UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE ESTUDIOS ESTADÍSTICOS

TESIS DOCTORAL

NUEVA METODOLOGÍA PARA IDENTIFICAR PATRONES DE CAMBIO CLIMÁTICO MEDIANTE ANÁLISIS CLÚSTER DE SERIES TEMPORALES CON REDUCCIÓN DE LA DIMENSIONALIDAD

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

ARNOBIO PALACIOS GUTIÉRREZ

DIRECTORES

Dr. JOSÉ LUIS VALENCIA DELFA
Dra. MARÍA VILLETA LÓPEZ

Nueva Metodología para Identificar Patrones de Cambio Climático mediante Análisis Clúster de Series Temporales con Reducción de la Dimensionalidad

Memoria para optar al grado de doctor presentada por

ARNOBIO PALACIOS GUTIÉRREZ

Doctorado en Análisis de Datos (Data Science)

Facultad de Estudios Estadísticos
UNIVERSIDAD COMPLUTENSE DE MADRID

Tesis dirigida por:
Dr. JOSÉ LUIS VALENCIA DELFA & Dra. MARÍA VILLETA LÓPEZ

MADRID, MAYO 2025

En algún lugar, algo increíble está esperando a ser descubierto

Carl Sagan

Si el principal objetivo del capitán fuera preservar su barco, lo mantendría en el puerto para siempre

Tomás de Aquino

Es sencillo hacer que las cosas sean complicadas, pero difícil hacer que sean sencillas

Friedrich Nietzsche

No hay imposibles. Lo que consideramos difícil se resume, tan solo, en una serie de acciones que se encadenan en un orden. Si se abordan con persistencia, permiten lograr ese todo de manera sencilla

Arnobio Palacios Gutiérrez

"Dedicada con mucho amor a mis padres, Ana Gutiérrez y Arnobio Palacios, y a mis hermanos, Liliana, Yuver (Q.E.P.D), Edinson y Sandra Milena, quienes siempre han confiado en mis capacidades y me han motivado hacia el éxito."

Agradecimientos

A lo largo del desarrollo de esta tesis, lo que más destaca son los momentos vividos. Estos innumerables momentos, con el paso del tiempo, se convirtieron en experiencias significativas que no sólo contribuyeron a la consecución de este objetivo, sino que también enriquecieron mi crecimiento personal, académico y profesional. Recuerdo con cariño y gratitud a todas las personas cuyos esfuerzos y apoyo dieron sentido y relevancia a esos momentos entrañables.

Quiero iniciar agradeciendo inmensamente a Dios porque siempre está a mi lado, y todo a su lado es perfecto. Siempre ha sido mi luz y sus infinitas bendiciones siempre me ayudan a lograr mis objetivos.

Reconozco con especial gratitud a mis directores de tesis, los doctores José Luis Valencia y María Villeta, quienes con su compromiso y paciencia me guiaron sabiamente en cada momento. Vuestra indiscutible profesionalidad y experiencia los convierte en unos supervisores ejemplares. De José Luis admiro la manera tan natural para generar estrategias, su capacidad de análisis y su generosidad, y de María la aplicación constante del orden y la lógica, así como el deber de reflexionar siempre sobre lo que se hace. De nuevo, gracias. Seguiré aplicando todas vuestras recomendaciones y espero que sigamos colaborando.

Agradezco de todo corazón a mi madre, Ana Sobeida, mi mayor ejemplo de persistencia y superación, a mi padre Arnobio y a mis hermanos Liliana, Yuver (Q.E.P.D), Edinson y Sandra Milena, quienes siempre han estado a mi lado y a quienes les dedico esta tesis. Ustedes siempre me han guiado por un buen camino, me han inculcado valores muy valiosos y han contribuido significativamente en mi formación integral, apoyándome incondicionalmente en cada momento. A pesar de la distancia, durante esta estancia, siempre estuvimos unidos, estuvieron pendientes de mí y de mi progreso, gracias por ser parte de mi vida y por todo.

A mis sobrinos, Juan Pablo, Eris Andrés, Juan Smith, Heylin Adriana, Dilson Adrián y Frank Stith, gracias por sus emotivos mensajes llenos de entusiasmo y buena energía, gracias por esas videollamadas divertidas y alentadoras, los quiero mucho; así, como ustedes me dicen “tío lo quiero mucho”.

A mi abuela, María Ernelinda (Q.E.P.D), quien siempre creyó en mí y me apoyó en todo lo que pudo. Siempre te tengo presente, a ti y a tus enseñanzas.

Un reconocimiento especial para mi mentora, la doctora Leidy Indira Hinestroza, quien me introdujo en el mundo de la investigación y moldeó mis primeros pasos. Desde que la conocí ha sido un apoyo incondicional para mí, sus enseñanzas y consejos han sido de gran valor para mi crecimiento. Gracias por su generosidad y sabiduría compartida, la quiero mucho.

Al Ministerio de Ciencia, Tecnología e Innovación de Colombia que, al brindarme esta oportunidad de transformación educativa mediante la asignación de una beca, contribuyó a mi formación, y espero que, con esto, a la de muchas personas.

A la Universidad Tecnológica del Chocó – Diego Luis Córdoba, por apoyar esta capacitación y aportar facilidad a los procesos para mi formación académica. Reconozco de manera especial, el apoyo brindado por el exrector David Emilio Mosquera Valencia, al haber confiado en mí y en mis capacidades, considerándome para obtener una comisión especial de estudios.

Expreso mi agradecimiento al Servicio de Desarrollos Climatológicos de la Agencia Estatal de Meteorología de España por la generación y distribución de los datos que hemos utilizado en este estudio, conjuntos de datos sólidos y bien estructurados.

