climáticos , como ciclones, sequías e intensas lluvias, anticipar y estudiar los cambios en las condiciones atmosféricas se ha convertido en una prioridad para países como Colombia, que cuentan con amplias áreas costeras. Estos eventos representan no solo un riesgo significativo para el medio ambiente y la seguridad, sino que también exigen un entendimiento profundo de las dinámicas atmosféricas. Las series de tiempo meteorológicas son herramientas clave en este contexto, ya que permiten el monitoreo continuo de variables climáticas, como la temperatura, la presión, la humedad y la precipitación, facilitando la identificación y estudio de patrones y anomalías que podrían anticipar eventos climáticos.

En este contexto, se abordaron las limitaciones actuales en la detección de anomalías en los datos meteorológicos de la Dirección General Marítima en Colombia, siguiendo la metodología CRoss Industry Standard Process for Data Mining (CRISP-DM). Se propuso un enfoque híbrido que combina un algoritmo estadístico diseñado para la detección de anomalías naturalmente imposibles relacionadas con sensores, con un método más robusto que permite detectar días completos como eventos anómalos, en el que se seleccionaron las series multivariadas mediante un análisis de correlación, donde se identificaron las variables que presentaban mayor interdependencia. Luego, se aplicó el clustering utilizando los algoritmos K-means y DBSCAN, con enfoques tanto locales como globales. Los mejores resultados de evaluación se obtuvieron con el enfoque global aplicado a la serie multivariada que incluye temperatura del aire y humedad relativa, mostrando un puntaje de silueta de 0.67 y un índice de Davies Bouldin 0.54 para DBSCAN.

Agradecimientos

Agradezco a Dios, mi familia, mis amigos y a los profesores que me brindaron su apoyo.

Índice general

Índice de figuras

Índice de cuadros

1 Introducción

En la actualidad, los datos meteorológicos juegan un papel importante en la comprensión y predicción de los patrones climáticos, lo que facilita la toma de decisiones informadas y preparaciones oportunas frente a fenómenos extremos. Sin embargo, la gran cantidad y complejidad de estos datos representan un desafío significativo, ya que las series temporales pueden contener desviaciones sutiles que, aunque inicialmente pasen desapercibidas, pueden ser indicativas de anomalías importantes. Estas anomalías, si no se detectan a tiempo, podrían desencadenar efectos adversos tanto en el medioambiente como en la sociedad, desde pérdidas económicas hasta desastres naturales. La identificación temprana de estas señales permite tomar medidas preventivas que ayudan a mitigar los impactos negativos, garantizando la resiliencia y la sostenibilidad frente a fenómenos climáticos extremos. Además, las técnicas avanzadas de análisis permiten no solo detectar estas señales, sino también optimizar los procesos de gestión y vigilancia, lo que contribuye a una mejor comprensión del clima y facilita respuestas más efectivas ante las variaciones climáticas.

En Colombia, la Dirección General Marítima (DIMAR) es la entidad encargada de gestionar y monitorear datos meteorológicos y oceanográficos. No obstante, su proceso actual para identificar anomalías presenta limitaciones que dificultan la detección precisa de estos valores. Este proyecto se basó en dichas limitaciones para desarrollar un modelo de detección automática de anomalías para variables meteorológicas, con el objetivo de fortalecer la capacidad de análisis y toma de decisiones basada en datos confiables. Los objetivos específicos incluyeron la selección de variables relevantes, la implementación de técnicas de preprocesamiento para series temporales, y la evaluación de modelos robustos para la detección de anomalías. Como resultado, se lograron identificar tanto anomalías asociadas a fallas en sensores como dias asociados a eventos climáticos inusuales. Este enfoque se aplicó a datos meteorológicos del litoral Caribe colombiano, demostrando su potencial para optimizar el análisis de datos en contextos similares y respaldando tanto la investigación científica en esta región como la resiliencia ante fenómenos climáticos extremos [1].

2 Definición del problema

2.1 Planteamiento del problema

En la actualidad, la creciente incidencia de fenómenos climáticos extremos ejerce una presión constante sobre la resiliencia global, afectando de manera directa a las comunidades y ecosistemas en Colombia [3]. En este marco nacional, la DIMAR resalta la urgencia de comprender y anticipar los cambios en las condiciones atmosféricas [4]. Esta comprensión se vuelve esencial no sólo como un medio para mitigar riesgos, sino también como un componente crucial en la planificación efectiva frente a la variabilidad climática específica de nuestro país. En medio de estos desafíos, la recopilación masiva de datos meteorológicos y oceanográficos ha surgido como un recurso invaluable para la investigación científica y la toma de decisiones informada. Estos extensos conjuntos de información proporcionan una visión detallada de los sistemas climáticos y oceánicos, brindando conocimientos valiosos que hasta hace poco estaban más allá de nuestro alcance. Sin embargo, la riqueza de datos, aunque valiosa, también presenta nuevos desafíos, destacando la importancia de abordar estratégicamente la complejidad inherente a estos conjuntos de información. [5].

En el ámbito de los desafíos inherentes, destaca la identificación de valores atípicos como un componente crucial para mantener la integridad de extensas bases de datos. Este proceso brinda una retroalimentación valiosa, permitiendo anticipar posibles irregularidades en la recopilación de datos y, consecuentemente, mejorando la gestión y procesamiento de la extracción de información climática. La relevancia crítica de este desafío radica en que la presencia de valores anómalos en variables físicas podría señalar áreas de mejora en los procesos o la instrumentación de observación, además de indicar eventos climáticos.

En este contexto, la problemática que abordaremos se enfoca en las limitaciones del proceso actualmente empleado por la DIMAR para identificar valores atípicos en los datos meteorológicos y oceanográficos. A pesar de que este procedimiento lleva a cabo la identificación, su eficacia se ve comprometida debido a la falta de salidas claras que respalden la toma de decisiones y a las dificultades asociadas con la interpretación de los resultados. La falta de un enfoque integral, que incorpore la detección automática de valores atípicos y considere la correlación entre variables, junto con mecanismos que faciliten la interpretación de anomalías, se presenta como un desafío crítico. Esta limitación, al carecer de un proceso completo, conduce a análisis más demorados y genera incertidumbres, afectando la capacidad de abordar eventos climáticos y realizar evaluaciones precisas.

2.2 Formulación del problema

La pregunta de investigación que guiará este proyecto aplicado esla siguiente:

Pregunta principal

¿Cómo es posible desarrollar un algoritmo de detección automática de anomalías en datos meteorológicos mediante métodos de análisis avanzados para mejorar la toma de decisiones?

Preguntas secundarias

- ¿Cuáles son las variables meteorológicas más relevantes para la detección de anomalías?
- ¿Cuáles son las técnicas que resultan más eficientes en la limpieza y preparación de datos meteorológicos, especialmente cuando se trabaja con series temporales, con un enfoque destacado en la identificación de los tipos específicos de anomalías?
- ¿Qué modelos son los más adecuados para la identificación de anomalías en datos meteorológicos?
- ¿Cómo evaluar la efectividad de un sistema de identificación de anomalías en datos meteorológicos?

3 Objetivos del proyecto

3.1 Objetivo General

Desarrollar un modelo de detección automática de anomalías en datos meteorológicos mediante métodos de análisis avanzados para mejorar la toma de decisiones.

3.2 Objetivos Específicos

1. Determinar las variables meteorológicas más relevantes para la detección de anomalías y establecer el intervalo temporal de las bases de datos a emplear.

2. Aplicar técnicas de preprocesamiento de datos meteorológicos, específicamente diseñadas para el manejo eficiente de series temporales, haciendo énfasis en la identificación de los tipos de anomalías especificados por los expertos de dominio.

3. Identificar e implementar un modelo de detección de anomalías en función de su eficacia según métricas de rendimiento previamente definidas.

4. Evaluar el rendimiento del modelo de detección de anomalías mediante la comparación con conjuntos de muestras previamente evaluadas.

4 Marco de referencia

4.1 Marco teórico

En esta sección, se presentan términos que resultan importantes para la comprensión de este proyecto, abordando aspectos específicos relacionados con la detección de anomalías en datos meteorológicos.

4.1.1 Dirección General Marítima

La DIMAR es la Autoridad Marítima Nacional en Colombia y opera en una extensa área de 928.660 km2, que representa el 44.85 % del territorio nacional, abarcando 2.900 km de costa en el Pacífico y el Caribe. La DIMAR tiene la responsabilidad de dirigir, coordinar y controlar las actividades marítimas para garantizar su seguridad [4]. En su compromiso con la seguridad marítima, la DIMAR estableció el Servicio Meteorológico Marino Nacional con el objetivo de estructurar tareas relacionadas con la medición, recepción, uso y transformación de datos meteorológicos y del nivel del mar en el territorio marítimo colombiano. Su objetivo es proporcionar información que fortalezca la seguridad de la navegación, la vida humana en el mar y diversas actividades marítimas a lo largo de los litorales y zonas insulares del país. Para llevar a cabo estas funciones, la DIMAR cuenta con dos Centros de Investigación, el Centro de Investigaciones Oceanográficas e Hidrográficas del Caribe (CIOH) y del Pacífico (CCCP), respaldados por el Centro Colombiano de Datos Oceanográficos (CECOLDO). Este último tiene como objetivo administrar y gestionar los datos de oceanografía física, química y biológica, meteorología marina y geoquímica marina producidos por los centros de investigación y otras instituciones del país que contribuyen al conocimiento del componente marino [4].

Climatología del Caribe Colombiano

Las condiciones climáticas del Caribe colombiano están determinadas por la interacción de estructuras océano-atmosféricas en diferentes escalas temporales y espaciales. Entre los principales fenómenos se encuentran la Vaguada Monzónica o Zona de Confluencia Intertropical (ZCIT), frentes fríos, ondas tropicales del este y ciclones tropicales. Además, influyen sistemas de altas presiones, como el Anticiclón Subtropical del Atlántico Norte (NASH), y bajas presiones locales, como la Baja del Darién. Otros fenómenos relevantes son los procesos de brisa mar-tierra, la precipitación convectiva mesoescalar y la influencia de la piscina cálida

del Atlántico [6].

La región del Caribe colombiano se caracteriza por su estacionalidad, influenciada por estructuras oceano-atmosféricas y experimenta tres épocas climáticas principales: la seca, de diciembre a abril; la de transición, de mayo a junio; y la húmeda, de julio a noviembre. La época seca, entre diciembre y abril, está marcada por la intensa actividad del NASH, que fortalece los vientos Alisios del noreste y genera condiciones de estabilidad atmosférica, reduciendo los volúmenes de precipitación en gran parte del litoral central y norte. En la época de transición, de mayo a junio, el debilitamiento del Anticiclón permite el ascenso progresivo de humedad por la Vaguada Monzónica, dando inicio a la temporada ciclónica, mientras que las ondas tropicales del este y la baja presión del Darién influyen en la variabilidad de las precipitaciones. Finalmente, la época húmeda, de julio a noviembre, es la de mayor actividad meteorológica, con la confluencia de la Vaguada Monzónica, la ZCIT y ciclones tropicales, registrando el pico de la temporada ciclónica, con la presencia de ondas tropicales y, ocasionalmente, frentes fríos [1].

Control de calidad de datos

Actualmente, la Dimar realiza un procedimiento de pruebas de calidad a nivel primario para todas sus variables, el cual conduce a la asignación de banderas de calidad recomendadas por el International Oceanographic Data and Information Exchange (IODE). Para este nivel de calidad, se aplican tres tipos de (Banderas de calidad) QF, que se muestran en la siguiente tabla 4.1. Las pruebas de nivel primario acordadas para el presente caso se agrupan en dos: prueba de instrumentos y sensores, y prueba de coherencia [7].

Tabla 4.1: Banderas de calidad

Código	**QF**	**Significado**
2	Desconocida	No se cuenta con suficiente información para determinar la calidad del dato.
4	Malo	Los datos han fallado en una o más de las pruebas de control de calidad documentadas.
9	Dato ausente	Datos faltantes en la serie de datos.

Prueba de instrumentos y sensores La prueba de instrumentos y sensores se compone de dos pruebas intermedias que se describen a continuación:

- **Validación de sintaxis:** Esta prueba se utiliza para descartar los mensajes recibidos que contienen una estructura desconocida o que el valor del parámetro o fecha de muestreo contenga caracteres no válidos.

- **Agregación temporal:** Esta prueba evalúa que el dato medido se encuentre dentro del periodo o ventana de tiempo esperada, es decir, de acuerdo con la agregación temporal configurada para cada variable en el sistema.

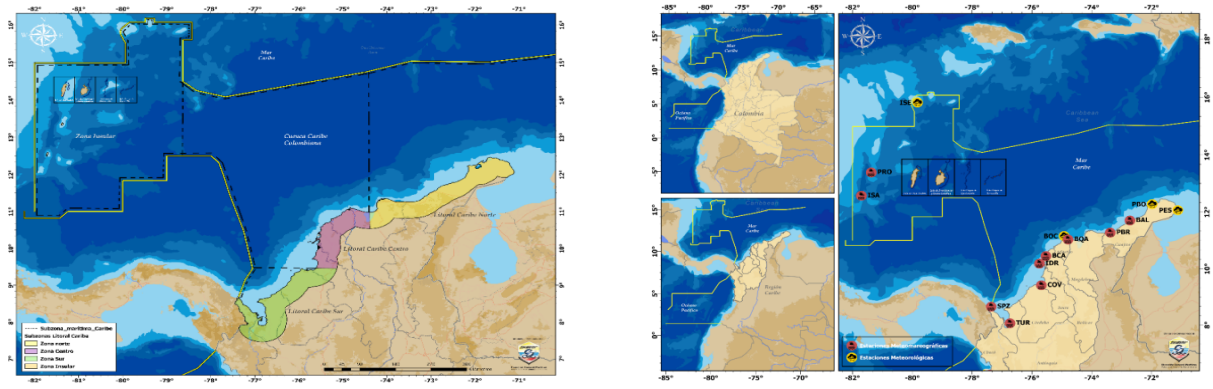
Prueba de coherencia La prueba de coherencia se compone de dos pruebas intermedias, a saber:

- **Datos faltantes:** En esta prueba se revisa toda la serie de datos en búsqueda de campos vacíos o de valores tales como [–10], [–05], [–15]. Estos últimos son equivalentes a datos faltantes. Una vez identificados estos campos, se llenan con el valor -99999 y se etiquetan con la bandera de calidad 9.
- **Valores imposibles:** Estos son valores naturalmente imposibles captados por los sensores.

  El control de calidad que la DIMAR aplica actualmente a todas sus variables se enfoca únicamente en asignar banderas de calidad, sin abordar de manera específica el estudio y la identificación detallada de anomalías asociadas a fenómenos climáticos. Esta carencia representa una limitación significativa y, al mismo tiempo, una oportunidad para avanzar en el desarrollo de metodologías que permitan identificar de manera preliminar datos o anomalías que requieran estudio adicional.

Red de Medición de Parámetros Oceanográficos y de Meteorología Marina (REDMPOMM)

La Dimar ha implementado la REDMPOMM con el objetivo de mejorar la observación y predicción de la variabilidad climática. Esta red está conformada por Boyas de Oleaje Direccional, Boyas Metoceanicas y Estaciones Meteomareográficas, distribuidas a lo largo del litoral Caribe y Pacífico colombiano, así como en el área insular. Estas estaciones tienen el propósito de medir variables como la temperatura del aire, la humedad relativa, la presión atmosférica, la precipitación, la radiación, la velocidad y dirección del viento, ya que proporcionan datos clave para el análisis del comportamiento climático en estas regiones, fundamentales para la vigilancia, predicción y toma de decisiones en el manejo de recursos naturales y la protección de la población ante fenómenos climáticos extremos. En este trabajo se utilizó la Estación Meteorológica Mareográfica Automática Satelital IDR, ubicada en el litoral Caribe central, con coordenadas de longitud 10.180 y latitud -75.750. Esta estación se encuentra en la Isla Naval, en la región caribeña central. La figura 4.1 a continuación muestra esta región y las estaciones pertenecientes a la REDMPOMM [1].



(a) Zonas del caribe colombiano

(b) Estaciones de la REDMPOMM

Figura 4.1: Región Caribe Colombiano tomada de [1]

4.1.2 Series de tiempo

Las series de tiempo se refieren a un conjunto de observaciones de una variable recopiladas secuencialmente a lo largo del tiempo. Estas observaciones pueden provenir de diversas fuentes, como tasas de interés semanales, precios de mercado diario o semanal, índices de precios mensuales, rendimientos anuales de cultivos, temperaturas diarias, precipitaciones mensuales, y ventas anuales de exportación. Las series temporales surgen de distintas áreas de aplicación, donde se busca obtener información a través del análisis de los datos para identificar patrones y comportamientos [8].

En el análisis de series temporales, se estudian las características del comportamiento de los datos, como las tendencias, la estacionalidad, los residuales y la presencia de anomalías. Un aspecto distintivo de las series temporales y sus modelos es que no se pueden asumir observaciones independientes provenientes de una misma población. Por lo tanto, un concepto clave en este tipo de análisis es que los modelos deben incorporar la dependencia entre los datos. Para llevar a cabo un análisis exitoso de series temporales, es necesario identificar modelos adecuados para el conjunto de observaciones, explorar el comportamiento de los datos y entender las diferentes componentes que lo componen [8].

Descomposición de series de tiempo

La descomposición de series temporales es una técnica fundamental que permite descomponer una serie temporal en sus componentes fundamentales: la tendencia T(t), la estacionalidad S(t) y la parte residual R(t). Esta metodología facilita la identificación de los patrones subyacentes y el análisis del comportamiento de los datos a lo largo del tiempo, proporcionando una visión más clara de las fuerzas que influyen en la serie. Al estudiar cada uno de estos componentes, es posible obtener información detallada que ayuda a interpretar los datos y a tomar decisiones fundamentadas basadas en patrones recurrentes [9]. A continuación, se explican cómo cada una de estas componentes contribuye a la interpretación de los datos:

Estacionariedad y tendencia Una serie temporal es estacionaria si sus propiedades estadísticas, como la media, la varianza y la autocorrelación, permanecen constantes a lo largo del tiempo. Esto implica que no hay tendencias significativas, y las fluctuaciones se mantienen alrededor de un nivel constante.

Estacionalidad La estacionalidad en una serie se refiere a patrones repetitivos y predecibles que ocurren en intervalos regulares de tiempo dentro de una serie temporal, como fluctuaciones anuales, mensuales o diarias. Estos patrones suelen estar asociados a factores cíclicos como estaciones del año, días de la semana o eventos periódicos.

Residual Son las diferencias entre los valores observados de una serie temporal y los valores esperados según un modelo. Representan el ruido o las fluctuaciones no explicadas por los componentes principales (tendencia y estacionalidad) y suelen analizarse para evaluar el ajuste del modelo.

Al aplicar la descomposición, se pueden distinguir dos enfoques principales: el multiplicativo y el aditivo. El enfoque multiplicativo se emplea cuando las componentes T(t) y S(t) interactúan entre sí, y las variaciones en la serie se expresan como el producto de estas componentes. Por otro lado, el enfoque aditivo se utiliza cuando la T(t) y la S(t) actúan de manera independiente, sumándose entre sí y a los residuos. La ecuación (4.1) representa el enfoque multiplicativo, mientras que la ecuación (4.2) corresponde al enfoque aditivo.

$$X(t) = T(t) \cdot S(t) \cdot R(t) \tag{4.1}$$

$$X(t) = T(t) + S(t) + R(t) \tag{4.2}$$

donde:

- $X(t)$: representa el valor observado en el tiempo t.
- $T(t)$: es la tendencia, que describe el cambio general a lo largo del tiempo.
- $S(t)$: es la estacionalidad, que describe las variaciones cíclicas en un patrón periódico.
- $R(t)$: son los residuos o el ruido, que representan las fluctuaciones no explicadas por la tendencia y la estacionalidad.

Tests estadísticos para validar la descomposición y estacionariedad

Dickey-Fuller Aumentada (ADF)

La prueba Dickey-Fuller Aumentada (ADF) es una herramienta estadística utilizada para verificar si una serie temporal es estacionaria, es decir, si sus propiedades estadísticas (como la media y la varianza) son constantes a lo largo del tiempo. La prueba es una extensión de la prueba Dickey-Fuller original, diseñada para manejar problemas de autocorrelación en los residuos al incluir términos de diferencia rezagados [10].

El nivel de significancia es importante en la interpretación de los resultados de la prueba ADF. Este nivel (comúnmente 1 %, 5 % o 10 %) se utiliza para determinar si se puede rechazar la hipótesis nula. En este contexto:

- **Hipótesis nula (H_0)**: La serie tiene una raíz unitaria, lo que indica que no es estacionaria.
- **Hipótesis alternativa (H_1)**: La serie es estacionaria.