A mis compañeras de doctorado, Gabriela y Milagros, por su cercanía, consejos y recomendaciones, todo ello de gran valor durante muchos momentos de esta formación. En mí tienen un amigo y espero que sigamos coincidiendo.

A todas esas personas que contribuyeron para que mi estancia en España fuera lo más agradable posible; desde los Matemáticos, George, Luis, Javi y Paula; desde los Outsiders, Farid, María Paola, Geovanny, Alexa, Esteban, Carolina, Onel y Julio; desde la Resi, Karen Jineth, Víctor Danilo, Dahiana, Ada, Milena, Sofia, Jessica, Luisa, Carolina y Jeffrey; Miguel Ángel quien siempre me abrió las puertas de su casa en Valencia; Johanna la Vice; Angie y Sirley; Pamela, Natalie, Julissa y Álvaro; y entre muchos más, los Magos del Balón. Gracias por todo, todos los momentos compartidos fueron muy valiosos, y saben que cuentan conmigo.

También agradezco con mucho aprecio a esas personas y amigos que, desde Colombia y otros países, siempre me hicieron saber que contaba con ellos y me enviaron mucho ánimo y apoyo emocional. Los estimo mucho, Frank, Heiler, Adriana, Silvia del Carmen, Emiliana, Alix del Carmen e hijas, Delfa María, Melisa, Anarist, Miladis, Regina, Lizeth, Escolástica del Rosario, Elisa Adriana, Angela Consuelo, Kelly, Fidel, Noretty, Pedroza, Carlos Alberto, David Fernando, Carlos Hernán, James Jeiser, John Fredy y Diany Yiseth, entre muchos más. Gracias por los buenos deseos y palabras de aliento.

Resumen

El clima y sus transformaciones constituyen un sistema dinámico de alta complejidad, en el que los niveles y variaciones de sus variables pueden presentar comportamientos significativamente heterogéneos a escala regional, incluso dentro de un mismo dominio geográfico. A pesar de la abundancia de estudios orientados al análisis del sistema climático, la mayoría han adoptado enfoques agregados que, al no considerar explícitamente la similitud de los patrones temporales, limitan la identificación de dinámicas climáticas regionales específicas.

Esta tesis se orienta a la caracterización e identificación de patrones climáticos regionales mediante la integración conceptual y metodológica de tres ejes fundamentales: cambio climático, regionalización y agrupamiento de series temporales. Este enfoque permite revelar estructuras latentes que permanecen ocultas en los análisis convencionales de tipo agregado.

La tesis se estructura a partir de tres artículos publicados en revistas indexadas en el Journal Citation Reports (JCR), en los cuales se proponen distintos procedimientos para el agrupamiento de series temporales climáticas, fundamentados en técnicas de reducción de dimensionalidad.

En una primera etapa de investigación, se diseñó un procedimiento para el agrupamiento de series temporales univariantes, que incorpora el uso del Análisis de Espectro Singular (SSA, por sus siglas en inglés) como técnica de preprocesamiento, permitiendo la extracción de componentes relevantes y la atenuación del ruido estructural.

Posteriormente, se desarrolló un método para el agrupamiento de series temporales multivariantes, basado en un enfoque de análisis multiescala que incorpora tanto la estructura secuencial temporal en la construcción de los vectores de características como la selección de atributos pertinentes desde una perspectiva informativa.

En una tercera fase, se propuso una metodología de regionalización climática que incorpora estimaciones proyectadas de cambio climático como variables exógenas en el proceso de agrupamiento, permitiendo además cuantificar la contribución relativa de variables geográficas a las dinámicas del cambio climático. Esta metodología emplea un modelo de Análisis Discriminante de Mínimos Cuadrados Parciales (PLS-DA, por sus siglas en inglés) de forma recursiva, integrando variables no climáticas con el objetivo de optimizar las agrupaciones mediante un índice sintético que equilibra métricas clave del desempeño del modelo.

En términos generales, los resultados obtenidos a lo largo de esta investigación evidencian la relevancia de considerar la interacción entre cambio climático, regionalización y análisis de series temporales para una comprensión más precisa de los sistemas climáticos. Los procedimientos metodológicos desarrollados constituyen un aporte significativo tanto al cuerpo teórico como al aplicado, y proporcionan un marco analítico robusto para el diseño de estrategias de adaptación y mitigación frente al cambio climático.

Abstract

Climate and its transformations constitute a highly complex and dynamic system, in which the levels and variations of climatic variables can exhibit significantly heterogeneous behaviours at the regional scale, even within a single geographic domain. Despite the abundance of studies focused on analysing the climate system, most have adopted aggregated approaches which, by not explicitly accounting for the similarity of temporal patterns, limit the identification of region-specific climatic dynamics.

This thesis is devoted to the characterization and identification of regional climatic patterns through the conceptual and methodological integration of three fundamental components: climate change, regionalization, and time series clustering. This approach enables the unveiling of latent structures that remain hidden in conventional aggregate analyses.

The thesis is structured around three articles published in journals indexed in the Journal Citation Reports (JCR), each of which proposes different procedures for clustering climatic time series, grounded in dimensionality reduction techniques.

In the first stage of research, a procedure for clustering univariate time series was designed, incorporating the use of Singular Spectrum Analysis (SSA) as a preprocessing technique. This allowed for the extraction of relevant components and the reduction of structural noise.

Subsequently, a method for clustering multivariate time series was developed, based on a multiscale analysis approach that considers both the sequential temporal structure in feature vector construction and the selection of informative attributes.

In the third stage, a climatic regionalization methodology was proposed, incorporating projected estimates of climate change as exogenous variables in the clustering process. This approach also enables the quantification of the relative contribution of geographic variables to climate change dynamics. The methodology employs a Partial Least Squares Discriminant Analysis (PLS-DA) model in a recursive manner, integrating non-climatic variables with the aim of optimizing the clustering outcomes through a synthetic index that balances key performance metrics.

Overall, the results of this research highlight the importance of considering the interaction between climate change, regionalization, and time series analysis in achieving a more accurate understanding of climate systems. The methodological procedures developed herein represent a significant contribution to both theoretical and applied domains, offering a robust analytical framework to support the design of climate change adaptation and mitigation strategies.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABLAS

Preámbulo

El cambio climático representa uno de los retos globales más apremiantes del siglo XXI, que afecta a la sociedad y a los ecosistemas. Dada su complejidad, comprender su dinámica y efectos específicos a nivel regional, es esencial para desarrollar estrategias eficaces de adaptación y mitigación. En este contexto y dado el potencial del aprendizaje no supervisado para identificar patrones ocultos, esta investigación se centra en el desarrollo de metodologías para detectar patrones de cambio climático a nivel regional, dada la importancia que tienen la agrupación de series temporales y la climatología en la regionalización del cambio climático, y que puede proporcionar herramientas valiosas para la toma de decisiones.

Entre otros aspectos, esta investigación surge de la necesidad de reducir la brecha entre los análisis climáticos agregados y la realidad de los fenómenos a escala local. Si bien es cierto que se ha avanzado en gran medida en la modelización de los sistemas climáticos amplios, gran parte de los estudios no captan con suficiente precisión las variaciones regionales finas, esenciales para abordar los impactos específicos del cambio climático. Aprovechando la agrupación de series temporales, esta tesis busca descubrir patrones de evolución temporal de variables climáticas, proporcionando una comprensión matizada de cómo varía espacialmente su dinámica. Para ello, esta tesis emplea enfoques avanzados como el SSA, el Análisis Multiescala y el PLS-DA. Estas metodologías aportan herramientas innovadoras en los procesos de agrupamiento, lo que contribuye a una mejor comprensión del fenómeno y sus implicaciones.

La finalidad de esta investigación, además de identificar patrones específicos de cambio climático, pretende desarrollar procedimientos para la modelización de sistemas climáticos que capturen variaciones regionales. Estos avances pueden contribuir a reducir las implicaciones del fenómeno en la sociedad y los ecosistemas. Los resultados buscan aportar al debate científico en el ámbito del análisis de datos y la climatología, al tiempo que proporcionan instrumentos prácticos para generar políticas específicas encaminadas a reducir los efectos del cambio climático para cada región.

El desarrollo de esta tesis se ha llevado a cabo en la Facultad de Estudios Estadísticos de la UCM, en el marco del Programa de Doctorado “Análisis de Datos (Data Science)”. Bajo la orientación de mis directores de tesis, se ha garantizado el rigor estadístico de los métodos de análisis que se han empleado para alcanzar los objetivos planteados. Asimismo, esta investigación también se realiza con el apoyo de una beca del Ministerio de Ciencias de Colombia y de la Universidad Tecnológica del Chocó, instituciones que han confiado en este trabajo brindando su respaldo.

Capítulo 1

Capítulo 1

Introducción

Resumen: *Este capítulo introduce los fundamentos de la investigación, explorando la interrelación entre cambio climático, regionalización y clustering de series temporales. Se destaca cómo estos tres conceptos se complementan para identificar y analizar patrones de cambio climático a nivel regional, subrayando la importancia de considerar las diferencias espaciales y la necesidad de herramientas analíticas precisas para interpretar las series temporales climáticas. Asimismo, se presentan los objetivos de la tesis y las contribuciones científicas derivadas de su desarrollo.*

1.1. Antecedentes y Motivación

En las últimas décadas, el debate científico sobre el cambio climático ha cobrado gran relevancia, en parte debido al predominio del escepticismo y la negación, sustentados en distintas percepciones e hipótesis sobre su existencia. Inicialmente, las estrategias negacionistas se enfocaban en rechazar por completo la realidad del cambio climático, pero esta postura se volvió insostenible ante la creciente evidencia científica (Van Valkengoed et al., 2021). Posteriormente, la discusión evolucionó hacia cuestiones sobre sus causas y efectos, incluyendo el grado de responsabilidad de la actividad humana y la magnitud de sus consecuencias (Van Valkengoed et al., 2021). En este contexto, una definición fundamental sobre este fenómeno es la del Panel Intergubernamental sobre el Cambio Climático (IPCC), que describe el **cambio climático** como una alteración identificable en el estado del clima, persistente por décadas o más, reflejada en variaciones en la media y/o la variabilidad de sus propiedades, ya sea debido a la variabilidad natural o a la influencia humana (IPCC, 2022).

En la actualidad, uno de los retos más importantes de la sociedad es evaluar los impactos del cambio climático, los cuales a medida que pasa el tiempo son más notables en las distintas regiones del planeta. Las variaciones de los patrones climáticos pueden generar impactos significativos en diversos sectores, como la seguridad alimentaria, la salud humana, la agricultura, los recursos hídricos y el suministro energético, entre otros. Según estudios científicos (de Lucena et al., 2009; O'neill & Ebi, 2009; Choularton et al., 2012; Schaeffer et al., 2012; Sathaye et al., 2013; Brown et al., 2015; Khan et al., 2021; Alsalal et al., 2024), estos efectos pueden estar asociados a cambios en la temperatura, la alteración de los regímenes de precipitaciones y la variabilidad climática a corto y largo plazo. La evidencia recopilada por organismos internacionales, como el IPCC, indica que estos factores pueden incrementar la vulnerabilidad de ciertos sistemas y poblaciones.