El resultado de la prueba se basa en comparar el estadístico de prueba calculado con los valores críticos específicos para el nivel de significancia elegido:

- Si el estadístico de prueba es menor que el valor crítico, se rechaza H_0, indicando que la serie es estacionaria.
- Si el estadístico de prueba no es menor que el valor crítico, no se rechaza H_0, lo que sugiere que la serie no es estacionaria.

Test de Ljung-Box
El test de Ljung-Box es un tipo de prueba estadística que permite determinar si alguna de las autocorrelaciones de un grupo de una serie temporal es distinta de cero. En lugar de probar la aleatoriedad en cada rezago distinto, prueba la aleatoriedad "general.[en] función de una serie de rezagos y, por lo tanto, es una prueba de combinación [10].

La prueba de Ljung–Box puede definirse como:

- H_0: Los datos no están correlacionados (es decir, las correlaciones en la población de la que se toma la muestra son 0, de modo que cualquier correlación observada en los datos resulta de la aleatoriedad del proceso de muestreo).
- H_a: Los datos muestran correlación serial.

La estadística de prueba es:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \tag{4.3}$$

donde n es el tamaño de la muestra, $\hat{\rho}_k$ es la autocorrelación de la muestra en el retardo k, y h es el número de retardos que se están probando. Bajo H_0, la estadística Q sigue asintóticamente una distribución $\chi^2(h)$. Para un nivel de significancia α, se rechaza H_0 si Q excede el valor crítico correspondiente de la distribución $\chi^2(h)$.

Test de Jarque-Bera En estadística, la prueba de Jarque-Bera es una prueba de bondad de ajuste para comprobar si una muestra de datos tiene la asimetría y la curtosis de una distribución normal [10].

La prueba estadística JB se define como:

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \tag{4.4}$$

Donde n es el número de observaciones (o grados de libertad en general); S es la asimetría de la muestra, K la curtosis de la muestra:

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}} \tag{4.5}$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{2}} \tag{4.6}$$

Donde $\hat{\mu}_3$ y $\hat{\mu}_4$ son las estimaciones de los momentos centrales tercer y cuarto, respectivamente, $\bar{x}$ es la media de la muestra y $\hat{\sigma}^2$ es la estimación del segundo momento central, la varianza.

El estadístico de Jarque-Bera se distribuye asintóticamente como una distribución chi cuadrado con dos grados de libertad y puede usarse para probar la hipótesis nula de que los datos pertenecen a una distribución normal. La hipótesis nula es una hipótesis conjunta de que la asimetría y el exceso de curtosis son nulos (asimetría = 0 y curtosis = 3). La prueba se puede utilizar en Modelos de Regresión para probar la hipótesis de Normalidad de los residuos. Para ello se utilizan los residuos estimados obtenidos por mínimos cuadrados. Los puntos críticos para muestras pequeñas se pueden calcular vía Monte Carlo.

4.1.3 Anomalías

Las anomalías se refieren a cualquier desviación del comportamiento normal en una serie temporal o conjunto de datos. Estas desviaciones pueden manifestarse de diversas formas, como anomalias contextuales, donde un punto de datos se desvía en función del contexto seleccionado, anomalias colectivas, donde un subconjunto completo de datos se aleja del patrón general, y anomalias puntuales o globales, donde un solo punto se desvía considerablemente del conjunto [11]. En el contexto climático, las anomalias pueden manifestarse como fenómenos extremos, como la presencia de ciclones, frentes fríos, períodos de sequía prolongados, o lluvias intensas. Estos eventos son desviaciones significativas del comportamiento climático normal y suelen reflejar cambios drásticos o leves en las condiciones atmosféricas.

Por otro lado, las fallas en los sensores tambien son una de las principales causas de anomalías en los datos recopilados, ya que generan resultados inesperados que pueden distorsionar el comportamiento esperado en las mediciones [12]. A continuación se presentan algunos de los tipos más comunes:

- Los spikes o picos se manifiestan como secuencias de puntos consecutivos que muestran un cambio brusco en los valores medidos. Estas anomalías rara vez aportan información útil, por lo que deben ser descartadas, lo que resulta en una pérdida temporal de datos del sensor. Las causas más frecuentes de estos spikes incluyen problemas de batería o desconexiones intermitentes en el hardware del sensor.

- Los outliers o valores atípicos son puntos de datos aislados que se desvían considerablemente del comportamiento general del conjunto de mediciones. Estos outliers pueden surgir de manera aleatoria e instantánea, y si no se detectan y eliminan, pueden sesgar la media, la varianza y otros parámetros estadísticos, afectando así la precisión del modelo. Las causas de estos valores atípicos suelen ser desconocidas, aunque en algunos casos pueden identificarse mediante el análisis del software del registrador de datos.

- Las stuck-at faults se caracterizan por series de valores que permanecen con variación casi nula durante períodos prolongados. Este tipo de falla conduce a la pérdida de información sobre las fluctuaciones reales de los datos, lo que puede justificar su descarte si no se ajustan al comportamiento esperado del sensor. Las causas más comunes suelen estar relacionadas con fallas en el hardware del sensor, problemas en la conexión o batería baja.

Detección de anomalías

La detección de anomalías en series de tiempo consiste en identificar patrones o puntos en los datos temporales que no se ajustan a las dinámicas o tendencias establecidas como normales. Este proceso utiliza técnicas analíticas y computacionales para monitorear las variaciones a lo largo del tiempo, considerando la estructura secuencial y las características propias de la serie, como estacionalidad, tendencias y ruido. Su relevancia radica en aplicaciones donde es importante detectar cambios inesperados, como en la predicción de fallos, la supervisión de sistemas críticos, el estudo de eventos climáticos y el análisis de comportamiento en datos temporales [13].

El proceso para identificar que técnica de detección de anomalias aplicar depende de varios aspectos que se muestran a continuación en la figura 4.2,las casillas en naranja muestran los enfoques que se tuvieron en cuenta en este trabajo.

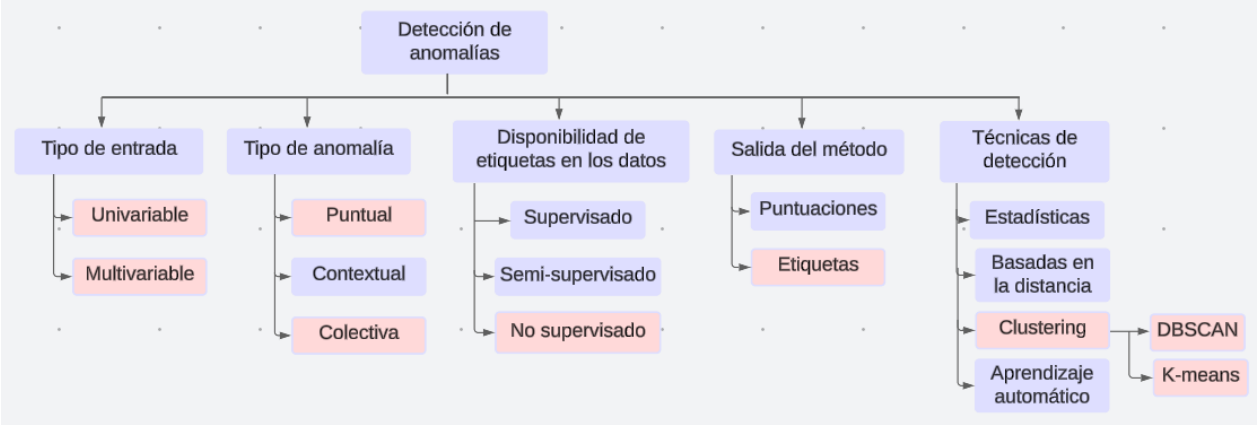


Figura 4.2: Aspectos importantes para la detección de anomalías

4.1.4 Clusterización

La clusterización es una técnica fundamental dentro del aprendizaje no supervisado que agrupa un conjunto de objetos de tal manera que los objetos dentro de un mismo grupo (o clúster) son más similares entre sí que a los de otros grupos. Esta técnica es particularmente valiosa en la detección de anomalías en series de tiempo, donde el objetivo es identificar patrones inusuales que pueden indicar eventos inesperados o fallos en los sistemas monitoreados. En este contexto, los algoritmos de clusterización permiten agrupar series de tiempo de que exhiben comportamientos similares, facilitando la identificación de puntos de datos que se desvían significativamente del comportamiento normal [14].

Cuando hablamos de clusterización en series de tiempo, no necesariamnete se pueden usar los mismos métodos que se aplican en datos transversales, donde se agrupan puntos individuales en un espacio fijo. En las series de tiempo, los datos no se organizan de manera independiente, sino en función del tiempo, lo que implica que los datos se ordenan de forma continua a lo largo del tiempo. Por lo tanto, en lugar de agrupar puntos aislados, lo que se busca es dividir la serie temporal en pequeños períodos o series, también conocidos como pasos temporales, que permiten capturar cómo evolucionan los datos a lo largo del tiempo. De esta manera, no se agruparán simplemente puntos aislados, sino series completas de tiempo que muestran la progresión de los datos. Utilizando métodos de clusterización o técnicas jerárquicas, se agruparán estas series en función de las medidas de similitud. De esta manera, cada clúster representará diferentes formas o patrones que toman los datos a lo largo del tiempo.

A continuación, se visualiza una serie dividida en tres series mas pequeñas4.3.

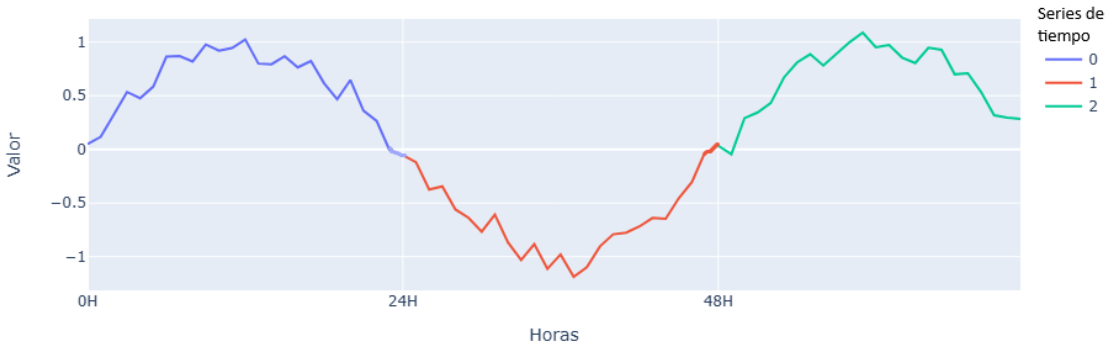


Figura 4.3: Serie de tiempo dividida

En la figura 4.3, se puede observar cómo la serie temporal ha sido segmentada en bloques de 24 datos (horas) cada uno, donde cada serie corresponde a un día específico. A partir de este proceso, al aplicar la clusterización, estas series de tiempo se agruparán en clusters según la similitud entre ellas y su comportamiento a lo largo del tiempo.

Medidas de similitud en la clusterización de series de tiempo

En la agrupación de series temporales, las medidas de similitud/disimilitud juegan un papel fundamental, ya que determinan la forma en que se mide la cercanía entre las series. Estas métricas son esenciales porque influyen directamente en la asignación de los puntos a los diferentes clústeres. Al calcular la distancia entre las series, es posible identificar patrones y agrupaciones basadas en la similitud temporal, lo que permite descubrir estructuras ocultas en los datos. Las medidas como la distancia de Hausdorff, Dynamic Time Warping, la distancia euclidiana, la deformación temporal dinámica, entre otras, son utilizadas para comparar series con distintos intervalos de muestreo y longitud. Cada una de estas métricas se adapta a diferentes características y propiedades de las series, como la traslación, el escalamiento o las derivadas temporales [15].

Dynamic Time Warping (DTW). La distancia DTW es una métrica que mide la similitud entre dos series temporales deformando el eje del tiempo de manera no lineal. Esta distancia permite capturar patrones temporales más allá de eventos simultáneos, lo que la diferencia de la distancia euclidiana, que solo compara puntos con marcas de tiempo coincidentes. DTW es particularmente útil cuando las series tienen diferentes longitudes y deben alinearse temporalmente para una comparación precisa. Esta medida es capaz de identificar similitudes entre series que, aunque pueden estar desplazadas en el tiempo, poseen estructuras subyacentes similares. Por lo tanto, DTW es una herramienta poderosa para el agrupamiento de series temporales, superando las limitaciones de otras métricas tradicionales [15].

Distancia euclidiana. La distancia euclidiana es una medida sencilla y fácil de entender que se utiliza ampliamente para calcular la diferencia entre puntos en un espacio

n-dimensional. Es útil cuando se desea evaluar la proximidad directa entre dos puntos basándose en sus coordenadas. Gracias a su simplicidad, la distancia euclidiana a menudo produce resultados rápidos y comprensibles, especialmente cuando se trabaja con datos que tienen una estructura lineal y cuando no es necesario considerar desplazamientos temporales. En algunos casos, ha demostrado ser efectiva para identificar grupos bien definidos, proporcionando una buena base para la clasificación inicial en problemas de clustering.

K - means

El algoritmo K-means es una técnica de agrupamiento utilizada en el aprendizaje no supervisado, donde el objetivo es dividir un conjunto de datos no etiquetados en un número predefinido de clusters. En este algoritmo, se selecciona el número K de clusters que se desea formar, y cada cluster está representado por un centroide, que es el punto medio o promedio de los datos en ese grupo. Los puntos de datos se asignan a su centroide más cercano, y el objetivo es minimizar la suma de las distancias entre los puntos y sus respectivos centroides.

El proceso iterativo del K-means comienza con la elección aleatoria de K puntos como centroides iniciales. Luego, cada punto de datos se asigna al centroide más cercano, y los centroides se recalculan como el punto medio de los datos en cada grupo. Este proceso se repite hasta que los centroides convergen, es decir, cuando ya no hay cambios significativos en la distribución de los datos entre los clusters. K-means es ampliamente utilizado en aplicaciones como segmentación de clientes, sistemas de recomendación, análisis de imagen, y reducción de dimensionalidad, ya que permite identificar patrones ocultos y descubrir grupos en los datos sin necesidad de etiquetas predefinidas[16].

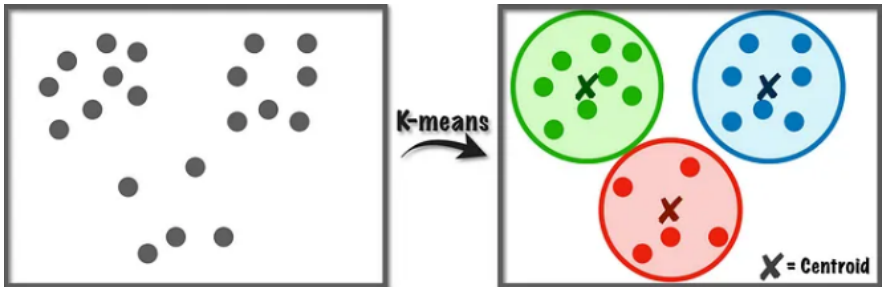


Figura 4.4: Clusterización con K-means

DBSCAN- Density-Based Spatial Clustering of Applications with Noise

Es un algoritmo no paramétrico de agrupamiento basado en densidad : dado un conjunto de puntos en algún espacio, agrupa los puntos que están muy agrupados (puntos con muchos vecinos cercanos) y marca como valores atípicos los puntos que se encuentran solos en regiones de baja densidad (aquellos cuyos vecinos más cercanos están demasiado lejos). DBSCAN utiliza dos parámetros principales: Número mínimo de puntos (minPts), que representa el número mínimo de puntos necesarios para que una región se considere densa, y distancia de vecindad (eps), que es la medida de distancia utilizada para determinar los puntos en la

vecindad de cualquier punto. DBSCAN se enfoca en dos conceptos fundamentales: alcance de densidad y conectividad de densidad. El alcance de densidad se refiere a la accesibilidad de un punto desde otro, indicando qué tan densamente se puede alcanzar un grupo, dependiendo de la distancia (eps) definida. Por su parte, la conectividad de densidad se basa en un enfoque de encadenamiento transitorio, donde los puntos se conectan a través de sus vecinos [17].

El proceso de agrupamiento en DBSCAN comienza seleccionando un punto de datos aleatorio que aún no ha sido visitado. A partir de este punto, se evalúa su vecindad dentro de una distancia eps. Si al menos minPts puntos se encuentran en esta vecindad, se inicia el agrupamiento. Todos los puntos dentro de esta distancia eps se agrupan bajo el mismo clúster, y este procedimiento se repite iterativamente para cada nuevo punto añadido, expandiendo el clúster hacia sus vecinos cercanos. Cuando se completa un clúster, el algoritmo retoma la búsqueda con un nuevo punto no visitado, permitiendo descubrir otros posibles grupos o identificar puntos como ruido. Este proceso continúa hasta que todos los puntos han sido visitados, clasificados y etiquetados como parte de un clúster o como ruido.

Evaluación en algoritmos de clusterización

La evaluación del desempeño de algoritmos de clusterización es importante para determinar la efectividad de las agrupaciones obtenidas. A diferencia de los métodos supervisados, donde se cuenta con etiquetas para comparar, la clusterización no supervisada presenta un desafío en este aspecto. Por lo tanto, se utilizan diversas métricas que permiten evaluar la calidad de los clústeres formados. Entre estas métricas, una de las más reconocidas es el coeficiente de silueta.

Puntaje de silueta. El puntaje de Silhouette es una métrica que se utiliza para evaluar la calidad de los resultados de agrupamiento de datos. Esta puntuación se calcula midiendo la similitud de cada punto de datos con el cluster al que pertenece y en qué medida se diferencia de otros clusters. Se define como:

$$s(i) = \frac{b(i) - a(i)}{\text{máx}(a(i), b(i))} \tag{4.7}$$

Donde:
- $a(i)$ es la distancia promedio entre el punto i y todos los demás puntos en el mismo clúster. - $b(i)$ es la distancia promedio entre el punto i y todos los puntos en el clúster más cercano.

El Coeficiente de Silueta mide la calidad de los agrupamientos al evaluar que tan bien o mal están emparejados los puntos dentro de su clúster en comparación con los clústeres vecinos. Varía entre -1 y $+1$, donde valores cercanos a $+1$ indican que los puntos están bien separados dentro de su clúster, mientras que valores cercanos a 1 sugieren un mal emparejamiento debido a la cercanía con otros clústeres. Si el coeficiente es superior o igual a 0.7 se considera “fuerte“, entre 0.5 y 0.7 “razonable“, y por debajo de 0.5 “débil“ [18].

Índice Davies-Bouldin. El Índice Davies-Bouldin es una métrica utilizada para evaluar la calidad de los agrupamientos (o clústeres) en análisis de datos. Este índice mide la dispersión interna de los clústeres y la separación entre ellos, buscando encontrar el equilibrio entre compactitud y separación. Se calcula como la media de las relaciones entre la dispersión interna de cada clúster y la distancia media al clúster más cercano [19].

$$DB(C) = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{d(C_i, C_j)} \right) \tag{4.8}$$

En donde:

$\mathcal{C}_i$: Centro del clúster i

$\mathcal{S}_i$: Desviación media de los puntos dentro del clúster i

$d(\mathcal{C}_i, \mathcal{C}_j)$: Distancia entre los centros de los clústeres i y j

k: Número total de clústeres.

En donde un Un índice bajo (menor a uno), indica que los clústeres están bien separados y compactos, lo que resulta en una mejor calidad del agrupamiento y un índice alto (mayor a uno) sugiere que los clústeres están solapados o mal separados, lo que indica una mala calidad en el agrupamiento.

4.1.5 Antecedentes

En esta sección, se revisan estudios relevantes que abordan aspectos relacionados con nuestra línea de investigación, analizando enfoques destacados para contextualizar esta propuesta en la detección de anomalías en datos meteorológicos.

En [20], se presenta un marco novedoso para la detección de anomalías en datos de series temporales relacionados con parámetros físicos y biogeoquímicos en entornos marinos. El enfoque se centra en la concentración de clorofila-a, temperatura de la superficie del mar y concentración de oxígeno disuelto. El marco propuesto utiliza un modelo de predicción SARIMA (Seasonal AutoRegressive Integrated Moving Average) para realizar pronósticos y detectar anomalías. Se realiza un análisis exploratorio preliminar, seguido de un preprocesamiento de las series temporales para abordar valores faltantes y outliers. La metodología incluye la estimación y evaluación del modelo, seguido de la fase de pronóstico y evaluación. Finalmente, se lleva a cabo la detección de anomalías, clasificándolas en niveles de gravedad, lo que resulta imporatnte para evaluar el impacto ambiental y apoyar la gestión de áreas marinas protegidas y granjas acuícolas. La contribución principal radica en la capacidad de clasificar la gravedad de las anomalías y su aplicabilidad en la evaluación del impacto de actividades humanas en entornos costeros.

En el estudio mostrado en [21], los investigadores desarrollaron RADIS, un marco de trabajo innovador que combina una red neuronal variacional autoencoder de memoria a corto y largo plazo (VAE-LSTM) con técnicas de preprocesamiento de datos, como la identificación de estados operativos normales a través de la generación de imágenes de series temporales y análisis de componentes conectados. Aplicaron este enfoque a datos de sensores de una planta de generación diésel en un buque cisterna, evaluando el rendimiento de la detección de anomalías mediante la inyección de ruido simulado en los datos. Los resultados demostraron la capacidad de RADIS para detectar y cuantificar anomalías en datos de maquinaria marina, destacando su aplicabilidad potencial en el monitoreo de equipos críticos en entornos marítimos. Aunque identificaron limitaciones y desafíos, como la dependencia de la calidad de los datos, sugieren futuras investigaciones, incluida la optimización continua, métodos de explicabilidad y la exploración de factores climáticos y de rendimiento para mejorar aún más este marco de trabajo.

En [22] los autores desarrollaron un enfoque innovador utilizando un modelo de predicción Wavelet Neural Network (WNN). y las variables de estudio fueron la salinidad superficial (SS) y la velocidad de corriente superficial (SCP). Se centraron en dos estrategias de detección: la Estrategia de Observación (OS) y la Estrategia de Predicción (PS). Descubrieron que PS fue eficaz para identificar anomalías persistentes causadas por fallas en equipos, mientras que OS fue más adecuada para anomalías ocasionales relacionadas con factores naturales. Se discutió la importancia del nivel de clasificación, y se encontró que un nivel del 99 proporciona un equilibrio razonable entre falsos negativos y falsos positivos. La confiabilidad del método se evaluó mediante curvas ROC, comparando el rendimiento del modelo WNN con modelos tradicionales de Redes Neuronales Artificiales (ANN). Los resultados indicaron que el WNN superó a los otros modelos en precisión y confiabilidad.

En el contexto de los antecedentes, estos estudios han abordado problemáticas similares a la que planteamos en este proyecto. El primer estudio se enfoca en la evaluación del impacto ambiental y la gestión de áreas marinas protegidas y granjas acuícolas, centrándose en la identificación de valores atípicos para tres variables, entre ellas: biogenicas y fisicas. Por otro lado, el segundo estudio se centra en la detección de anomalías en maquinaria marina, ofreciendo una perspectiva diferente al considerar aspectos relacionados con equipos marinos. El tercer estudio se asemeja más a esta propuesta al identificar anomalías debidas a fallas en equipos y ocasionadas por factores naturales, aunque se limita a dos variables físicas.

A diferencia de estudios previos, esta propuesta combina un enfoque univariable y multivariable con un método híbrido basado en un algoritmo estadístico, seguido por un algoritmo de clusterización. Este enfoque permite no solo la detección de anomalías causadas por sensores,sino también la identificación de anomalias asociadas a eventos climáticos.Además, se enfoca específicamente en el litoral Caribe colombiano, adaptándose a las particularidades y características únicas de esta región.

5 Metodología

Para llevar a cabo este proyecto, se implementó la metodología CRISP-DM (CRoss Industry Standard Process for Data Mining), desarrollada por Wirth y Hipp en [2]. Este enfoque ofrece un marco estructurado y flexible compuesto por seis fases clave: comprensión del problema, entendimiento de los datos, preparación de los datos, modelado, evaluación e implementación. Esta metodología sirvió como guía para organizar y ejecutar cada etapa del análisis, facilitando la correcta aplicación de los procedimientos analíticos y asegurando la obtención de resultados confiables. Como se muestra en la Figura 5.1, el ciclo correspondiente ilustra cómo se estructuró y ejecutó el análisis.

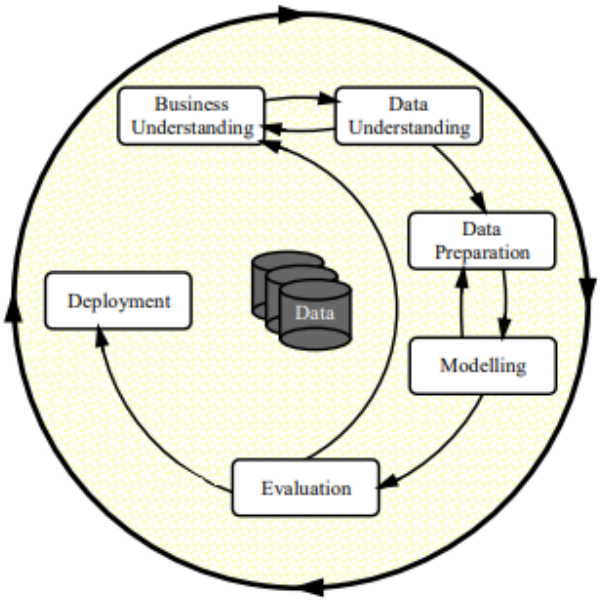


Figura 5.1: Fases del Modelo del Proceso CRISP-DM [2]

5.1 Entendimiento de los datos

Esta sección está alineada con el objetivo específico 1, ya que busca identificar las variables meteorológicas clave y determinar el intervalo temporal de las bases de datos a utilizar. Para ello, se realizó un análisis preliminar de los datos disponibles, evaluando su calidad y características principales, lo que permitió definir las variables más relevantes para el estudio de anomalías y seleccionar el periodo adecuado para el análisis.

5.1.1 Recolección de los datos

Se analizaron bases de datos meteorológicas que incluyeron varias variables, tales como la temperatura del aire, la humedad relativa, la presión atmosférica, la radiación, la velocidad y dirección del viento, y la precipitación. Estos datos fueron recolectados por la Estación IDR, que forma parte de la RedMpomm y se encuentra ubicada en Isla Naval, como se mostró en la figura 4.1.

Las bases de datos fueron proporcionados por el Cecoldo en formato CSV y cubrieron el periodo de 2014 a 2022, con la excepción de la serie de radiación, que abarcó desde 2018 hasta 2023. Cada una de las bases representaba un año, sumando un total de nueve bases por variable. Los registros presentaron diferentes frecuencias: cuatro series tenían intervalos de una hora, mientras que tres series estaban registradas cada 10 minutos. Cada base incluyó los siguientes componentes:

- ID Estación: Código alfanumérico de 10 dígitos que identifica la estación.
- Cordenadas: Latitud y longitud en formato numérico de 7 dígitos.
- Fecha: Formato MM/DD/YYYY que indica la fecha de registro.
- Hora: Formato HH:MM:SS que señala la hora de registro.
- Valor: Valor medido por el sensor.
- Bandera (QF): Código numérico de 1 dígito que indica la calidad de los datos.

5.1.2 Análisis exploratorio de los datos (EDA)

El EDA desempeñó un papel importante en este proyecto al facilitar la identificación de patrones y anomalías, así como la comprensión de los datos. Este proceso fue importante para guiar las decisiones de preprocesamiento y validar los resultados obtenidos en etapas posteriores.

Para realizar el EDA, fue necesario llevar a cabo algunos ajustes estructurales en las bases de datos. Estas modificaciones aseguraron la coherencia de la información. A continuación, se describen los pasos implementados:

1. Unificación de Nombres de Columnas: Se revisaron y estandarizaron los nombres de las columnas para cada variable y año, facilitando su integración eficiente.

2. Verificación de Marca de Tiempo: Se aseguró que la marca de tiempo fuera consistente a lo largo de todo el intervalo de datos en cada base.

3. Corrección Tipos de Datos: Se verificó que las columnas de los valores de los sensores fueran de tipo numérico y las columnas de fecha y hora se ajustaron al tipo DateTime para mejorar su manipulación y análisis.

4. Correción de duplicados: Se identificaron y eliminaron los registros duplicados.

5. Integración de Bases de Datos: Se procedió a unir todas las bases de datos correspondientes a una misma variable meteorológica, consolidando la información de manera coherente y organizada.

Despues de realizar estos ajustes a las bases de datos se procedió a relizar el EDA

Visualización de las series de tiempo

Para continuar con el análisis, se visualizó cada serie de tiempo con el objetivo de comprender su comportamiento a lo largo del tiempo, figura 5.2 . En estas visualizaciones, el eje vertical representa los valores medidos para cada variable meteorológica, mientras que el eje horizontal corresponde al componente temporal.

Se presentaron dos versiones de las gráficas: una representación estándar y otra con un enfoque de "zoom". Esta doble visualización se empleó para manejar el efecto de los valores extremos, que pueden distorsionar la percepción del patrón general de la serie. En la primera figura 5.2, se incluyen todos los valores, incluyendo los extremos o comportamientos anormales, los cuales se marcaron con circulos negros. Sin embargo, en la segunda versión, Figura 5.3, se ajusta la escala del eje vertical para enfocarse en el comportamiento más general de la serie, excluyendo los valores extremos.

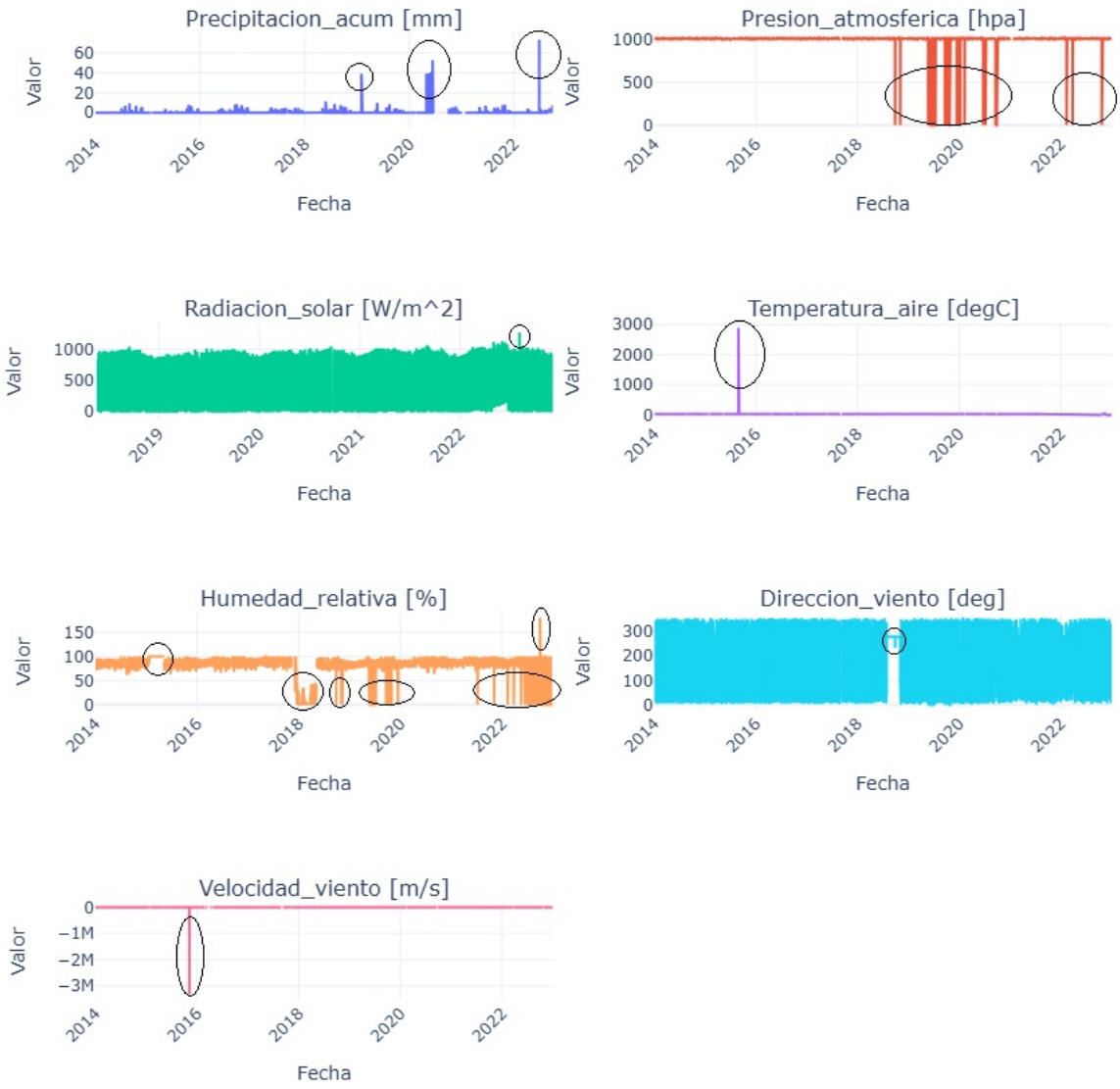


Figura 5.2: Series Temporales de las variables meteorologicas.

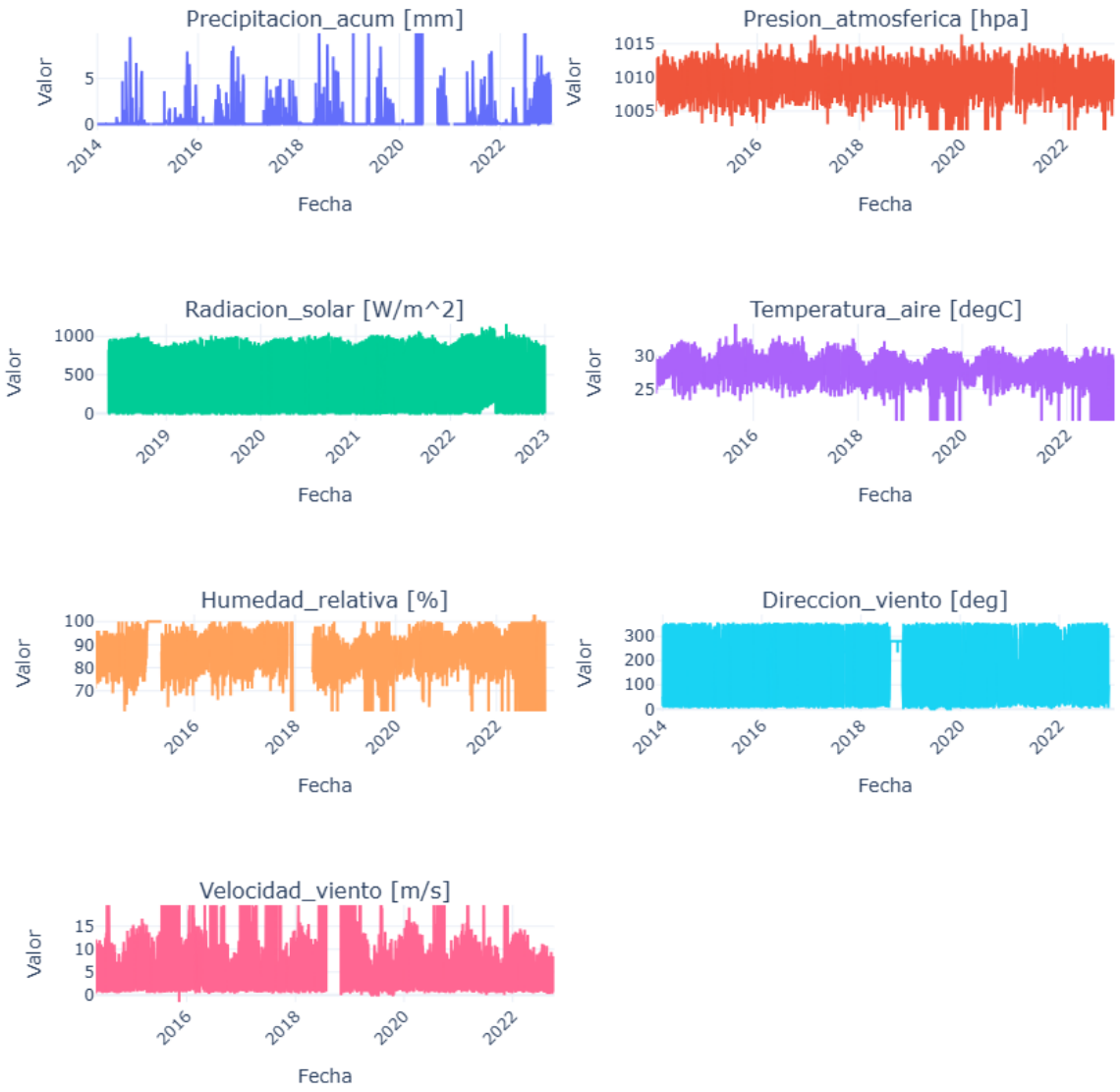


Figura 5.3: Visualización ajustada de las series de tiempo: Comportamiento sin valores extremos

A partir de las representaciones visuales de las series, se evidenciaron dos aspectos clave:

Valores extremos: En la figura 5.2, se observaron valores atípicos extremos que se apartan significativamente del comportamiento general de la serie temporal. Estos valores fueron considerados anomalías debido a que no se ajustan a el patrón esperado de la serie, lo que dificulta una interpretación precisa del comportamiento normal de los datos. Un ejemplo claro de estos valores extremos se observó en el gráfico de temperatura en la Figura 5.2, donde alrededor del año 2016 hay un valor cercano a los 3000 grados Celsius, un valor completamente incompatible con la física de la variable y con la distribución esperada. Este tipo de anomalías fue fácilmente identificable visualmente en la mayoría de las variables analizadas. Para asegurar la identificación y evaluación de estas anomalías, se complementó el análisis con gráficos de boxplots como los que se muestran en la Figura5.4.

Con el objetivo de comparar todas las variables en un mismo gráfico, se utilizaron boxplots para visualizar las distribuciones e identificar las anomalías . Sin embargo, las diferencias en la magnitud de las variables dificultaban una comparación directa. Para abordar este problema, se realizó un proceso de normalización de los datos, que consistió en escalar las variables a una misma unidad de medida. Una vez que los datos fueron normalizados, los boxplots confirmaron la presencia de anomalías en la mayoría de las variables, con la excepción de la dirección del viento.

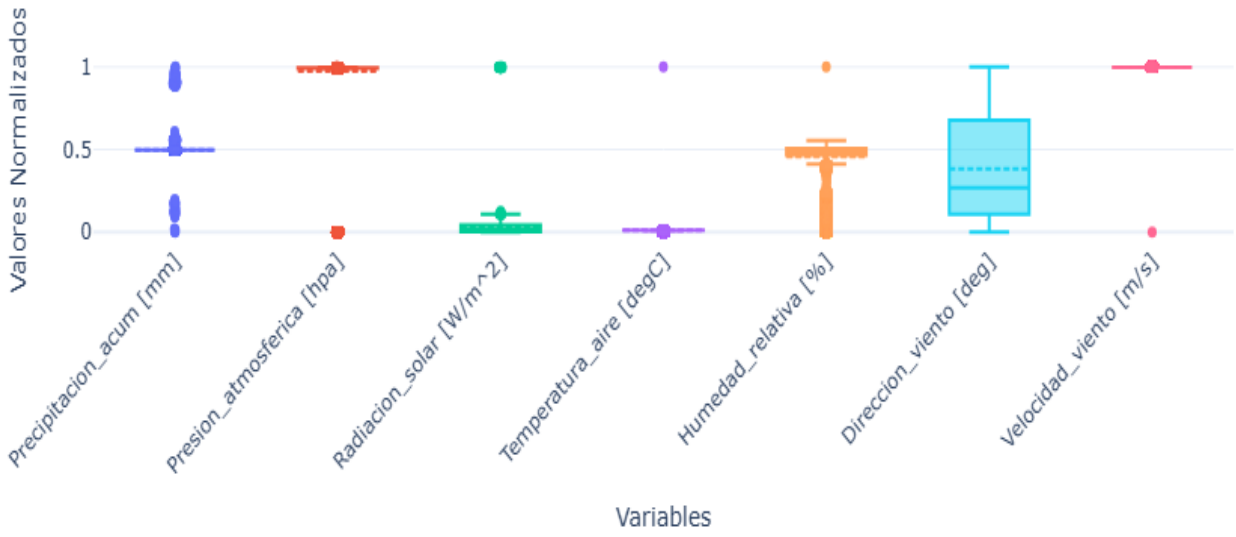


Figura 5.4: Análisis de anomalías en Variables Meteorológicas Normalizadas mediante Boxplots

Espacios vacíos (valores NaN): Se detectó la presencia de espacios vacíos en las gráficas, correspondientes a valores NaN (Not a Number) o datos faltantes. Estos huecos en la serie temporal indicaron la ausencia de información en determinadas observaciones, y fueron claramente visibles en los gráficos. Un ejemplo de estos espacios vacíos se muestra en la figura 5.5, donde se puede observar la falta de datos en septiembre de 2017 y nuevamente en noviembre del mismo año. Además, se aprecia que estos huecos suelen presentarse en forma de conglomerados de valores perdidos, es decir, secuencias de NaNs consecutivos. Un comportamiento anómalo también es visible en el período comprendido entre enero y mayo de 2018, donde los datos presentan irregularidades notables.