A medida que los fenómenos meteorológicos y climáticos se hacen más erráticos y extremos, comprender sus variaciones es crucial para desarrollar estrategias de adaptación conformes a los contextos regionales, puesto que a escala regional, estos problemas y sus impactos pueden variar considerablemente debido a la influencia de factores físicos como la topografía, la distribución de la vegetación, las urbanizaciones y las masas de agua, que afectan el clima de la superficie (Chen et al., 2006; Teodoro et al., 2021). En este contexto, la **regionalización** surge como una herramienta vital, que no solo facilita la identificación de patrones espaciales y temporales en las variaciones climáticas, sino que también permite detectar cambios significativos con mayor precisión (Laepple & Huybers, 2014).

En general, para comprender realmente la dinámica y los patrones temporales del clima en diferentes regiones de un territorio, se recurre al potente método analítico de la **agrupación de series temporales**, en ocasiones también mencionado como clustering de series temporales. Uno de los métodos multivariantes de regionalización climática más utilizados (Li et al., 2022). Esta metodología permite identificar patrones recurrentes en conjuntos de datos climáticos, facilitando la

caracterización de regiones con comportamientos similares (Aghabozorgi et al., 2015). De modo que, al agrupar patrones temporales similares, la agrupación de series temporales ofrece una visión detallada de los comportamientos climáticos únicos de regiones distintas, permitiendo desvelar perspectivas que de otro modo quedarían oscurecidas en los análisis climáticos agregados.

La intersección entre estos tres aspectos tan interesantes permite hacer contribuciones clave tales como:

- Identificación y análisis de tendencias y anomalías climáticas específicas, posibilitando un análisis más detallado de los cambios regionales.
- Desarrollo de estrategias de adaptación más específicas y eficaces adaptadas a los comportamientos climáticos particulares de cada región.
- Mejora de la precisión en los modelos predictivos al identificar regiones con patrones climáticos históricos similares.
- Fomento de la investigación interdisciplinar al tender un puente entre la climatología, la ciencia de los datos y la planificación regional.

Por tanto, para la identificación de patrones de cambio climático en un contexto de regionalización, el principal interés se centra precisamente en la agrupación de series temporales de información climática, uno de los retos actuales de la minería de datos. Una serie temporal se clasifica como datos dinámicos, ya que sus características se organizan cronológicamente. Al querer agrupar series de tiempo, realmente queremos agrupar objetos complejos, pues este desafío supone pensar en la alta dimensionalidad, el orden temporal y el ruido que, por definición, es impredecible y representa las fluctuaciones aleatorias que no siguen un patrón especifico. Estos tres factores constituyen el mayor problema de la agrupación de series temporales (Aghabozorgi et al., 2015), y conllevan un compromiso entre la velocidad en el desempeño de los algoritmos y la calidad de los resultados.

Las revisiones más impactantes (Warren Liao, 2005; Rani & Sikka, 2012; Aghabozorgi et al., 2015; Alqahtani et al., 2021; Ergüner Özkoç, 2021) realizadas sobre la agrupación de las series temporales, coinciden en que principalmente, los diferentes trabajos e investigaciones adelantadas para abordar este asunto se basan en una de tres categorías o enfoques existentes para agrupar series temporales:

- Métodos basados en los datos brutos, que trabajan directamente con las series temporales sin transformación previa.
- Métodos basados en características relevantes extraídas de los datos brutos antes de aplicar los algoritmos de agrupamiento.
- Métodos basados en modelos matemáticos construidos a partir de datos brutos.

De acuerdo con Aghabozorgi et al. (2015), los enfoques de agrupación de series temporales basados en características y en modelos son clave para la reducción de la dimensionalidad, aspecto fundamental en el proceso de agrupamiento de series temporales. En esta tesis se emplean técnicas de representación de datos para

abordar desafíos como la alta dimensionalidad, el orden temporal y el ruido de las series temporales climáticas.

Para identificar y analizar variaciones climáticas locales mediante la regionalización de patrones climáticos, diversos estudios (Hsu & Li, 2010; Roushangar & Alizadeh, 2018; Sharghi et al., 2018; Liu et al., 2019; Li et al., 2022) han implementado con éxito técnicas de representación de series temporales aplicadas a datos climáticos. Aunque este enfoque ha demostrado ser efectivo, también presenta ciertas limitaciones en los procesos de agrupamiento. La mayoría de los trabajos utilizan índices basados en valores medios, acumulados por períodos, tendencias, cambios de variación u otros derivados de ellos (P. Shi et al., 2016; Roushangar & Alizadeh, 2018; Ilbay-Yupa et al., 2021; Abbasi et al., 2022). Sin embargo, estos índices pueden verse influenciados por valores extremos, dependencia de los datos o la selección de variables. Además, los estudios climáticos suelen centrarse en tendencias climáticas promedio, lo que dificulta identificar cambios inusuales (Gebremichael et al., 2022), afectando significativamente los resultados del agrupamiento. Asimismo, la falta de parámetros y procedimientos de inferencia limita la capacidad de estos enfoques para explicar las diferencias entre los grupos de forma general. Por ello, incluir técnicas de preprocesamiento que gestionen adecuadamente los componentes intrínsecos de las series temporales, junto con procedimientos de inferencia o modelado tras el preprocesamiento, podría mejorar el rendimiento y la utilidad de los análisis de agrupamiento.