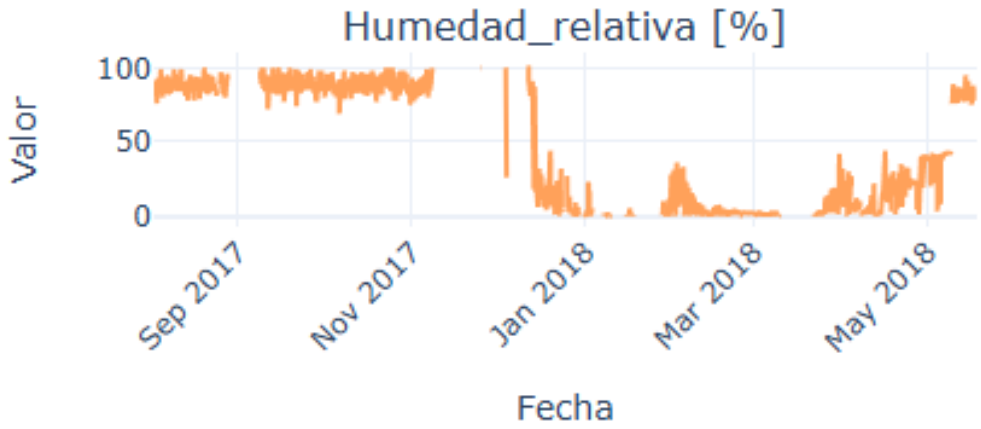


Figura 5.5: Valores perdidos Nan

Para visualizar mejor los datos perdidos Nan, se realizó un análisis más detallado, representando el porcentaje de datos faltantes para cada variable mediante el gráfico de barras en la figura 5.6, lo que permitió visualizar de manera clara la distribución de los valores ausentes en el conjunto de datos.

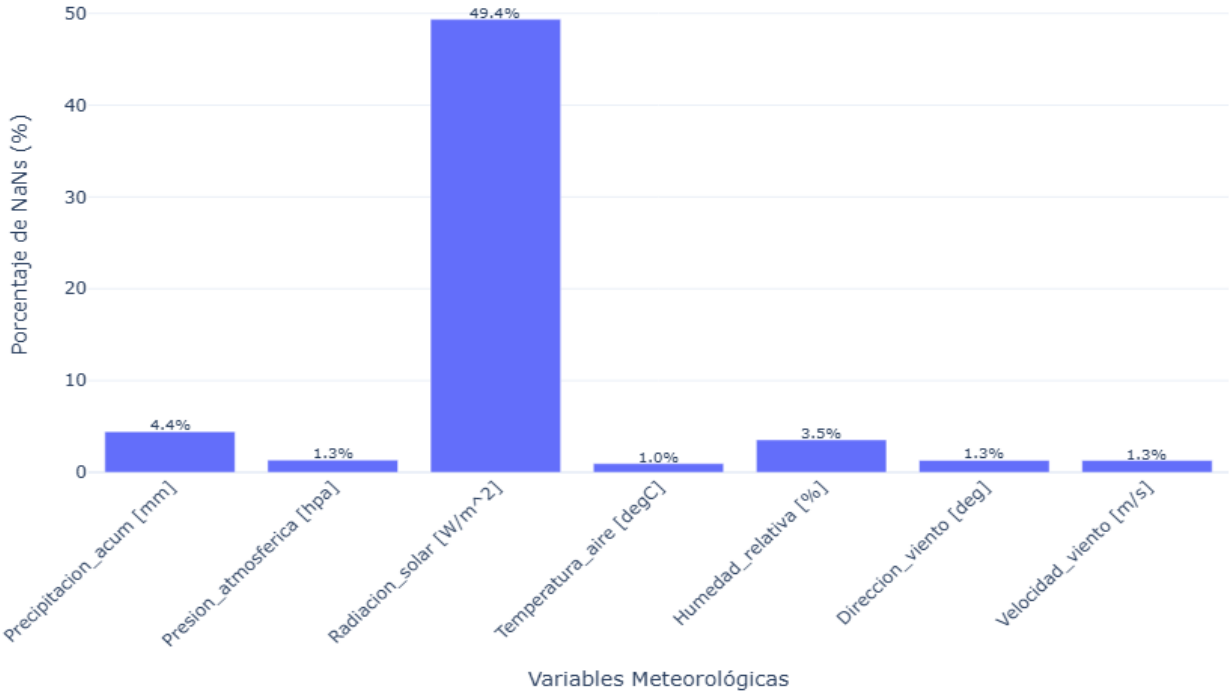


Figura 5.6: Gráfico de barras datos Nan

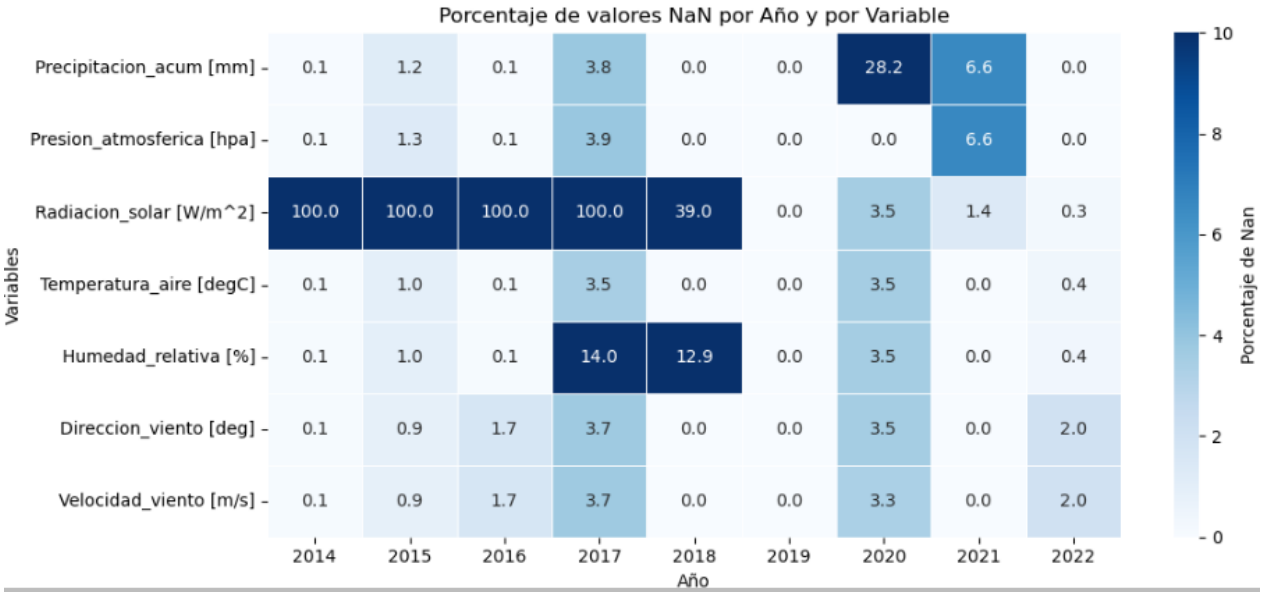


Figura 5.7: Mapa de calor datos Nan por año

El análisis realizado a través del gráfico de barras mostró que la variable Radiación presenta un porcentaje considerable de datos faltantes, alcanzando aproximadamente el 49 % de los registros, lo que representa casi la mitad de los valores. Este hallazgo fue confirmado por la visualización de la serie temporal de Radiación en la figura 5.2, donde se puede observar que los datos disponibles para esta variable comienzan a partir de 2018. En comparación con Radiación, las demás variables también presentaron valores NaN, aunque en su mayoría estos fueron inferiores al 5 %. Además, el mapa de calor mostrado en la figura 5.7 ilustra el porcentaje anual de datos faltantes para cada variable. En él, se destacan los años 2017 y 2018 como los más críticos para la variable Humedad, y el año 2020 para Precipitación, donde los datos faltantes superan el 10 %. Cabe mencionar que para la variable Radiación, los años entre 2014 y 2017 presentaron un 100 % de datos faltantes, lo que corrobora el análisis anterior para esta variable.

Estacionalidad y Estacionariedad

Para el análisis de la estacionalidad y la estacionariedad de las variables, primero se utilizó la Prueba de Dickey-Fuller Aumentada (ADF) para determinar si las series eran estacionarias o no. Posteriormente, se procedió con la descomposición de cada serie temporal utilizando un enfoque aditivo. Este enfoque es adecuado cuando las componentes de la serie, como la tendencia, la estacionalidad y los residuos, se combinan de manera lineal. Al analizar las series, se observó que las fluctuaciones estacionales no aumentan en proporción a los valores de la serie, lo que justifica la selección del modelo aditivo como el más adecuado para describir este comportamiento.

Antes de aplicar la descomposición, se optó por realizar un remuestreo de los datos. Para ello, se utilizó la media de los valores en intervalos diarios y mensuales, lo que permitió suavizar las series y reducir el ruido, destacando así las verdaderas tendencias y patrones estacionales. Al comparar ambos enfoques, primero con los datos diarios y luego con los mensuales, se observó que la estacionalidad anual se reflejaba de manera más clara y precisa en los valores mensuales, lo que hizo que este intervalo fuera el más adecuado para la descomposición de cada serie. Después de la descomposición, se aplicaron dos pruebas fundamentales. La prueba Jarque-Bera, que permitió evaluar si los residuos seguían una distribución normal, y el test de Ljung-Box, que ayudó a evaluar la autocorrelación de los residuos. Estas pruebas permitieron validar la calidad de la descomposición al asegurar que los residuos fueran independientes y se ajustaran adecuadamente a una distribución normal. A continuación, se muestra la descomposición de la humedad en la figura 5.8. Las descomposiciones correspondientes a las demás variables se encuentran detalladas en el Anexo A.

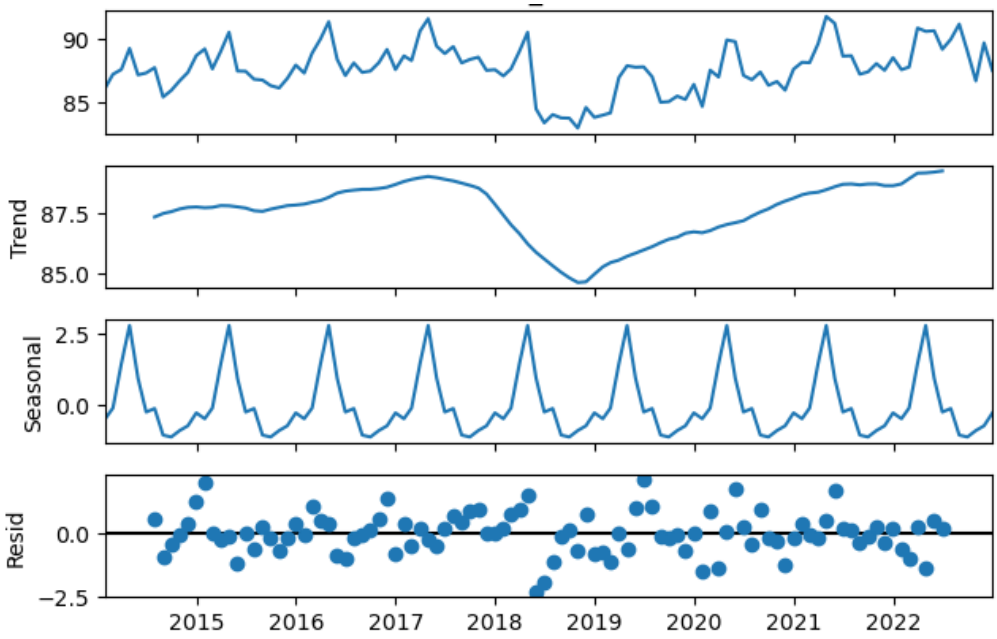


Figura 5.8: Descomposición de la Humedad relativa

En general los resultados de este análisis mostraron que todas las series presentaron estacionalidad, y todas fueron clasificadas como estacionarias según la prueba ADF, además sin evidencia de una tendencia definida. Para centrarnos mejor en las características de la estacionalidad, en la tabla 5.1 se presentan los meses en los que se observaron los picos máximos y mínimos según las estacionalidades. Además, se registran los valores de los p-values provenientes de ambas pruebas para verificar la calidad de la descomposición.

Tabla 5.1: Mínimos y máximos de la estacionalidad y P-values

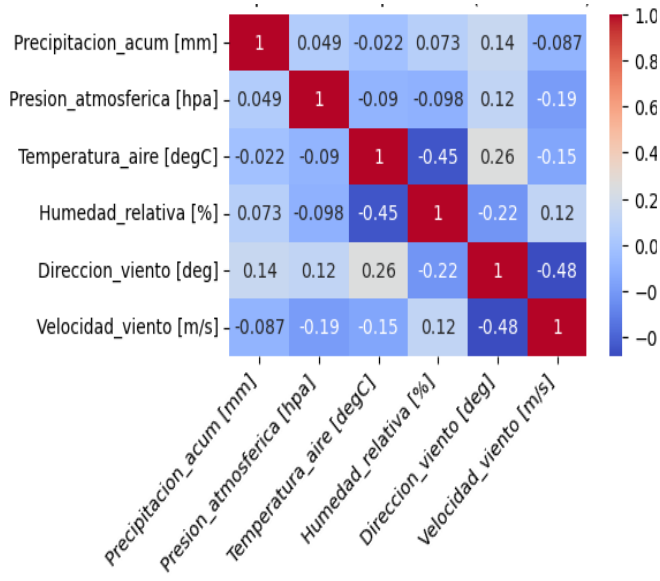
Variable	**Mínimo (mes)**	**Máximo (mes)**	**Jarque-Bera (p-value)**	**Ljung-Box (p-value)**
Humedad	06,09	4	0.9	0.06
Temperatura	3	06,09	0.16	0.063
Precipitación	6	10	0	0
Dirección del viento	3	10	0.45	0.5
Velocidad del viento	10	2	0.84	0.001
Presión	04,11	01,06	0.4	0.06
Radiación	10	07,02	0.2	0.3

Según la tabla 5.1, los p-values indican que, para la mayoría de las variables, los residuos se comportaron de manera coherente, ya que todos superan el umbral de significancia de 0.05. En estos casos, no se encuentra suficiente evidencia para rechazar la hipótesis nula, lo que sugiere que los residuos siguen una distribución normal. Sin embargo, hay un caso crítico para la variable precipitación, donde el modelo de descomposición no se ajusta adecuadamente. Esto se refleja en los p-values tanto del test Jarque-Bera como del test Ljung-Box, que indican que los residuos no siguen una distribución normal ni presentan una autocorrelación adecuada, lo que limita la precisión y validez de la estacionalidad observada en este caso.

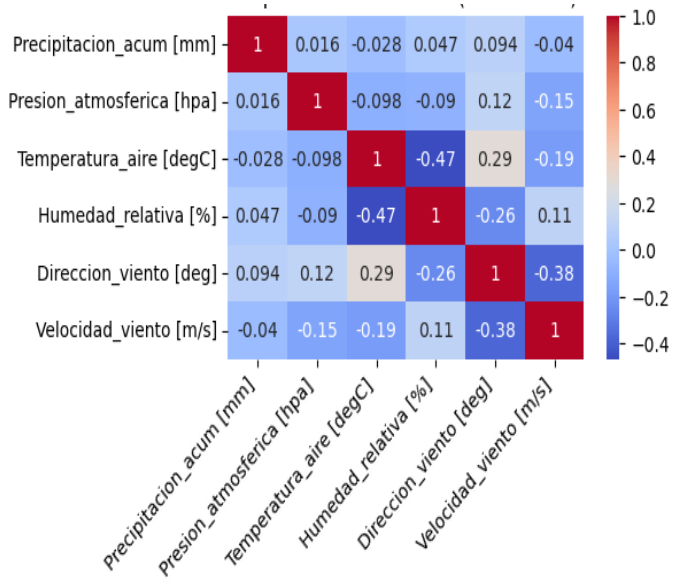
Correlación

Para realizar el análisis de correlación entre las variables, se utilizaron las correlaciones de Spearman y Pearson. La medida de Spearman fue elegida debido a su capacidad para identificar la relación monotónica entre las variables, permitiendo observar patrones generales en el comportamiento, incluso cuando no existe una relación lineal directa. Por otro lado, la correlación de Pearson fue seleccionada para cuantificar la relación lineal entre las variables.

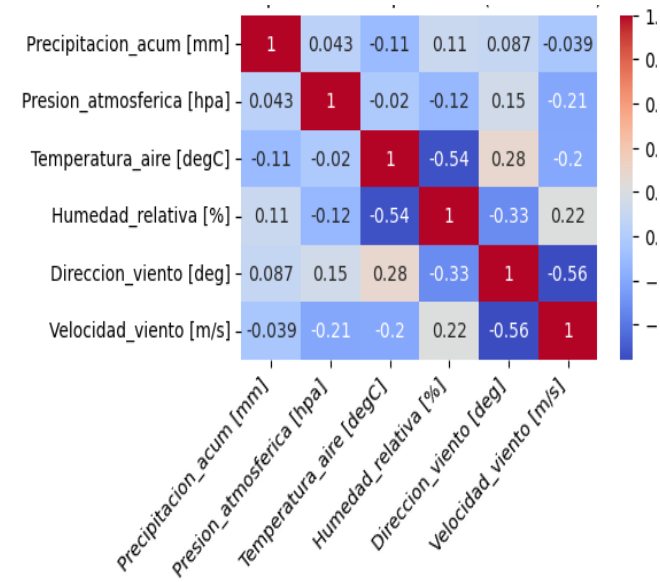
Primero, se calcularon las correlaciones para todos los periodos disponibles, como se muestra en la figura 5.9a y 5.9b. Luego, se aplicaron las mismas correlaciones únicamente a los periodos más consistentes, es decir los que obtuvieron menos datos Nan 5.9c y 5.9d, identificados previamente durante el análisis de valores faltantes en la sección anterior, este enfoque tuvo como objetivo evaluar cómo varían las correlaciones cuando se eliminan las incertidumbres asociadas con los datos imputados.



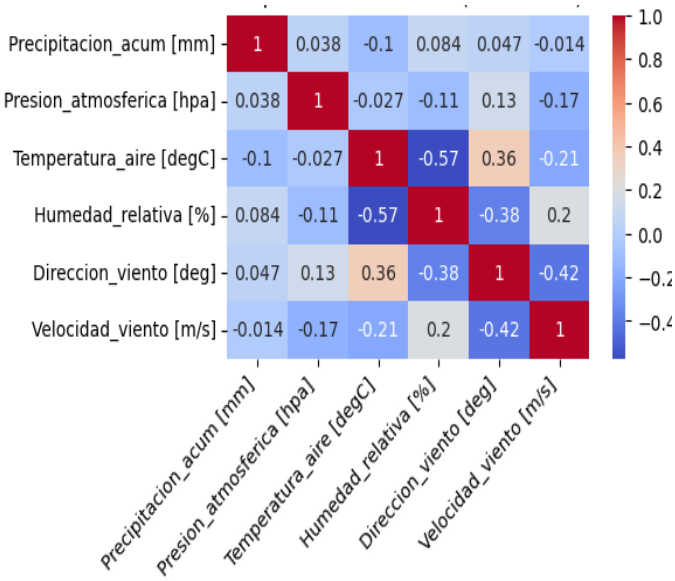
(a) Spearman 2014-2022



(b) Pearson 2014-2022



(c) Spearman 2014-2017



(d) Pearson 2014-2017

Figura 5.9: Análisis de correlaciones

A partir del análisis de las correlaciones, se observó que las variables más representativas son la temperatura y la humedad, las cuales muestran una relación inversa, es decir, cuando una variable aumenta, la otra tiende a disminuir. Además, se destacan las correlaciones inversas entre dirección del viento y velocidad del viento, así como entre dirección del viento y humedad. Por otro lado, se evidenció que la radiación solar, aunque con una menor influencia, también presenta una relación inversa con la humedad y directa con la temperatura. Al evaluar el analisis de correlaciones centrado en los períodos más consistentes, se observó un leve aumento en las correlaciones, aunque el cambio no fue drástico.

5.2 Preparación de los datos

En línea con el objetivo específico 2, que busca aplicar técnicas de preprocesamiento para el manejo eficiente de series temporales y la identificación de anomalías, se llevó a cabo la preparación de los datos. Tras realizar un análisis exploratorio, se identificaron problemas como valores faltantes (NaN) y valores atípicos extremos, los cuales fueron corregidos para garantizar la calidad de los datos. Este proceso fue esencial para asegurar que los datos estuvieran listos para el modelado.

La preparación incluyó varias tareas, tales como:

1. Re-muestreo de los datos
2. Manejo de valores extremos: Primera fase de detección de anomalías
3. Tratamiento de los valores Nan
4. Normalización de los datos
5. Selección de variables para el modelamiento

5.2.1 Remuestreo de los datos

El remuestreo de las variables se llevó a cabo con el fin de unificar las marcas temporales de los datos. Al trabajar con múltiples fuentes de información, las variables tenían frecuencias de medición distintas, lo que dificultaba su análisis conjunto. Este ajuste resultó importanate para la detección de anomalías multivariables, ya que, para que las anomalías detectadas en una variable pudieran ser comparadas con las de las demás, las observaciones debían estar alineadas temporalmente. En la siguiente tabla 5.2, se presenta la marca temporal inicial de las variables.

En la Tabla 5.2 se observa que las tres variables (velocidad del viento, dirección del viento y precipitación) presentaban una marca temporal con una frecuencia de 10 minutos. Sin embargo, solo para las variables de velocidad del viento y dirección del viento se optó por un remuestreo basado en el cálculo del promedio de cada conjunto de seis mediciones consecutivas. Por otro lado, para la variable precipitación, debido a su naturaleza acumulativa, se

Tabla 5.2: Marca de tiempo de las variables.

Variable	**Marca temporal**
Temperatura del aire	Horaria
Humedad del aire	Horaria
Presión barométrica	Horaria
Radiación global	Horaria
Velocidad del viento	10 min
Dirección del viento	10 min
Precipitación acumulada	10 min

utilizó la suma de los valores dentro de cada intervalo horario, ya que cada medición refleja la cantidad total de precipitación acumulada durante ese tiempo. Esta transformación permitió asegurar la consistencia temporal entre todas las variables del estudio. Este cambio se implementó considerando los siguientes aspectos:

- Uso de la media para el remuestreo: La media fue seleccionada como método de remuestreo debido a su capacidad para representar el comportamiento promedio de los datos en cada intervalo de tiempo, lo que es particularmente útil para la detección de anomalías. Aunque los valores atípicos pueden influir en la media, esta característica resulta beneficiosa en este contexto, ya que permite que las anomalías potenciales se reflejen directamente en el cálculo y sean identificables en el análisis posterior. De esta manera, la media no solo suaviza las variaciones normales, sino que también conserva información clave sobre desviaciones significativas, alineándose con los objetivos del estudio.

- Uso del acumulado para el remuestreo: La acumulación o suma se implementó para la variable de precipitación debido a la naturaleza de la medición de este sensor. A diferencia de las variables de viento, que miden valores instantáneos o de corta duración, el sensor de precipitación registra la cantidad total de lluvia acumulada durante un intervalo de tiempo. Por lo tanto, en lugar de utilizar el promedio, como en el caso del viento, se optó por la suma de los valores en cada intervalo para reflejar con precisión el total de precipitación acumulada en ese período. Este enfoque asegura que la información sobre la precipitación se mantenga coherente con la forma en que los sensores registran este fenómeno, respetando su naturaleza acumulativa.

- Unificación de la temporalidad: Al homogenizar las marcas temporales a una hora, se facilita la integración y el análisis conjunto de todas las variables.

- Relevancia para la detección de anomalías diarias: El objetivo del análisis es evaluar patrones y detectar anomalías que ocurren a lo largo de los días. Una marca temporal de 10 minutos genera demasiados puntos de datos dentro de cada hora, lo que no solo puede introducir redundancia, sino que también puede dificultar la identificación de tendencias relevantes. En cambio, una marca horaria ofrece una vista más clara

y manejable del comportamiento de las variables durante el día, lo que resulta más adecuado para el propósito analítico.

5.2.2 Manejo de valores extremos: Primera fase de detección de anomalías

Una vez ajustados los datos para garantizar la consistencia en las marcas de tiempo, se identificaron los valores extremos en las series temporales, como se muestra en la figura 5.2. Para abordar estos valores, nos apoyamos en el enfoque detallado en el estudio *"Sensor Network Data Fault Types"* [12], que proporciona una base sólida para analizar fallas en sensores. Según este estudio, las fallas más comunes incluyen outliers, stuck-at faults y spikes, que suelen manifestarse como desviaciones notables respecto al comportamiento esperado del sistema.

Siguiendo este enfoque, se diseñó un modelo estadístico específico para la detección de este tipo de fallas presentes en la series analizadas. A continuación, se detallan las técnicas utilizadas para abordar cada tipo de falla.

Z-Score

Como primer paso en el análisis, se filtraron los valores que excedían los rangos medibles de cada sensor,tabla 5.3, ya que estos puntos representan anomalías iniciales asociadas a las limitaciones físicas de los dispositivos. Sin embargo, estos valores no se descartaron directamente, sino que se incluyeron en el análisis posterior para determinar si corresponden a outliers o spikes. Para detectar outliers y spikes en las series temporales, se aplicó Z-score, un método que calcula la desviación estándar de cada valor respecto a la media de la serie. Si el valor absoluto del Z-score de un punto excede un umbral predefinido, ese punto es identificado como una anomalía. Este enfoque resulta útil para detectar tanto outliers como spikes, aunque estos presenten comportamientos diferentes. Los outliers son puntos aislados que se desvían significativamente de la media, mientras que los spikes corresponden a secuencias consecutivas de cambios bruscos. Sin embargo, ambos tipos de anomalías generan desviaciones extremas y, por lo tanto, pueden ser detectados de manera similar mediante el Z-score.

Para diferenciar entre outliers y spikes, se evaluaron los Z-scores de los puntos adyacentes, lo que permitió clasificar las anomalías según su comportamiento, ya sea aislado o en secuencias. El siguiente paso consistió en ajustar el umbral del Z-score de manera iterativa para cada variable, probando diferentes valores y evaluando cómo estos afectaban la detección de anomalías. Después de varias pruebas, se seleccionó el umbral que ofrecía el mejor equilibrio entre la detección precisa de anomalías y la minimización de falsos positivos. Los umbrales definitivos para cada variable se presentan en la tabla 5.3.

Variación cero

El método de Variación Cero fue adaptado para detectar stuck-at faults en las series temporales, los cuales se caracterizan por valores constantes o sin variación durante períodos prolongados. Para aplicar esta técnica, se calculó la diferencia entre los valores consecutivos de la serie temporal. Cuando esta diferencia era igual a cero, se consideraba que el dato representaba un "stuck-at fault", siempre que no fuera nulo. Sin embargo, para evitar que fluctuaciones naturales de los datos fueran erróneamente etiquetadas como anomalías, se estableció un umbral de longitud mínimo. Solo las secuencias de variación cero que mantenían una constancia significativa durante un número predefinido de puntos consecutivos fueron clasificadas como "stuck-at faults". Esto permitió una detección más precisa y específica de este tipo de fallos.

El proceso de adaptación del umbral de longitud para las secuencias de variación cero fue iterativo y específico para cada variable. Se probaron diferentes valores de umbral para cada conjunto de datos, ajustándolos hasta encontrar aquel que permitiera distinguir de manera precisa entre las secuencias normales y las anomalías. En algunos casos, los umbrales bajos resultaron en la detección de comportamientos normales, mientras que en otros revelaron anomalías. Con cada ajuste, las secuencias detectadas que superaban el umbral empezaron a mostrar una clara desviación de los patrones habituales de la serie, reflejando comportamientos anómalos más evidentes.

A continuación, se presentan los umbrales seleccionados para cada serie en la siguiente tabla 5.3.

Tabla 5.3: Umbrales y rangos medibles de los sensores.

Variable	**Umbral (V-cero)**	**Umbral (Z-score)**	**Rango del sensor**
Humedad relativa [%]	18	1.2	(0, 100)
Precipitación acumulada [mm]	3000	6	(0, 50)
Presión atmosférica [hPa]	5	0.8	(500, 1100)
Radiación solar [W/m^2]	40	3	(0, 1400)
Temperatura aire [°C]	10	1	(-10, 60)
Dirección viento [deg]	9	3	(0, 359)
Velocidad viento [m/s]	10	1	(0, 80)

En las siguientes figuras se presentan las anomalías causadas por fallas en los sensores en cada variable, detectadas tras aplicar los métodos de Z-score y variación cero.

Figura 5.10: Humedad Relativa

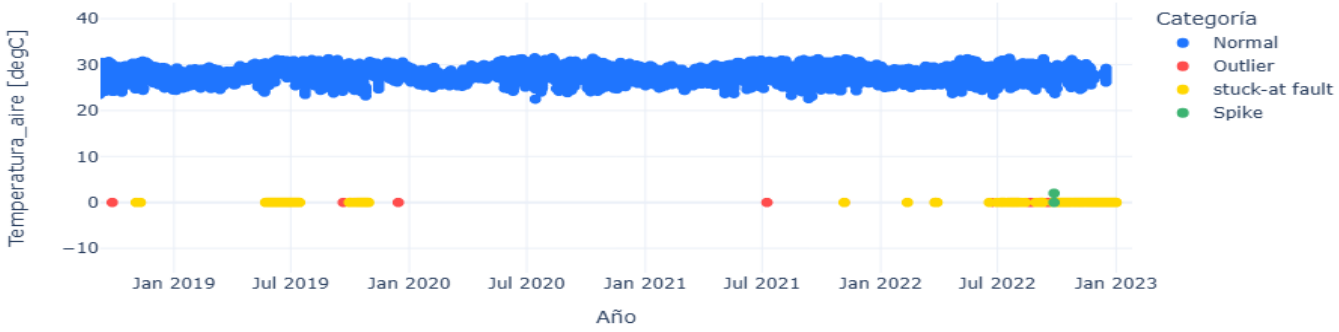


Figura 5.11: Temperatura

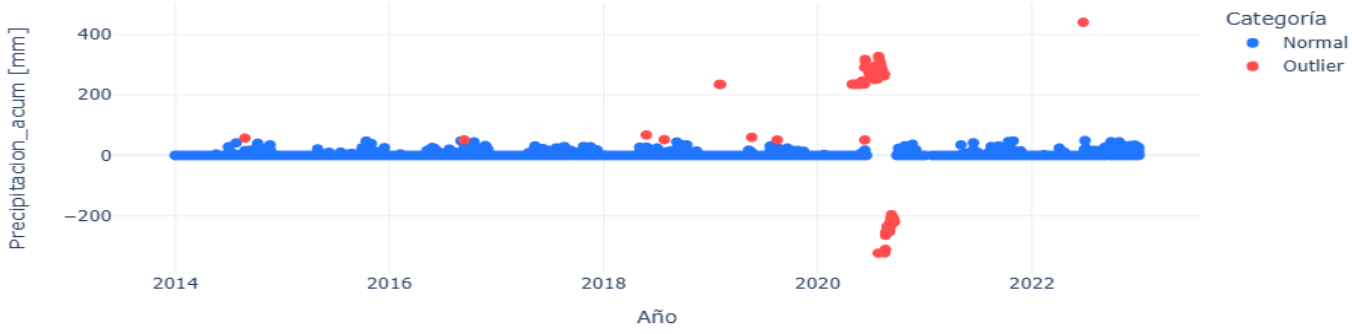


Figura 5.12: Precipitación

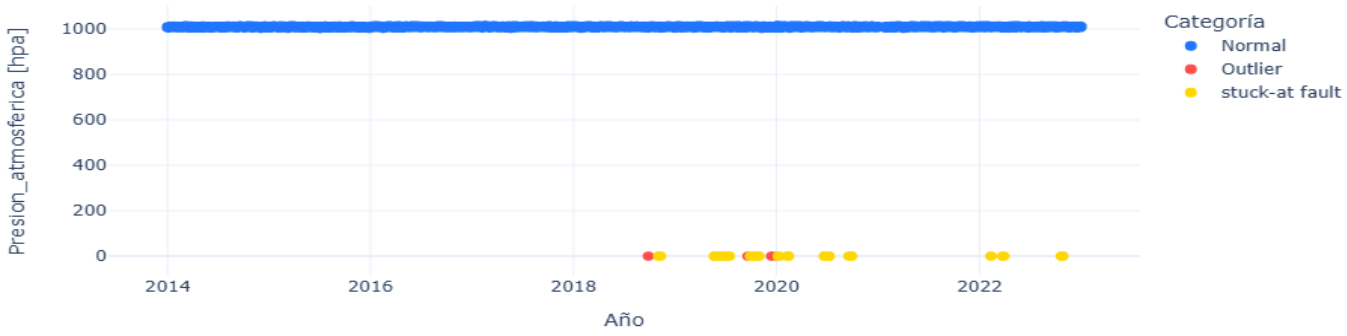


Figura 5.13: Presión atmosférica

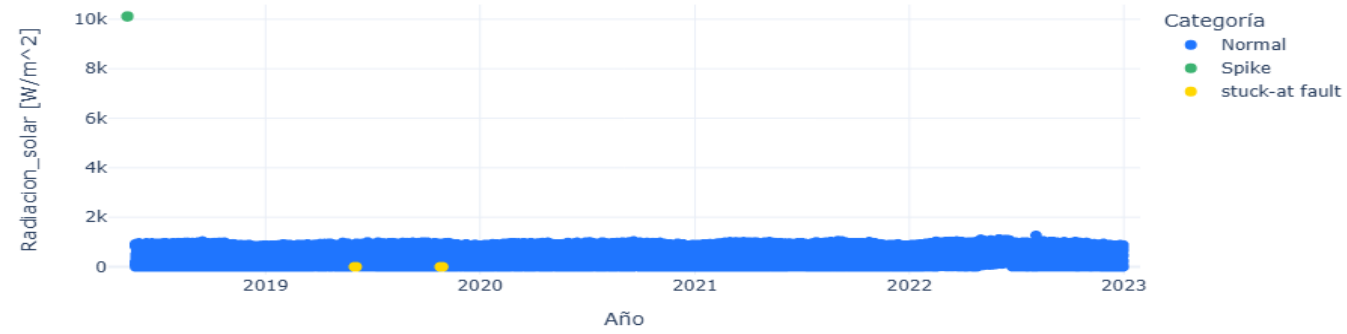


Figura 5.14: Radiación Solar

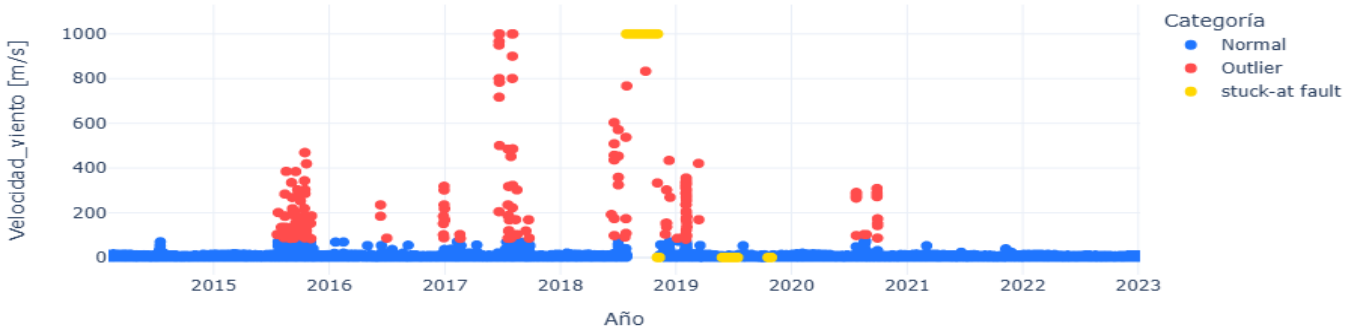


Figura 5.15: Velocidad del viento

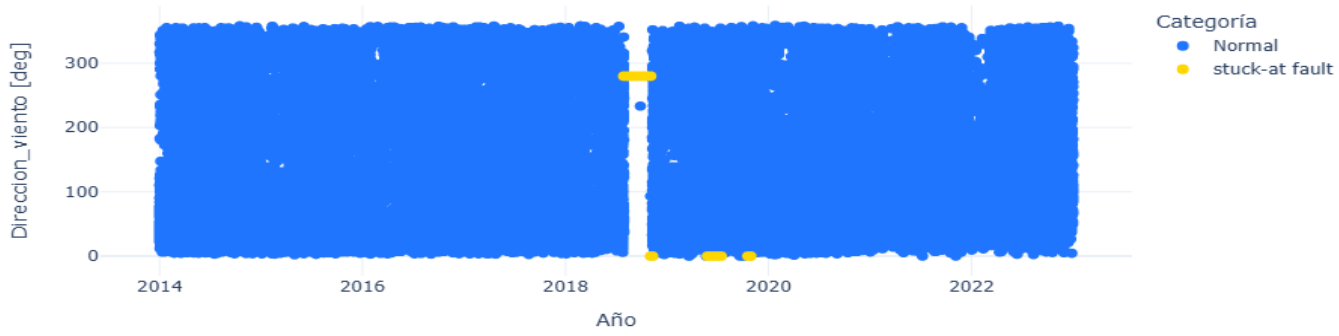


Figura 5.16: Dirección del viento

Los gráficos mostrados anteriormente permiten observar la detección de las anomalías causadas por fallas en los sensores para cada una de las variables. En particular, se destaca el gráfico de la humedad relativa, donde es evidente la presencia de los tres tipos de anomalías. Se puede apreciar que estos valores fueron correctamente identificados de acuerdo con su comportamiento. Lo mismo ocurrió con las otras variables, donde las anomalías fueron detectadas de manera adecuada.

Evaluación del modelo estadístico

Esta sección está alineada con los objetivos específicos 3 y 4, que buscan identificar e implementar un modelo de detección de anomalías y evaluar su rendimiento comparándolo con muestras previamente evaluadas. La evaluación del modelo se realizó utilizando las etiquetas de calidad desarrolladas por la DIMAR, que sirvieron como un evaluador aproximado para el algoritmo. Estas etiquetas reflejan únicamente la calidad de los datos y no representan directamente anomalías. Por ejemplo, la etiqueta de calidad (4), como se muestra en la tabla 4.1 , clasifica los datos como "Malos". Sin embargo, esta categoría no distingue entre distintos tipos de anomalías, como picos abruptos (spikes), valores atípicos (outliers) o fallos persistentes (stuck-at faults).

A pesar de esta limitación, para todas las avriables se observó una precisión aproximada del 85 % al comparar los resultados del algoritmo con las etiquetas de calidad (4), lo que indica que la mayoría de los datos clasificados como "Malos"por la DIMAR corresponden a anomalías identificadas por nuestro modelo. Esta evaluación aproximada permitió obtener una idea general sobre la validez del algoritmo en la detección de anomalías.

5.2.3 Tratamiento de los valores Nan

Antes de iniciar el tratamiento de los valores Nan o perdidos, se realizó un paso adicional de procesamiento relacionado con los datos anómalos detectados en la sección anterior. Estas

anomalías, asociadas a fallas en los sensores, fueron eliminadas porque alteraban el comportamiento o la naturaleza normal de las variables, lo que podría afectar el análisis posterior. Esta eliminación incrementó la cantidad de valores Nan, pero resultó muy importante para garantizar la validez de los resultados. A continuación en en la figura 5.17 se presentan las nuevas cantidades de valores perdidos tras esta etapa de limpieza.