Por otro lado, al considerar que los procesos físicos que influyen en las variables climáticas a menudo operan en un amplio rango de escalas de tiempo (Tessier et al., 1996), se han desarrollado diferentes estudios basados en análisis multiescala. Téngase en cuenta que el término “escala” tiene numerosas definiciones, cuya connotación varía según el campo de estudio. En el contexto de la climatología, se pueden distinguir tres tipos de escala: *escala espacial*, que hace referencia a la región donde se miden y definen los parámetros climáticos; *escala temporal*, que se relaciona con el tiempo característico del proceso investigado, abarcando desde períodos subdiarios hasta escalas de décadas o multidecadales; y *escala topológica*, que describe la disposición de una red, desde nodos individuales hasta su estructura completa (Agarwal et al., 2018). De acuerdo con Agarwal et al. (2017), en esta investigación, la filosofía de un enfoque multiescala se basa en la premisa de que cualquier proceso (en este caso, climatológico) de interés siempre puede representarse a varias escalas temporales con distintas complejidades. Por lo que aquí, con **análisis multiescala** nos referimos a técnicas que examinan patrones en diferentes escalas temporales, es decir; técnicas que se basan en la descomposición de datos en múltiples escalas temporales para identificar patrones y tendencias a corto, mediano y largo plazo. Esto permite obtener agrupaciones de series con comportamientos similares, considerando variaciones en diferentes niveles de granularidad temporal. La mayoría de los estudios en esta línea giran en torno a conceptos como la transformada wavelet y la entropía multiescala (Agarwal et al., 2016; Roushangar & Alizadeh, 2018; Li et al., 2022). Estas metodologías pueden ser complejas de aplicar especialmente para grandes conjuntos de datos y para series

temporales largas. Además, la mayoría de estos enfoques no tienen en cuenta las covariables durante el proceso de agrupamiento y se centran en el análisis de precipitaciones (Agarwal et al., 2016; Sehgal et al., 2018; Gao & Shang, 2019). En consecuencia, la introducción de procedimientos de análisis multiescala eficaces y más sencillos que tengan en cuenta las covariables para el agrupamiento, pueden ser de utilidad para la regionalización climática.

Por último, teniendo en cuenta que las técnicas de aprendizaje no supervisado, utilizadas en la regionalización climática, requieren de la especificación a priori del número de grupos y que su óptimo se selecciona, generalmente, haciendo uso de índices de validación y considerando en el proceso de agrupamiento únicamente variables climáticas, explorar otras estrategias que permitan optimizar el proceso de agrupamiento teniendo en cuenta el efecto de otras variables no climáticas, resulta de interés para la modelización del cambio climático.

1.2. Alcance

Así, el principal objetivo de esta tesis es el de definir procedimientos para identificar patrones de cambio climático mediante la agrupación de series temporales basándonos en la reducción de la dimensionalidad y abordando las problemáticas indicadas anteriormente. Para ello, desde nuestra investigación, se trabaja para optimizar el proceso de clustering desde la representación, proponiendo la aplicación de técnicas como el SSA. Esta es una metodología bien desarrollada, que mejora la precisión de los resultados de agrupación, gracias a su capacidad para eliminar el ruido de las series, una de sus principales aplicaciones, junto con el análisis de perfiles espectrales (Golyandina & Korobeynikov, 2014). La introducción de la aplicación de SSA en el proceso de agrupamiento de las series climáticas junto con la inclusión de parámetros de tendencia en los vectores de características, permite integrar parámetros estacionales y de autocorrelación estimados mediante el modelado de los componentes que gobiernan la naturaleza de las series. Así mismo, en este estudio se propone también un enfoque de agrupamiento novedoso basado en análisis multiescala más fácil de aplicar que los métodos mencionados anteriormente, y que además de tener en cuenta las precipitaciones, tendrá en cuenta las temperaturas extremas (máximas y mínimas), permitiendo un enfoque de agrupación de series temporales multivariantes, fácilmente adaptable a otras variables climáticas, siempre que se realice una selección cuidadosa de los parámetros o características relevantes en distintas escalas para cada tipo de variable. Por ejemplo, considerando la asimetría de la distribución de las temperaturas y de las precipitaciones, en nuestras aplicaciones resulta más factible tener en cuenta medianas en lugar de medias para la construcción de vectores de características, ya que el estimador de la mediana refleja una representación más robusta de estas variables climáticas que el estimador de la media.

Este método se ha propuesto a partir del enfoque de análisis multiescala de series financieras univariantes de Shi et al. (2021), el cual se adapta en esta tesis para definir una medida de distancia entre series temporales multivariantes. Así mismo,

se propone un nuevo enfoque de regionalización climática, en el que se incorporan indicadores cuantificables del cambio climático por cada serie temporal meteorológica como input para el proceso de agrupamiento, y que además tiene en cuenta variables no climáticas (geográficas) mediante un proceso de discriminación que permite cuantificar su impacto en el cambio climático y optimizar el proceso de agrupamiento mediante un índice generado que equilibra las principales métricas del modelo.

En general, nuestros enfoques, a diferencia de la gran mayoría que emplean cantidades reducidas de series temporales, emplean miles de series temporales, además se aplican a series de tiempo asociadas a áreas muy pequeñas, de hasta $5\times5km^2$, lo que constituye una aportación novedosa dentro de la literatura existente.

1.3. Objetivos

Basándose en la evidencia científica que confirma la existencia del cambio climático, este estudio plantea como hipótesis que las variaciones climáticas dentro de un dominio geográfico particular presentan diferencias entre distintas zonas. De modo que, el objetivo principal de la presente investigación es generar procedimientos para identificar patrones de cambio climático mediante procesos de regionalización climática basados en clustering de series temporales con reducción de la dimensionalidad. Este enfoque permite analizar los comportamientos climáticos y establecer diferencias a nivel subregional. Por otro lado, se busca fortalecer las capacidades en el aprendizaje no supervisado en machine learning, fomentando el uso de técnicas de preprocesamiento capaces de manejar la alta dimensionalidad, el ruido y el orden temporal. Además, se pretende optimizar los procesos de agrupamiento, mejorando su precisión y utilidad para la regionalización climática.

A continuación, se presentan los objetivos generales y sus fases específicas que han guiado el desarrollo de esta tesis.

Objetivo I: Desarrollar un procedimiento de agrupamiento de series temporales univariadas basado en la extracción de características de tendencia, estacionalidad y ruido mediante SSA, facilitando la reducción de dimensionalidad para la identificación de patrones de cambio en la temperatura máxima.

En esta fase del estudio se propone un método para agrupar series temporales univariantes, lo que permitirá:

1) Descomposición mediante SSA para extraer componentes de las series temporales.

2) Modelado de las componentes extraídas para definir los vectores de características de cada serie temporal.

3) Definición de prototipos de clústeres y análisis de patrones de temperaturas máximas.

Objetivo II: Analizar la evolución de temperaturas extremas y precipitaciones, identificando regiones climáticas diferenciadas y desarrollando un procedimiento de agrupación de series temporales multivariantes que integre múltiples escalas de tiempo y preserve el orden temporal en la reducción de dimensionalidad.

Esta fase de agrupamiento de series temporales multivariantes permitirá lo siguiente:

4) Integración de varias escalas temporales en los procesos de reducción de la dimensionalidad de series temporales.

5) Preservación del orden temporal en la construcción de los vectores de características de las series.

6) Definición de una distancia para series temporales multivariantes que tenga en cuenta diferentes escalas temporales en la medida de las similitudes.

7) Análisis de patrones de temperaturas máximas, mínimas y de precipitaciones.

Objetivo III: Proponer un método de regionalización climática basado en la evaluación del impacto de variables geográficas sobre los patrones de cambio climático, incorporando de manera recursiva factores no climáticos en el agrupamiento de series temporales para optimizar la regionalización.

En esta fase, en la que se incorporan de forma recursiva variables no climáticas al agrupamiento de series temporales, se permitirá:

8) Incorporación de indicadores cuantificables del cambio climático por cada serie temporal como input para la agrupamiento.

9) Obtención de diferentes configuraciones de clustering en base a la definición de una distancia para series temporales multivariantes que permita asignar diferentes pesos a las variables del modelo.

10) Uso de PLS-DA para describir cualitativamente los clústeres y analizar la influencia de las variables geográficas.

11) Evaluación del impacto de variables geográficas en los patrones climáticos.

12) Desarrollo de un nuevo índice que optimice el agrupamiento equilibrando métricas clave del modelo.

13) Aplicación a variaciones de temperaturas y precipitación.

En general, aplicamos distintos algoritmos de clustering convencionales en cada procedimiento. Dado que la agrupación de series temporales permite reconocer cambios dinámicos mediante la detección de correlaciones entre las series de tiempo, así como identificar los patrones que las gobiernan (Aghabozorgi et al., 2015), una vez obtenidas las agrupaciones finales, se definieron los prototipos de

los clústeres resultantes y se analizaron sus diferencias climáticas. Este análisis permitió interpretar la evolución y variabilidad del clima en diversas áreas de las regiones estudiadas, considerando la representación de las series temporales y distintos aspectos clave, como índices y técnicas de análisis de información climática, relevantes para describir los patrones de cambio.

Para la aplicación de los métodos propuestos, empleamos dos bases de datos elaboradas por el Servicio de Desarrollo Climatológico de la Agencia Estatal de Meteorología de España (AEMET). La primera, utilizada en la etapa inicial, fue obtenida mediante interpolación espacial por kriging a partir de todos los datos termo-pluviométricos de la Base de Datos Climática Histórica de la AEMET. Se señala que el número de observaciones disponibles no es constante a lo largo del tiempo, ya que depende de la fecha. Sin embargo, dentro del rango temporal considerado, las observaciones termométricas oscilaron entre aproximadamente 150 y 2000 (Luna et al., 2008). La base de datos generada contiene 1,776 series temporales registradas entre 1931 y 2009, distribuidas en la Península Ibérica con una resolución de $25\times25\text{km}^2$. La segunda base de datos, empleada en las siguientes etapas del estudio, fue generada mediante interpolación óptima a partir de todas las observaciones disponibles en el Banco Nacional de Datos de AEMET (1,800 estaciones termométricas y 3,236 pluviométricas). Esta base incluye 16,156 series temporales registradas entre 1951 y 2021, distribuidas en España con una resolución de $5\times5\text{km}^2$.

Como indica la AEMET, la interpolación estadística es un método de estimación lineal que maneja adecuadamente distribuciones irregulares de observaciones, minimizando el error de interpolación mediante una correcta formulación de las estadísticas de error que sustentan la construcción del campo analizado. La interpolación espacial por kriging es un método geoestadístico que estima valores en ubicaciones no muestreadas a partir de datos conocidos, considerando la correlación espacial entre los puntos (Oliver & Webster, 1990). Este método asume que los puntos cercanos presentan valores más similares que los distantes y, para una estimación óptima, minimiza el error de predicción asignando pesos a los datos conocidos. Para más detalles sobre el método, véase Oliver & Webster (1990). Por su parte, la interpolación óptima, también denominada interpolación estadística debido a las aproximaciones e hipótesis asumidas en su implementación, permite interpolar observaciones meteorológicas en una rejilla de análisis, ponderando los datos según la distancia al punto de rejilla y las características del error. Además, emplea funciones de estructura isotrópicas para reducir el error de interpolación. Más detalles sobre su algoritmo pueden consultarse en la sección 2.2 del documento Nota Técnica 24 de AEMET.