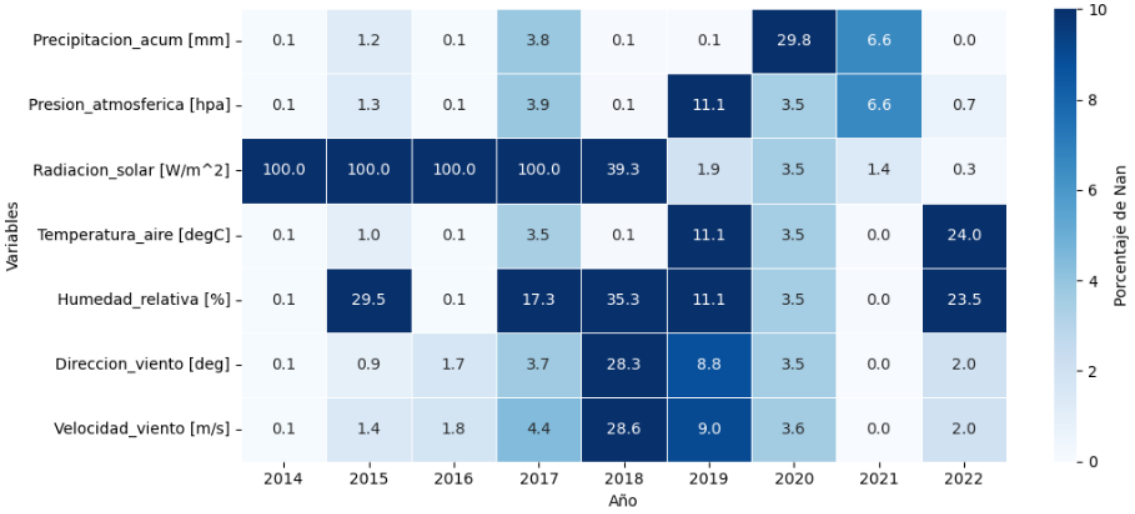


Figura 5.17: Mapa de calor de Porcentaje de datos Nan por variable y por año

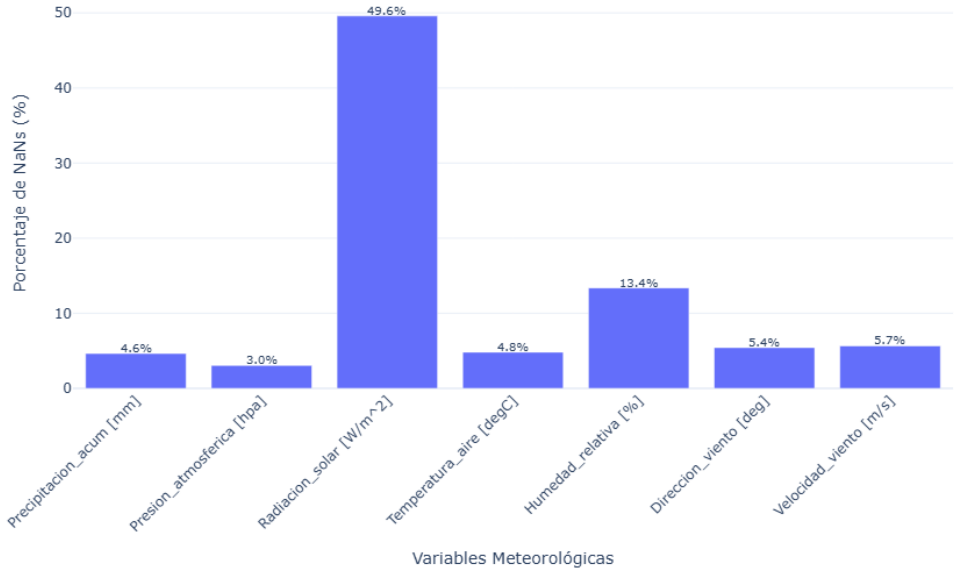


Figura 5.18: Gráfico de barras datos Nan

A partir de las figuras 5.17 y 5.18 se observó un aumento significativo en los datos faltantes (NaN) para la mayoría de las variables, especialmente en los años 2018 y 2019. Sin embargo, en el caso de la precipitación, la mayor pérdida de datos ocurrió en 2020. En general, la mayoría de las variables obtuvieron un porcentaje de datos faltantes cercano al 5 %, a excepción de la humedad relativa, que destaca como la más afectada, con un 13 % de datos perdidos.

Imputación de punto cercano con la mediana

Para realizar la imputación de los datos faltantes, nos basamos en el estudio: Aplicación de ciencia de datos para la reconstrucción de series de tiempo de variables meteorológicas en Islas del Rosario (Caribe colombiano) [23], en el que los autores trabajaron con datos de la misma estación meteorológica utilizada en este trabajo. Este enfoque fue particularmente importante, ya que aseguró que la metodología aplicada se ajustara a las características y condiciones de las series de tiempo en cuestión.

El método implementado fue la imputación de punto cercano con la mediana, que organiza los valores cercanos para seleccionar la mediana como el reemplazo del valor faltante, lo que garantiza que el valor imputado esté dentro del rango de los datos observados, reduciendo la probabilidad de replicar o amplificar anomalías.

El proceso de imputación siguió los siguientes pasos:

1. Se agruparon los datos de cada variable por hora del día y por día del año, para cada año de la serie temporal.
2. Se calculó la mediana para cada combinación específica de día y hora a lo largo de todos los años disponibles.
3. Se utilizó la mediana calculada para imputar los valores faltantes en los años correspondientes.

Después de realizar las imputaciones en cada variable, se obtuvieron series temporales completas y consistentes, como se observa en la figura 5.19, donde se aprecia una armonía entre los datos imputados y los datos originales. Sin embargo, es importante destacar que los periodos imputados, aunque muestran un comportamiento similar al de los datos reales, presentan un nivel de incertidumbre mayor en comparación con estos últimos.

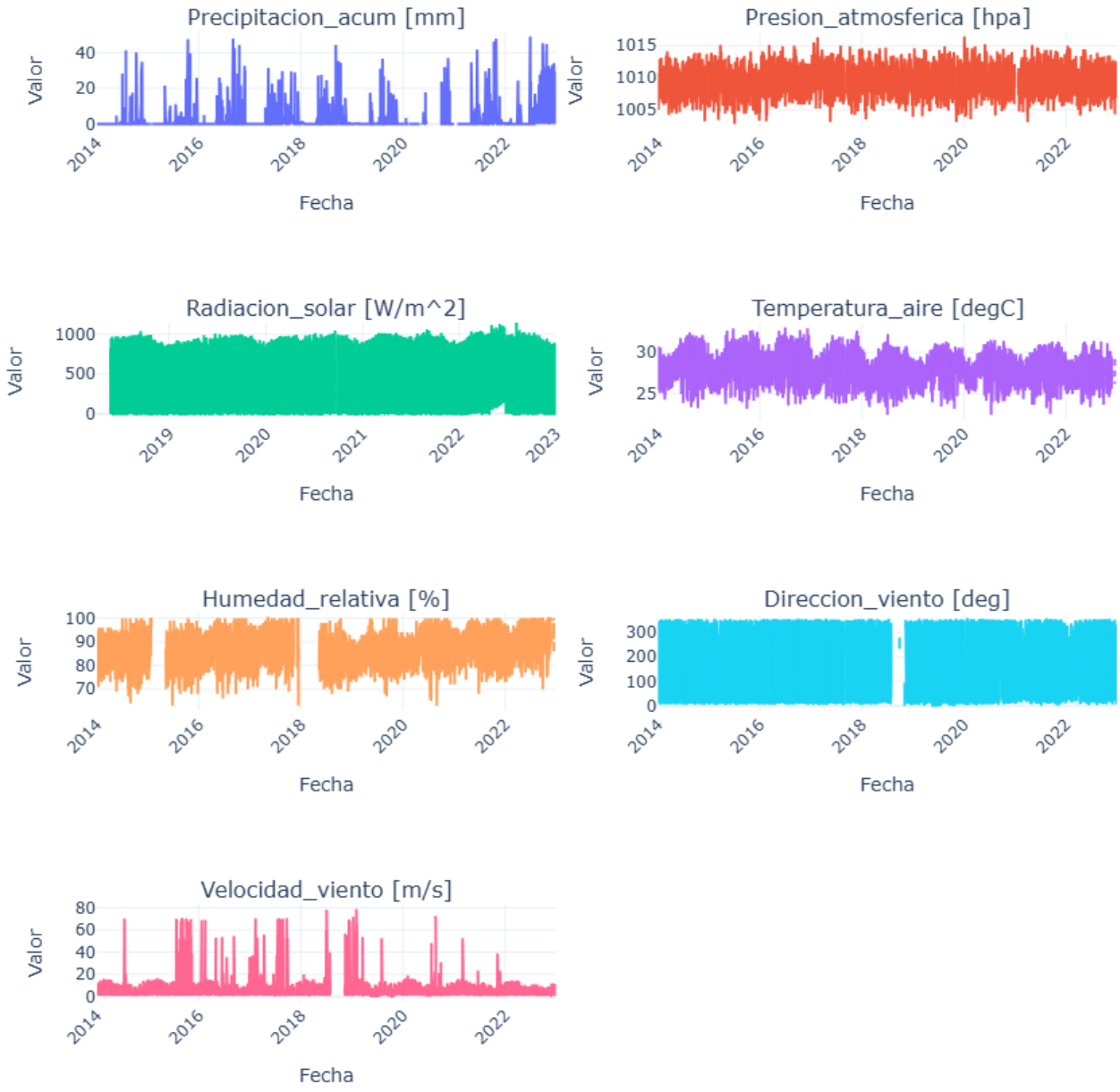


Figura 5.19: Series imputadas

5.2.4 Normalización de los datos

Después de limpiar y organizar los datos, se procedió a la normalización utilizando el Z-score, con el objetivo de estandarizar los valores y evitar que las diferencias en la magnitud de las variables afectaran los resultados del análisis. Esta técnica permitió convertir los datos en un formato centrado en la media, con dispersión ajustada, facilitando su comparación

directa. En la etapa siguiente, se aplicaron modelos de agrupamiento basados en distancia, los cuales requieren de una métrica precisa para medir la similitud entre los puntos. Al operar con distancias, fue fundamental que los datos estuvieran normalizados, esto garantizó que las variables se encontraran en un mismo rango, permitiendo que las técnicas de agrupamiento se ajustaran bien a los datos y no se generaran sesgos debido a las escalas diferentes.

5.2.5 Selección de variables para el modelamiento

Con base en el análisis de correlación realizado en 5.1.2, donde se identificó la correlación entre todas las series de tiempo, se seleccionaron los grupos de variables que presentaban una correlación Spearman mínima de al menos 0.40 entre sí. Se consideró importante trabajar con variables correlacionadas ya que permiten que el modelo capture las dinámicas naturales del sistema y evitan la inclusión de información irrelevante que podría añadir ruido al análisis. El enfoque consiste en aprovechar las relaciones esperadas entre las variables: si en condiciones normales estas muestran un comportamiento correlacionado, cualquier desviación significativa de esa relación puede ser considerada una anomalía. Por ejemplo, si dos variables que suelen estar relacionadas no presentan el patrón esperado en ciertos momentos, esos días pueden indicar eventos atípicos asociados a fenómenos externos o cambios inusuales en el sistema. Este método no solo mejora la detección de anomalías relevantes, sino que también facilita la interpretación de los resultados al centrarse en patrones que reflejan las dinámicas reales. La series multivariadas seleccionadas fueron las siguientes:

- Tempertura del aire y Humedad relativa (TA,HR)
- Dirección del viento y Velocidad del viento (DV,VV)

5.3 Modelamiento

Esta sección está alineada con el objetivo específico 3, que busca identificar e implementar un modelo de detección de anomalías basado en su eficacia según métricas de rendimiento. El proceso de modelamiento se centró en la aplicación de dos técnicas de clustering, utilizando los algoritmos *K-Means* y *DBSCAN*. El modelamiento se llevó a cabo en varias etapas, que se enfocaron en la selección de datos, la elección de las métricas de similitud, la definición de los hiperparámetros, y la aplicación de los modelos a diferentes ventanas temporales.

5.3.1 Matriz de pasos de tiempo

Con el objetivo de identificar días anómalos, se diseñó una matriz de pasos de tiempo que permitió observar patrones a lo largo del tiempo, en lugar de enfocarse únicamente en puntos aislados. Este enfoque fue importante porque permitió captar de manera más precisa y significativa los eventos anómalos al considerar el comportamiento de las variables a lo largo de un período de tiempo, en lugar de limitarse a analizar puntos aislados que podrían ser simples fluctuaciones o ruido. La matriz de pasos de tiempo se estructura con filas que representan cada uno de los días, y columnas que corresponden a las distintas horas del día.

Cada celda de la matriz almacena un vector que contiene los valores de todas las variables consideradas, permitiendo así un análisis multivariado que captura la evolución conjunta de los datos a lo largo del tiempo.

Por ejemplo para un año completo de datos multivariados (temperatura y humedad relativa), la matriz de pasos de tiempo tuvo la siguiente dimensión:

Dimensión total: (365, 24, 2)

- 365 filas: Representan cada uno de los días del año.
- 24 columnas: Corresponden a las 24 horas de cada día.
- 2 valores: Cada celda almacena un vector con los valores simultáneos de temperatura y humedad relativa para ese día/hora.

5.3.2 Medidas de similitud

Una vez estructuradas las series multivariadas en formato de matrices de pasos de tiempo, se seleccionaron las métricas de similitud para el ajuste de los algoritmos de clustering. Las métricas elegidas fueron a Distancia Temporal de Warping (DTW) y la distancia Euclidiana. Estas dos métricas fueron elegidas porque proporcionan diferentes perspectivas para modelar las similitudes entre las series. permitiendo una evaluación más completa de los patrones temporales desde dos enfoques distintos de medidas de similitud, que se implementarían en la grilla de hiperparametros.

5.3.3 Grilla de hiperparametros

La siguiente etapa consistió en definir las grillas de hiperparámetros para los algoritmos *K-Means* y *DBSCAN*.

K-Means

- Número de clusters: Desde 2 hasta 15.
- Número de inicializaciones: 10, 20
- Número de iteraciones máximas: 10, 20
- Medidas de similitud: DTW (Dynamic Time Warping) y Distancia Euclidiana

DBSCAN

- Valores de epsilon(eps): Desde 1 hasta 50 (con un incremento de 1).
- Valores de minsamples: Desde 2 hasta 50 (con un incremento de 1).
- Medidas de similitud: DTW (Dynamic Time Warping) y Distancia Euclidiana

5.3.4 Ventanas de tiempo

En esta sección se establecieron dos enfoques de ventanas de tiempo para la aplicación de los algoritmos de clustering, cada una diseñada con un propósito específico acorde al tipo de anomalias que se buscaban identificar:

- Ventana anual o local: Este enfoque considera el análisis de datos agrupados por año, lo que permite detectar patrones y anomalias a gran escala que se manifiestan de manera cíclica o estacional en el transcurso de un año.
- Ventana global: En este caso, se agrupan todos los datos sin distinción temporal, aplicando los modelos de clustering de forma completa sobre el conjunto global de información disponible. Esto es útil para identificar anomalias que abarcan todo el período analizado, proporcionando un panorama general de posibles eventos atípicos que podrían ocurrir en cualquier momento.

5.3.5 Recursos técnicos utilizados

Para este trabajo se utilizó el lenguaje de programación Python y la plataforma Jupyter Lab. El hardware empleado contó con las siguientes especificaciones: RTX 3090, AMD 5800x3d, 64 GB de RAM y 1 TB de HDD.

Las librerías utilizadas para la preparación de las series de tiempo, modelamiento y evaluación de series temporales fueron:

- Numpy y Pandas para la manipulación y limpieza de datos.
- Tslearn para el preprocesamiento y modelamiento de series temporales.
- Sklearn para los algoritmos de clustering y evaluación.
- Seaborn, Matplotlib y Plotly para las visualizaciones.
- Statsmodels y Scipy para análisis estadísticos y operaciones avanzadas.

5.4 Aplicación y evaluación de los modelos

Esta sección está alineada con el objetivo específico 4, que busca evaluar el rendimiento del modelo de detección de anomalías. Una vez definidos los modelos, se procedió a su aplicación sobre cada serie de tiempo multivariada, utilizando las ventanas temporales previamente establecidas. Cada modelo se ejecutó con la grilla de hiperparámetros correspondiente, ajustando los parámetros para obtener los mejores resultados en cada caso, la selección de los mejores hiperparámetros se realizó tomando en cuenta dos métricas de evaluación de algoritmos de clustering. La primera de ellas fue la métrica de la silueta, que permitió evaluar la cohesión y separación de los clústeres generados. Además, se utilizó el índice de Davies-Bouldin, que evalúa la calidad del clustering considerando tanto la dispersión interna de los clústeres como

la distancia entre ellos, lo que complementó la evaluación de la separación y compactación de los grupos identificados.

De salida, para cada día se obtuvo una etiqueta de cluster, lo que permitió clasificar los datos en función de su comportamiento. Los clusters con un menor número de puntos fueron identificados como representativos de comportamientos raros o anómalos. Un análisis más profundo de estos datos anómalos permitirá evaluar su posible correspondencia a eventos climáticos específicos, proporcionando información valiosa para entender las interacciones entre los patrones anómalos y los fenómenos meteorológicos recurrentes, mientras que el cluster mayoritario correspondió a los datos normales, facilitando así la distinción entre anomalías y patrones habituales en las series de tiempo.

6 Resultados y Análisis

A continuación, se presentan los mejores resultados obtenidos tras la aplicación de los algoritmos de clustering sobre las series de tiempo multivariadas. En las siguientes tablas se muestran los detalles sobre los clústeres generados, métricas de evaluación como el índice de Davies-Bouldin y el Silhouette Score, y los hiperparámetros óptimos encontrados durante la optimización de los modelos. Cabe destacar que los clusters con menor número de series fueron los que se consideron como comportamientos anómalos, evidenciando anomalías que requieren una atención especial para su análisis en el área.

6.1 DBSCAN

Los resultados obtenidos con el algoritmo DBSCAN siempre fueron más efectivos cuando se utilizó como métrica de similitud el DTW. Por lo tanto, los siguientes resultados presentados en las tablas corresponden a los obtenidos después de aplicar el algoritmo utilizando esta métrica de similitud.

Tabla 6.1: Clustering con DBSCAN Y ventana global

Series Multivariadas	**TA y HR**	**VV y DV**
Clusters	2	2
Cantidad en cada Cluster	97 %, 3 %	70 %, 30 %
EPS	2	4
min samples	28	135
Puntaje de Silueta	0.67	0.33
Índice Davies-Bouldin	0.54	1.16
Tiempo de ejecución	10min	10.5min

Tabla 6.2: Clustering con ventana anual para (Humedad y Temperatura)

Año	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
Clusters	2	2	2	2	2	2	2	2	2
Cantidad en cada cluster	360,5	360,5	350,15	352,13	362,3	363,2	354,12	352,13	357,8
EPS	2	2	2	2	2	2	2	2	2
Min Samples	10	2	28	24	16	2	11	12	15
Puntaje de Silueta	0.64	0.62	0.59	0.5	0.52	0.4	0.43	0.52	0.53
Índice Davies-Bouldin	0.39	0.47	0.53	0.66	0.49	0.95	0.74	0.64	0.56

Tabla 6.3: Clustering con ventana anual para (Velocidad y dirección del viento)

Año	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
Clusters	2	2	2	2	2	2	2	2	2
Cantidad en cada cluster	263, 102	252, 113	228, 138	159, 206	364, 1	240, 125	199, 167	253, 112	202, 163
EPS	4	4	4	4	6	4	4	4	4
Min Samples	44	49	32	47	2	41	46	39	41
Puntaje de Silueta	0.41	0.38	0.34	0.26	0.25	0.35	0.34	0.40	0.35
Índice Davies-Bouldin	0.99	1.06	1.17	1.42	0.56	1.14	1.17	1.01	1.13

6.1.1 Análisis de resultados con enfoque ventana global

En la Tabla 6.3, se presentan las métricas de evaluación para la serie multivariada (TA y HR), evidenciando una buena calidad en la clusterización. El Silhouette Score alcanzado fue de 0.67, lo que indica que los datos están adecuadamente agrupados dentro de sus respectivos clústeres, con una separación notable entre ellos. Este valor, cercano al ideal de 1, sugiere que los puntos están significativamente más cerca del centroide de su propio clúster que del de clústeres vecinos. Esto respalda la idea de un agrupamiento compacto y bien definido. Por otro lado, el Davies-Bouldin Index (DBI) fue 0.54, un valor bajo (menor que 1) que respalda la calidad de los clústeres, ya que minimiza la dispersión interna de los datos dentro de cada grupo y maximiza la distancia entre los clústeres. La coherencia entre ambas métricas evidencia que el número de clústeres seleccionado es adecuado para capturar la estructura subyacente de los datos, proporcionando una agrupación confiable y útil para la identificación de las anomalías.