1.4. Organización de la tesis

La tesis está organizada en 7 capítulos y la metodología utilizada se describe en los capítulos correspondientes.

En el capítulo 2, se exponen los principales conceptos sobre la agrupación de series temporales y la reducción de la dimensionalidad. Asimismo, se definen los

algoritmos de clustering utilizados junto con los índices de validación. Las técnicas de preprocesamiento empleadas en los procesos de agrupamiento se indican en cada capítulo, al igual que las técnicas de análisis climático utilizadas para comprender los patrones de cambio climático.

El capítulo 3 aborda un procedimiento definido para la agrupación de series temporales univariantes de información climática, haciendo uso del SSA en el proceso de descomposición de las series, se aplican modelos de regresión para modelar sus componentes y se construyen vectores característicos para la representación y agrupación, con un enfoque en el análisis de patrones de temperatura máxima en la Península Ibérica.

El capítulo 4 da paso a la agrupación de series temporales multivariantes, detallando un enfoque de análisis multiescala que da especial importancia al orden temporal de las características de las series por escalas y que permite analizar en profundidad patrones de temperaturas extremas y de precipitaciones en España teniendo en cuenta diferentes índices climáticos. De igual forma, se detalla la distancia propuesta entre series multivariantes en múltiples escalas.

En el capítulo 5, se presenta un nuevo enfoque de regionalización climática que, además de considerar como input medidas cuantificables de los cambios del clima para la agrupación, introduce el uso del PLS-DA. Esta técnica optimiza los resultados de la agrupación permitiendo definir un nuevo índice que equilibra las principales métricas del modelo. En este capítulo, se detalla cómo, mediante la discriminación PLS-DA, este enfoque permite cuantificar el efecto de variables geográficas sobre el cambio climático, empleando diversas configuraciones de clustering obtenidas con varios métodos y con diferentes matrices de distancias.

El capítulo 6 recoge una síntesis y una discusión integradora de los hallazgos obtenidos.

Finalmente, el capítulo7 expone las conclusiones principales de la investigación y plantea posibles líneas de estudio futuras.

Además, se incluye el Anexo I, que presenta algunos de los resúmenes de contribuciones presentadas en simposios y congresos internacionales.

1.5. Principales contribuciones

Las principales contribuciones científicas generadas mediante el desarrollo de esta tesis son las siguientes:

Contribución 1: Palacios Gutiérrez, A., Valencia Delfa, J.L. & Villeta López, M. Time series clustering using trend, seasonal and autoregressive components to identify maximum temperature patterns in the Iberian Peninsula. Environ Ecol Stat 30, 421–442 (2023). https://doi.org/10.1007/s10651-023-00572-9

Contribución 2: Palacios-Gutiérrez, A., Valencia-Delfa, J.L. & Villeta, M. Identification of extreme temperature and precipitation patterns in Spain based on multiscale

analysis of time series. Nat Hazards 121, 7991–8009 (2025). https://doi.org/10.1007/s11069-024-07082-2

Contribución 3: Palacios-Gutiérrez, A., Valencia-Delfa, J.L. & Villeta, M. Quantifying Impact of Geographical Variables on Climate Change Patterns over Spain by Time Series Clustering. Earth Syst Environ (2025). https://doi.org/10.1007/s41748-025-00568-4

Contribución 4: Palacios Gutiérrez A., Valencia Delfa J.L. Time Series Clustering using Trend, Seasonal and Autoregressive Components: Patterns of Change of Maximum Temperature in Iberian Peninsula. Publications of the Institute of Geophysics, Polish Academy of Sciences. vol. 443 (E-13), 2022, pp. 47–54. DOI: 10.25171/InstGeoph_PAS_Publs-2022-042

Capítulo 2

Capítulo 2

Agrupación de series temporales mediante reducción de la dimensionalidad y conceptos relacionados

Resumen: *En este capítulo, se introducen algunos conceptos fundamentales, relacionados con las series temporales y con los análisis clúster, utilizados a lo largo de esta tesis. A continuación, se describe en detalle el enfoque principal del estudio: la agrupación de series temporales con reducción de la dimensionalidad. Finalmente, se presentan los diferentes algoritmos de clustering utilizados y los índices de validación aplicados para optimizar los resultados de las agrupaciones finales de las series temporales climáticas.*

2.1. Series temporales

Como se indicó en el capítulo anterior, una de las herramientas de análisis multivariantes que nos permite descubrir patrones interesantes sobre el clima y sus cambios, es la agrupación de datos de series temporales. Un tipo de datos bien interesantes y con muchas aplicaciones. Comenzamos presentando brevemente una descripción de las series temporales y sus componentes para aclarar su estructura, antes de discutir el enfoque de agrupación empleado para este tipo de datos.

Se entiende por Serie Temporal una secuencia de observaciones o valores que una variable toma a lo largo del tiempo de manera secuencial. En Box et al. (2015) se indica que una característica intrínseca de las series temporales es que, típicamente, las observaciones adyacentes son dependientes. La naturaleza de esta dependencia entre las observaciones de una serie temporal es de considerable interés práctico, por lo que el análisis de las series temporales se enmarca en técnicas para el análisis de dicha dependencia. Principalmente, el tratamiento y análisis de las series temporales se puede considerar a partir de tres grandes grupos de técnicas (Granger, 1989; Mauricio, 2007; González Velasco & del Puerto García, 2009; Box et al., 2015; García Díaz, 2016):

a. Métodos de descomposición que, en los estudios clásicos, dan lugar a analizar el patrón de comportamiento de una serie mediante las componentes que la gobiernan. Cabe señalar que, al hablar de los componentes de una serie temporal, nos referimos a los elementos fundamentales que estructuran la evolución de los datos a lo largo del tiempo, los cuales se describen en detalle dos párrafos más abajo.

b. Métodos de suavizado, que pueden ser clasificados en dos grupos, métodos de medias móviles y métodos de alisamiento exponencial para reducir la variabilidad y resaltar tendencias.

c. Modelos ARIMA, que son una generalización de los métodos anteriores, en donde se considera que el verdadero modelo matemático para una serie temporal es el concepto de proceso estocástico, el cual se entiende como una sucesión de variables aleatorias que evolucionan con el tiempo.