En contraste, al aplicar este mismo enfoque de ventana global a la serie multivariada (VV, DV), tabla 6.3, se obtuvieron resultados menos satisfactorios. El Silhouette Score fue de 0.33, un valor que indica un agrupamiento débil, ya que es significativamente menor al umbral de

0.5, lo que sugiere que muchos puntos están cerca de los límites entre clústeres. Además, el Davies-Bouldin Index fue de 1.16, un valor superior a 1 que señala una menor calidad en los clústeres debido a una mayor dispersión dentro de los mismos y a una separación insuficiente entre los grupos. Estos resultados indican que el agrupamiento para la serie (VV, DV) no fue tan efectivo como para la serie (TA, HR). Sin embargo, es importante destacar que en ambos casos las métricas de evaluación son consistentes y se respaldan mutuamente, lo que refuerza la validez de las conclusiones derivadas.

A continuación se muestran los días agrupados según su respectivo clúster y variable.

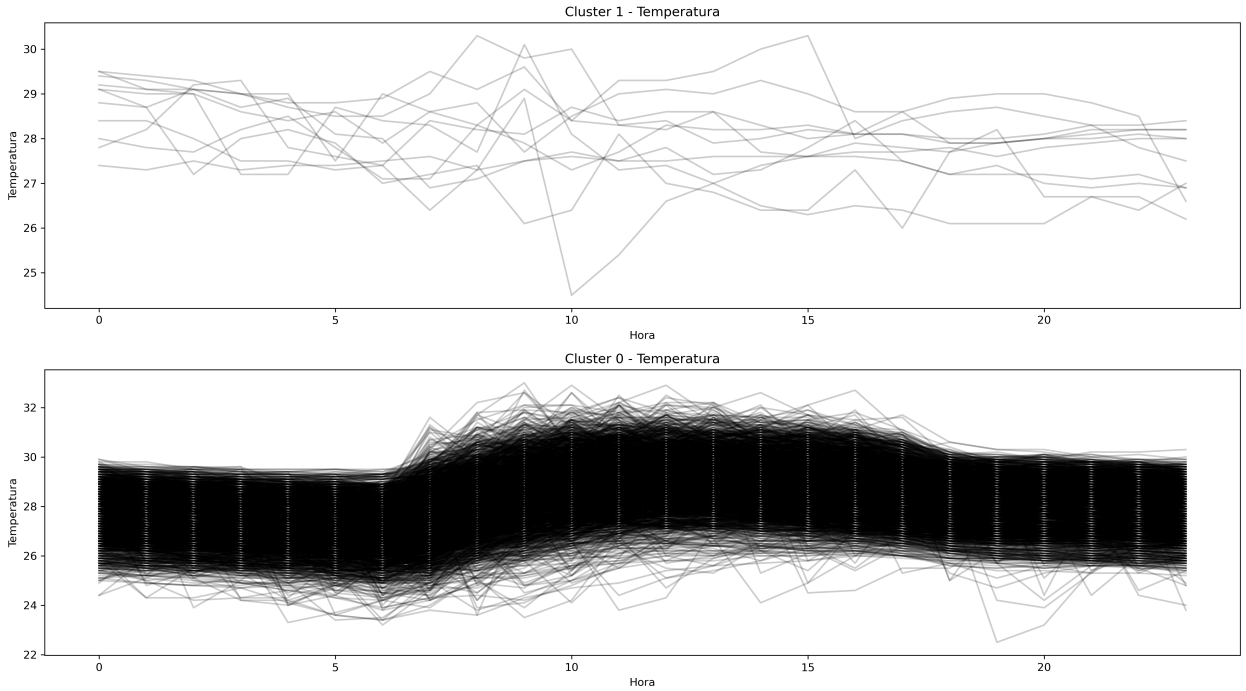


Figura 6.1: Clústerización de temperatura: Patrones diarios

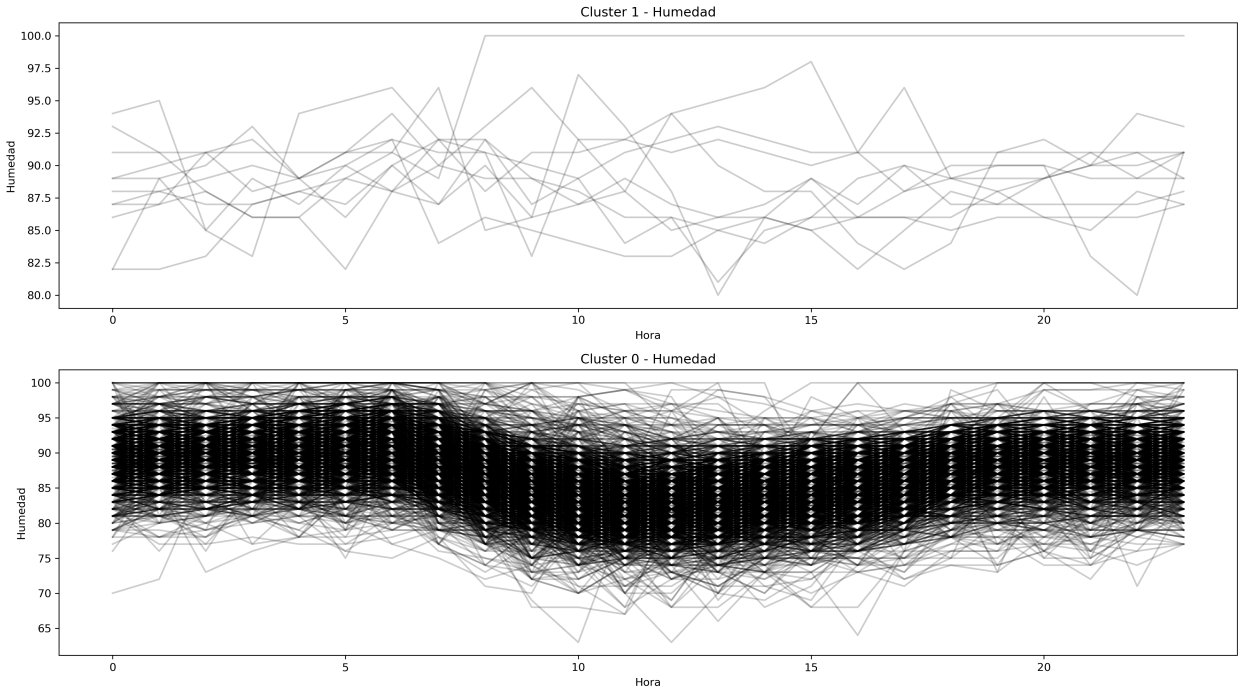


Figura 6.2: Clústerización de Humedad: Patrones diarios

A partir de las gráficas en las figuras 6.1 y 6.2, se puede observar que el clúster predominante, en este caso el clúster 0, representa la normalidad y un comportamiento consistente en las series diarias de humedad y temperatura. En este clúster, se evidencia una clara estacionalidad diaria, y se puede apreciar la evidente relación entre ambas variables y los intervalos normales que se marcan. Por otro lado, el clúster 1, que representa las anomalías, muestra comportamientos atípicos que no se ajustan a la normalidad en ambas variables. En este clúster, las series parecen estar más dispersas, con una mayor presencia de ruido o incluso mostrando periodos de poca variación, donde los intervalos entre los cambios de humedad y temperatura difieren claramente del patrón normal visto en el cluster 0.

6.1.2 Análisis de resultados enfoque ventana local o anual

A partir de la tabla 6.2, para los años estudiados, se obtuvo un puntaje promedio de Silhouette de 0.52, con un valor mayor o igual a 0.5 en 7 de los 9 años analizados. Esto sugiere un agrupamiento razonable en cada periodo, apoyado también por un índice promedio de Davies-Bouldin de 0.60, donde todos los valores individuales son inferiores a 1, reflejando una separación adecuada entre los clústeres. En este sentido, el algoritmo DBSCAN mostró un buen desempeño en el enfoque anual para la detección de días anómalos en la serie multivariada de temperatura y humedad. Por otro lado, en la serie multivariada de velocidad y dirección del viento, el puntaje promedio de Silhouette fue de 0.35, y el índice Davies-Bouldin alcanzó un valor de 1.07, ambos indicadores que evidencian un agrupamiento débil. Esto sugiere que DBSCAN tuvo un desempeño limitado en la identificación de días anómalos.

En general, DBSCAN mostró un desempeño sólido tanto en el enfoque global como en el local para la serie multivariada de temperatura y humedad. Sin embargo, en ambos casos, tuvo un bajo rendimiento en la detección de anomalías en la serie multivariada de velocidad y dirección del viento. Este desempeño limitado podría atribuirse a la complejidad inherente en el manejo y la reducción de dimensionalidad de las variables relacionadas con el viento, las cuales son altamente sensibles y están registradas con una frecuencia temporal de 10 minutos. Estas características pueden dificultar la identificación de patrones claros y la distinción entre anomalías y variaciones naturales.

6.2 K-Means

Los resultados obtenidos con el algoritmo K-Means siempre fueron más efectivos cuando se utilizó como métrica de similitud el DTW. Por lo tanto, los siguientes resultados presentados en las tablas corresponden a los obtenidos después de aplicar el algoritmo utilizando esta métrica de similitud.

Tabla 6.4: Clustering con K-Means y ventana global

Series Multivariadas	**TA Y HR**	**VV Y DV**
Número de clusters	2	3
Iteraciones máximas	10	10
Inicializaciones	10	10
Puntaje de silueta	0.26	0.3
Índice Davis-Bouldin	1.2	1.1

Tabla 6.5: Clustering con ventana anual para (Humedad y Temperatura)

Año	2014	2015	2016	2017	2018	2019	2020	2021	2022
# Clusters	3	3	2	3	2	2	2	2	2
Iteraciones máximas	20	10	20	10	20	10	10	20	10
Inicializaciones	10	10	20	10	20	10	10	20	10
Puntaje de silueta	0.24	0.24	0.4	0.27	0.26	0.23	0.26	0.28	0.29
Índice Davis-Bouldin	1.39	1.4	1.3	1.36	1.56	1.65	1.54	1.54	1.47

Tabla 6.6: Clustering con ventana anual para (Velocidad y Dirección del Viento)

Año	2014	2015	2016	2017	2018	2019	2020	2021	2022
# Clusters	2	3	4	6	2	2	4	3	6
Iteraciones máximas	10	10	10	10	20	10	20	10	10
Inicializaciones	10	10	10	10	10	10	10	10	10
Puntaje de silueta	0.34	0.30	0.28	0.30	0.30	0.31	0.29	0.31	0.28
Índice Davis-Bouldin	1.60	1.35	1.36	1.26	1.40	1.42	1.39	1.38	1.28

6.2.1 Análisis de resultados

A partir de las tablas 6.4, 6.5 y 6.6, se observa que para todos los enfoques, ya sea anual o global, y en ambas series multivariadas, los valores obtenidos para el puntaje de silueta son menores a 0.5, mientras que el índice Davis-Bouldin supera los 1. Esto indica que el rendimiento del algoritmo K-means no fue óptimo en la clusterización, ya que los clusters formados no logran representar fielmente la estructura subyacente de los datos. A pesar de que el algoritmo logra identificar algunos clusters, la calidad de los mismos no es lo suficientemente buena, evidenciado por la dispersión de los resultados.

Uno de los factores detrás del bajo rendimiento de K-means radica en su dependencia de la inicialización de los centroides, lo que lo hace susceptible a quedar atrapado en soluciones locales y no alcanzar una convergencia adecuada hacia el patrón real de los datos. Adicionalmente, este algoritmo asume que los clusters son convexos, lo que puede ser una limitación cuando se trabaja con datos multivariados que presentan estructuras irregulares. Además, la dimensionalidad también juega un papel crítico: en datos con alta dimensionalidad, la distancia entre puntos tiende a ser menos informativa, haciendo que la separación de clusters sea inexacta.

Sumado a esto, las limitaciones computacionales impidieron realizar una evaluación más exhaustiva, restringiendo el número de iteraciones y inicializaciones exploradas. Frente a esta situación, DBSCAN se muestra como una alternativa más robusta, ya que no depende tanto de la inicialización y es capaz de identificar clusters con formas arbitrarias, lo que lo hace más adecuado para manejar datos multivariados complejos. Por lo tanto, K-means no resultó ser el método más adecuado para este caso, demostrando que su uso requiere ajustes más cuidadosos en cuanto a configuración y contexto para obtener resultados confiables.

7 Conclusiones

- Con el enfoque híbrido implementado en este trabajo, se obtuvieron resultados satisfactorios para la serie temporal multivariada de temperatura del aire y humedad relativa, alcanzando un puntaje de silueta de 0.67 y un índice Davies-Bouldin de 0.33 por medio de DBSCAN. Esto brindó una buena base para la identificación de días anómalos globales, destacando aquellos días que podrían necesitar un análisis más detallado por el analista en el área y mejorar la toma de decisiones, dandole cumplimiento al objetivo general y el objetivo especifico 3 en este proyecto .
- Con el propósito de alcanzar el objetivo de identificar las variables meteorológicas más relevantes para la detección de anomalías, el proceso de Análisis Exploratorio de Datos jugó un papel importante desde la primera fase. Durante esta etapa, se evaluaron aspectos como la identificación de datos perdidos y las características inherentes a las series temporales, lo que sentó las bases para la primera selección de variables. En la segunda fase, el análisis de correlación fue de gran importancia al seleccionar aquellas variables que serían sometidas a técnicas de clusterización. Este enfoque no solo permitió destacar las relaciones significativas entre las variables, facilitando la interpretación de cualquier desviación en sus patrones esperados, sino que también ayudó a filtrar aquellas variables menos relevantes que podrían introducir ruido innecesario. De este modo, se garantizaron resultados más precisos y fiables, optimizando la selección de variables para un análisis más detallado.
- La calidad del proceso de preparación y limpieza de las series de tiempo fue fundamental, ya que constituyó la base para obtener resultados confiables en la detección de anomalías mediante técnicas de clusterización. El proceso adecuado aseguró que los datos estuvieran libres de inconsistencias, valores atípicos y errores que podrían afectar negativamente los análisis posteriores, lo que permitió identificar las anomalías de manera más efectiva, alineándose con el objetivo de aplicar técnicas avanzadas de preprocesamiento para mejorar la calidad del análisis y la interpretación de las seires meteorológicas.
- El enfoque de identificar días anómalos, en lugar de centrarse únicamente en datos puntuales, fue beneficioso para los investigadores del área, ya que al etiquetar un día completo como anómalo, facilitó la comprensión del contexto y ofreció una visión más clara y accesible de los patrones o comportamientos inusuales, permitiendo distinguir fácilmente días con comportamientos atípicos sostenidos a lo largo del tiempo, en contraste con valores puntuales que podrían confundirse con ruido o eventos aleatorios,

mejorando significativamente la capacidad para evaluar el impacto real de estas irregularidades climaticas. .

- El proceso de detección de anomalías en este trabajo fue mas robusto, en comparación con el llevado a cabo por la DIMAR, se trató de un enfoque hibrido y multivariable que permitió identificar tres tipos de anomalías relacionadas con fallos en sensores y la deteccion de días anómalos que determinan cuándo un día merece una revisión. A diferencia de la DIMAR, que no se enfoca específicamente en la detección de anomalías, sino en la clasificación de calidad de los datos, proporcionando únicamente tres etiquetas de calidad, entre las cuales se incluye solamente una etiqueta de error, basada principalmente en el sobrepaso de los límites de medición de los sensores. A pesar de esta limitación,para dar cumplimiento al ultimo objetivo, dicho proceso sirvió como referencia inicial para evaluar la primera etapa de detección en nuestro trabajo, destacando las ventajas de un enfoque más detallado y orientado a la identificación de anomalías.

- La experiencia y el asesoramiento del equipo de expertos de DIMAR fueron clave para este trabajo. Su conocimiento sobre variables meteorológicas, fenómenos climáticos y procesos de control de calidad en las mediciones tuvo impacto en el proceso de preparación y modelamiento de las series,permitiendo comprender mejor los datos y contextualizar las posibles anomalías en las series temporales analizadas.

Bibliografía

[1] D. H. Moyano and S. G. Montes, *Caracterización meteomarina anual del Caribe Colombiano 2021*, Área de Comunicaciones Estratégicas Acoes, Ed., Cartagena, Bolívar, Colombia, 2022, disponible en: https://www.dimar.mil.co.

[2] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," in *Presented at the European Commission ESPRIT Program.* Ulm, Germany and Tübingen, Germany: DaimlerChrysler Research & Technology and Wilhelm-Schickard-Institute, 2000, project number 24959, partly sponsored by the European Commission under the ESPRIT program.

[3] M. de Ambiente y Desarrollo Sostenible, "Impacto del cambio climático en colombia," Ministerio de Ambiente y Desarrollo Sostenible, 2025, accedido el 27 de febrero de 2025. [En línea]. Disponible: https://archivo.minambiente.gov.co/index.php/cambio-climatico/que-es-cambio-climatico/impacto-del-cambio-climatico-en-colombia.

[4] Dirección General Marítima de Colombia, *Manual Metodológico: Información Oceanográfica y de Meteorología Marina*, 2023.

[5] D. G. Marítima, "¿qué es dimar? - misión y visión," Available: https://www.dimar.mil.co/, 2023, [Online].

[6] S. Hastenrath, "The intertropical convergence zone of the eastern pacific revisited," *International Journal of Climatology*, vol. 22, pp. 347–356, 2002.

[7] Dirección General Marítima (Dimar), *Manual de calidad de datos de estaciones meteorológicas automáticas satelitales*, versión 1 ed. Bogotá D.C., Colombia: Editorial Dimar, 2023. [Online]. Available: https://doi.org/10.26640/cecoldo.general_00004

[8] J. D. Cryer and K.-S. Chan, *Time Series Analysis With Applications in R.* Springer, 2008.

[9] G. Dudek, "Std: A seasonal-trend-dispersion decomposition of time series," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 10, pp. 10 339–10 350, 2023.

[10] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 2nd ed. New York, NY, USA: Springer-Verlag, 2002, springer Texts in Statistics.

[11] R. Foorthuis, "On the nature and types of anomalies: a review of deviations in data," *International Journal of Data Science and Analytics*, vol. 12, p. 297–331, 2021.

[12] K. Ni, N. Ramanathan, M. N. H. Chehade, L. Balzano, S. Nair, S. Zahedi, E. Kohler, G. Pottie, M. Hansen, and M. Srivastava, "Sensor network data fault types," *ACM Transactions on Sensor Networks*, vol. 5, no. 3, p. Article 25, 2009. [Online]. Available: http://doi.acm.org/10.1145/1525856.1525863

[13] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. Article 15, 58 pages, 2009. [Online]. Available: http://doi.acm.org/10.1145/1541880.1541882

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD*, vol. 96, 1996, pp. 226–231.

[15] M. Müller, "Dynamic time warping," *Information Retrieval for Music and Motion*, vol. 2, pp. 69–84, 2007.

[16] F. Ahmad, R. V. Babu, and N. Mazzocca, "A novel k-means clustering algorithm based on genetic algorithm for time series anomaly detection," *Applied Sciences*, vol. 11, no. 15, p. 6889, 2021.

[17] H.-P. Kriegel, E. Schubert, and A. Dorer, "Density-based clustering: A review of recent advances," in *Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM)*, 2020, pp. 473–482.

[18] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[19] S. Petrovic, "A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters," *Proceedings of the 11th Nordic Workshop on Secure IT Systems*, vol. 2006, pp. 53–64, 2006.

[20] N. D. Buono *et al.*, "Detecting anomalies in marine data: A framework for time series analysis," in *Machine Learning, Optimization, and Data Science*, ser. Lecture Notes in Computer Science, G. N. et al., Ed. Cham: Springer, 2023, vol. 13810, p. 36.

[21] C. Velasco-Gallego and I. Lazakis, "Radis: A real-time anomaly detection intelligent system for fault diagnosis of marine machinery," *Expert Systems with Applications*, vol. 204, p. 117634, October 2022. [Online]. Available: https://doi.org/10.1016/j.eswa.2022.117634

[22] Y. Wang, L. Han, W. Liu, S. Yang, and Y. Gao, "Study on wavelet neural network based anomaly detection in ocean observing data series," *Ocean Engineering*, vol. 186, p. 106129, August 2019. [Online]. Available: https://doi.org/10.1016/j.oceaneng.2019.106129

[23] C. C. Vargas, J. Quintero-Ibáñez, and A. Solanilla, "Aplicación de ciencia de datos para la reconstrucción de series de tiempo de variables meteorológicas en islas del rosario (caribe colombiano) entre los años 2013–2021," *Boletín Científico CIOH*, vol. 41, no. 2, pp. 67–80, 2022. [Online]. Available: https://doi.org/10.26640/22159045.2022.604

A Anexo: Descomposición de series

A continuación, se muestran las figuras correspondientes a las descomposiciones de las series temporales.

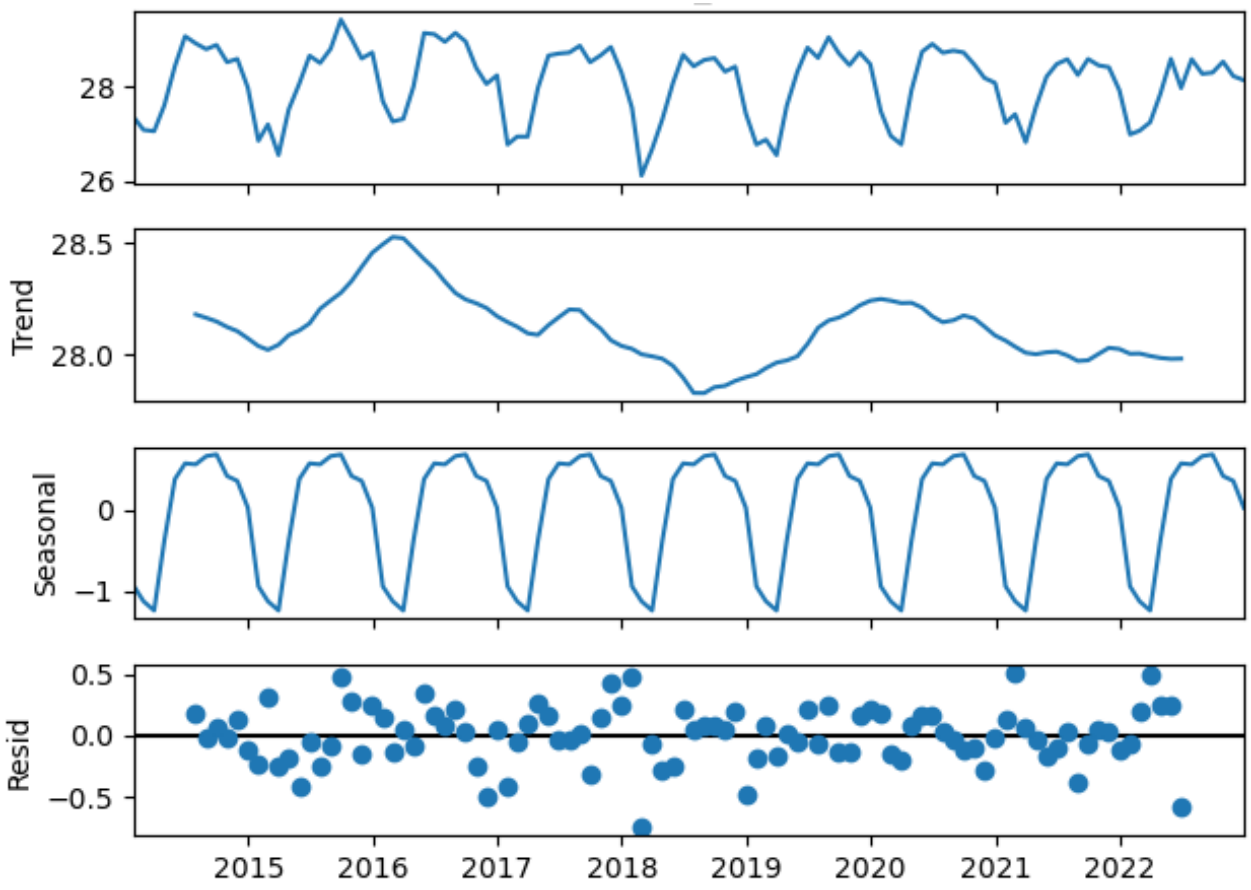


Figura A.1: Descomposición de la Temperatura del aire

Figura A.2: Descomposición de la precipitación

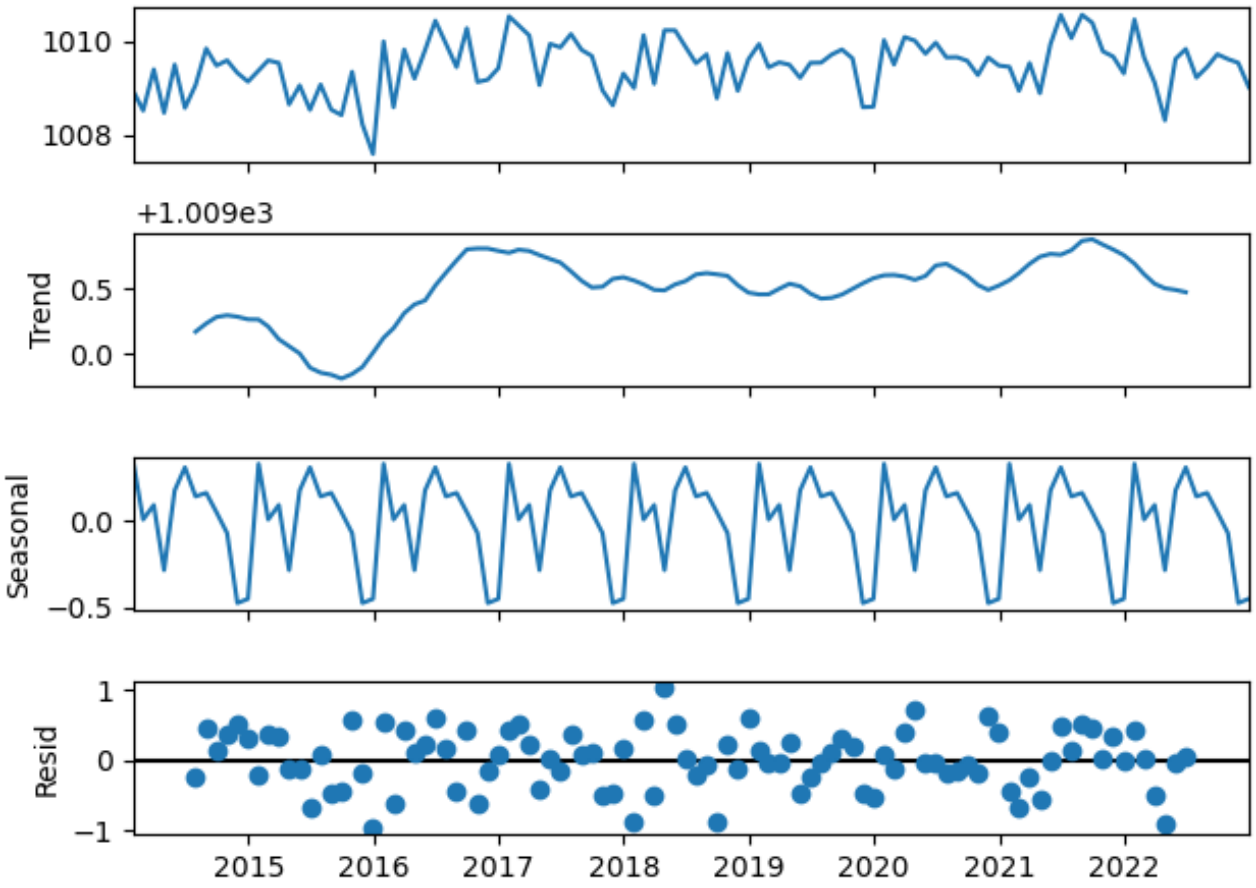


Figura A.3: Descomposición de la Presión atmosférica

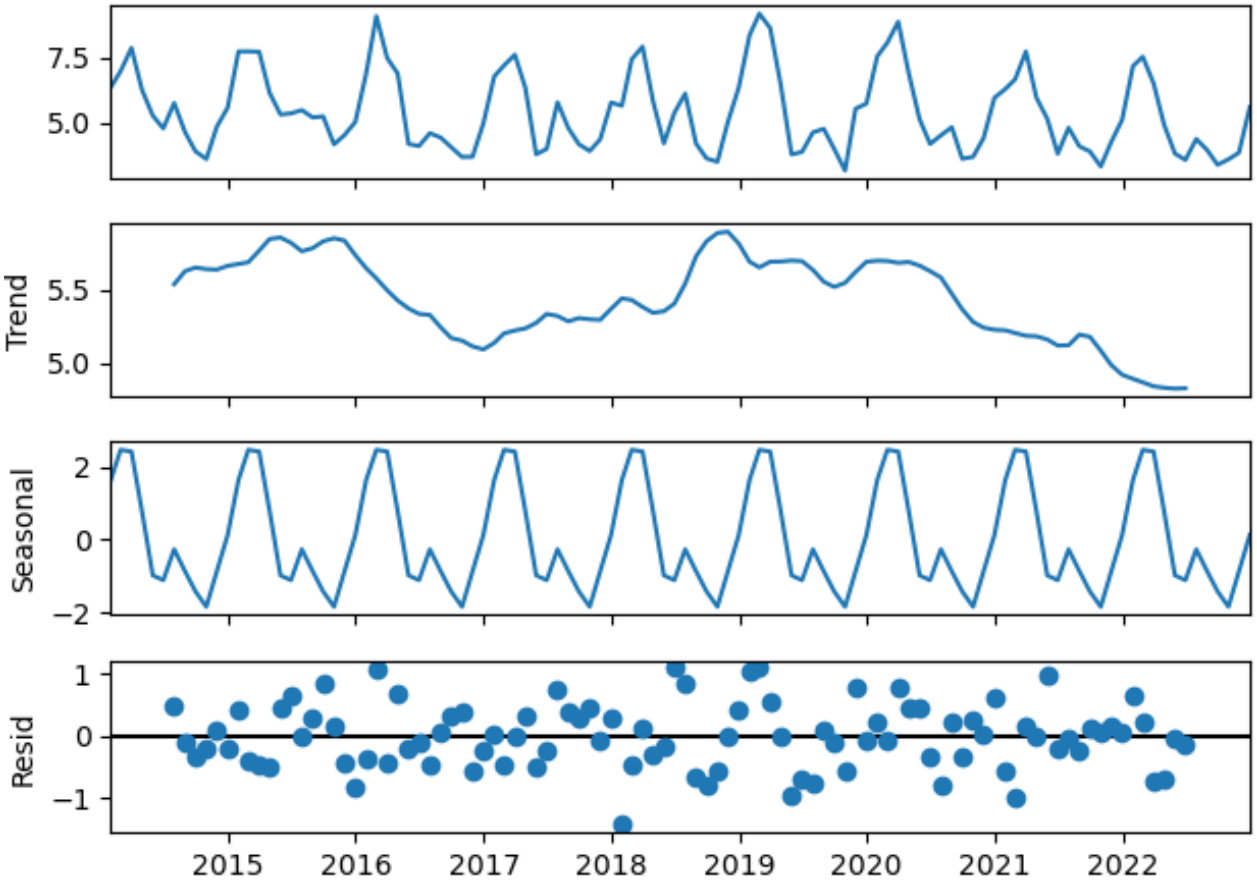


Figura A.4: Descomposición de la Velocidad del viento

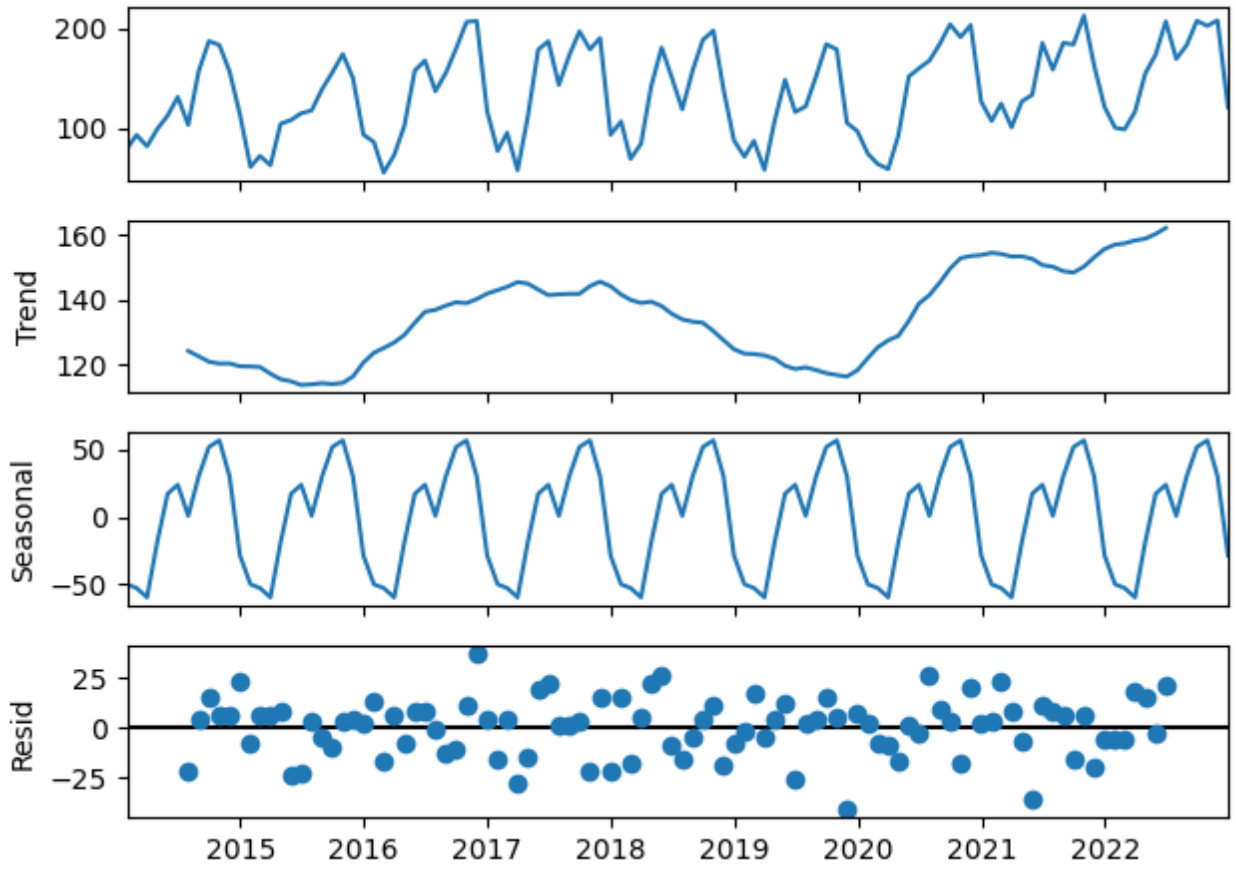


Figura A.5: Descomposición de la TDirección del viento

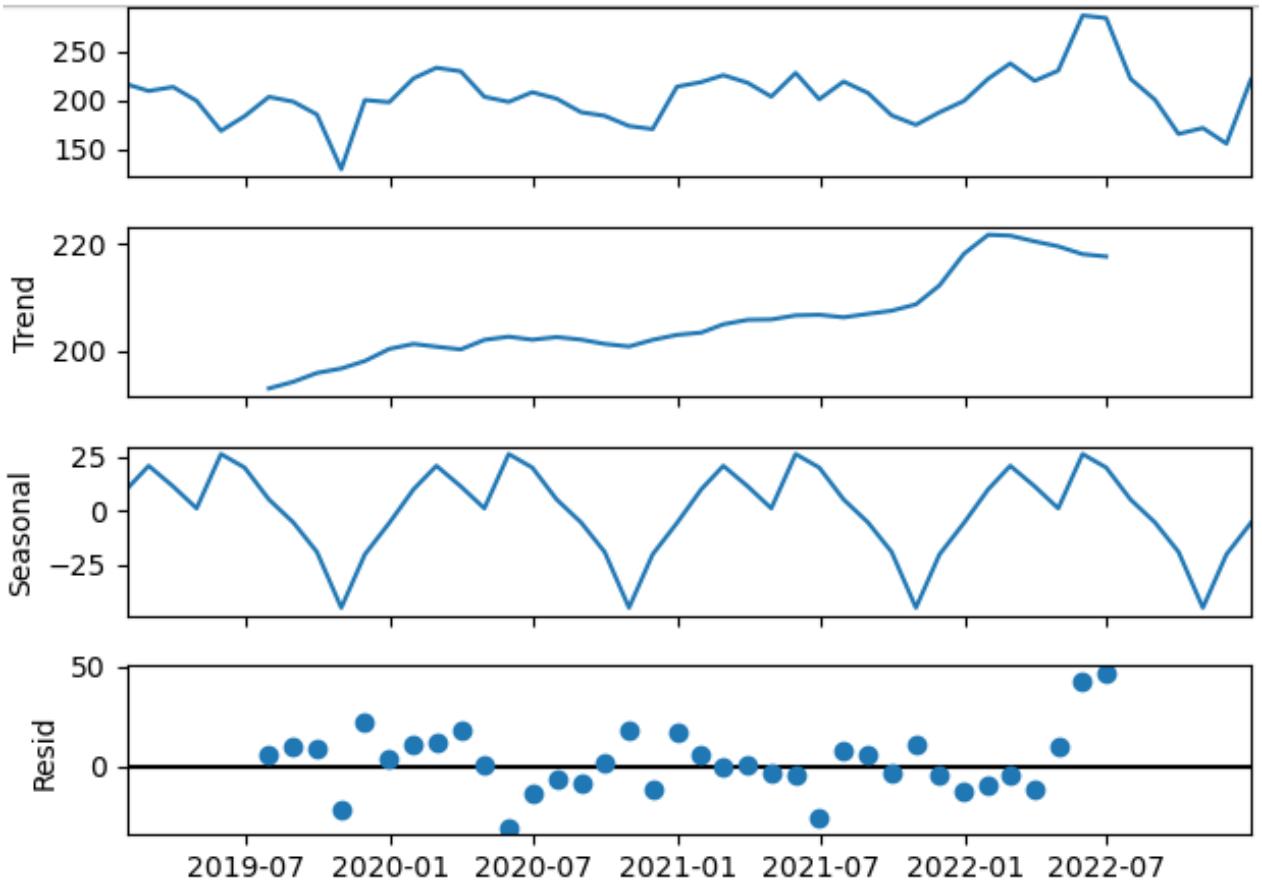


Figura A.6: Descomposición de la Radiación solar

B Guía de uso

B.1 Requisitos previos

1. Las series temporales deben contar con una marca de tiempo con frecuencia horaria para garantizar el correcto funcionamiento del algoritmo.

2. Cada variable debe estar representada por una única serie temporal que combine los datos de todos los años, manteniendo la temporalidad completa. No se deben presentar múltiples series separadas por año; todas deben consolidarse en una sola serie por variable.

B.2 Instalación del entorno

B.2.1 Instalación de Python

1. Descargar la versión más reciente de Python desde el sitio oficial: `https://www.python.org/downloads/`

2. Durante la instalación, asegurarse de seleccionar la opción "Add Python to PATH" antes de hacer clic en "Install Now" para facilitar el acceso a Python desde la terminal.

3. Verificar la instalación abriendo una terminal y ejecutando:
`python -version` o `python3 -version`

B.2.2 Instalación de Jupyter Notebook

1. Abrir una terminal (CMD, PowerShell, o Terminal en macOS/Linux).

2. Instalar Jupyter Notebook ejecutando: `pip install notebook`

3. Verificar la instalación con: `jupyter notebook -version`

B.2.3 Instalación de las librerías necesarias

1. Instalar las librerías requeridas para el algoritmo ejecutando en la terminal: `pip install numpy pandas tslearn scikit-learn seaborn matplotlib plotly statsmodels scipy`

B.3 Configuración inicial

B.3.1 Abrir Jupyter Notebook

1. Inicia Jupyter Notebook desde la terminal ejecutando: `jupyter notebook`

2. Se abrirá una ventana en el navegador. Ahora localizar y abrir el archivo `deteccion_anomalias.ipynb` desde la ubicación donde fue guardado.

B.4 Configuración de la ruta de los datos

1. En el primer bloque de código del archivo `deteccion_anomalias.ipynb`, se debe localizar la variable `direccion`.

2. Asignar la ruta completa donde se encuentran las series temporales. Por ejemplo: `direccion = Ç:/ruta/a/tus/datos/series_temporales.csv"`

B.5 Ejecución del algoritmo

B.5.1 Ejecutar todo el archivo

1. Una vez configurada la variable `direccion`, selecciona la opción **Cell >Run All** (Ejecutar Todo) en la barra de herramientas del notebook.

2. Alternativamente, se pueden ejecutar los bloques uno por uno seleccionándolos y presionando `Shift + Enter`.

B.6 Resultados generados

1. El algoritmo generará gráficos que muestran las anomalías univariables y las detectadas por el enfoque de clusterización.

2. Además, se generará un archivo de salida con las etiquetas asignadas respecto a los dos enfoques.