En general, una serie temporal es gobernada por cuatro componentes (Granger, 1989; Box et al., 2015; García Díaz, 2016):

a. Tendencia, que indica la dirección general del cambio en los datos a lo largo del tiempo, ya sea en aumento, disminución o estabilidad.

b. Variación Estacional, que refleja patrones repetitivos o ciclos que ocurren en intervalos regulares, como días, semanas, meses o años.

c. Factor Cíclico, que refiere variaciones que ocurren en intervalos irregulares y a largo plazo, a menudo influenciados por factores económicos o ambientales.

d. Variación Irregular, o ruido, proveniente de fluctuaciones aleatorias que no siguen un patrón específico y pueden ser causadas por factores impredecibles.

Cuando la secuencia de observaciones que define la serie temporal es ordenada en el tiempo en torno a una variable o característica, se hace referencia a una serie univariante o escalar, pero cuando la secuencia se ordena en el tiempo sobre varias características de una unidad observable en diferentes momentos, entonces hablamos de una serie multivariante o vectorial.

Habitualmente, una serie temporal univariante, matemáticamente, se representa como: $x_1, x_2, x_3, \dots, x_N$; $(x_t: t = 1, \dots, N)$; $(x_t)_1^N$, siendo x_t la observación de orden t $(1 \leq t \leq N)$ de la serie y N el número de observaciones de la serie completa, que indica el tamaño o longitud de la serie. Las observaciones $x_1, x_2, x_3, \dots, x_N$ de la serie temporal pueden recogerse en un vector columna $\boldsymbol{x} \equiv [x_1, x_2, x_3, \dots, x_N]'$ de orden $N \times 1$.

Y por su parte, la denotación matemática de una serie temporal multivariante es: $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \dots, \boldsymbol{x}_N$; $(\boldsymbol{x}_t: t = 1, \dots, N)$; $(\boldsymbol{x}_t)_1^N$, donde $\boldsymbol{x}_t \equiv [x_{t1}, x_{t2}, x_{t3}, \dots, x_{tM}]'$ $(M \geq 2)$ es la observación de orden t $(1 \leq t \leq N)$ de la serie y N el número de observaciones de la serie completa. Las observaciones $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \dots, \boldsymbol{x}_N$ de la serie pueden recogerse en una matriz $\boldsymbol{X}$ de orden $N \times M$:

$$\boldsymbol{X} \equiv \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \boldsymbol{x}_3' \\ \vdots \\ \boldsymbol{x}_N' \end{bmatrix} \equiv \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1M} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \cdots & x_{NM} \end{bmatrix},$$

donde x_{tj} es la observación de orden t $(1 \leq t \leq N)$ de la variable o característica j $(1 \leq i \leq N)$, que es la misma en cada tiempo t.

Es importante mencionar que, en ocasiones, los datos de series temporales pueden estar disponibles sobre varias variables de interés relacionadas, y que en tales casos un análisis más informativo y eficaz es a menudo posible considerando las series individuales como componentes de una serie temporal multivariada o vectorial analizando las series conjuntamente (Box et al., 2015). Este supuesto se verá en los procedimientos propuestos que consideran análisis de series multivariadas y que se incluyen en capítulos posteriores para estudiar las relaciones dinámicas entre varias series temporales de la matriz $\boldsymbol{X}$.

2.2. Generalidades del Análisis Clúster

El Análisis Clúster hace referencia a una variedad de métodos matemáticos que pueden ser empleados para determinar qué objetos de un conjunto son similares. Estos métodos agrupan automáticamente los objetos con descripciones similares en el mismo clúster (Diday & Simon, 1976; Kaufman & Rousseeuw, 1990; Romesburg, 2004).

Dado un conjunto P de observaciones, el Análisis Clúster consiste en realizar una partición del conjunto P en varios subconjuntos o clústeres $P_1, P_2, \dots, P_k$, donde cada observación es asignada a algún clúster por alguna técnica de clasificación.

El Análisis Clúster y sus técnicas funcionan de acuerdo con la forma en que se presentan los datos u observaciones. Así pues, el objetivo puede ser la obtención de modelos para la minería de datos tanto en un contexto supervisado como no supervisado. Agrupar significa dividir los datos en grupos significativos, y esta tarea es la que referencia el aprendizaje no supervisado, mientras que la clasificación, asociada al aprendizaje supervisado, tiene como objetivo asignar objetos particulares a estos grupos (Caruso et al., 2018). La diferencia radica en que, en el aprendizaje supervisado se definen las clases y la información se utiliza para clasificar las observaciones futuras, es decir; se cuenta con individuos u observaciones previamente etiquetados, mientras que en el aprendizaje no supervisado no existe un sistema de clasificación preasignado. Este último sugiere que se busca una estructura detrás de los datos que se quiere revelar. De ahí, es posible comprender que el Análisis Clúster también se conoce como aprendizaje no supervisado, como lo refieren Wierzchoń & Kłopotek (2018), ya que, aquí se carece de la información sobre la pertenencia de los objetos a las clases, y no se conoce, cuántas clases debería haber realmente, aunque la aplicación mecánica de algunos algoritmos permite la división de cualquier conjunto arbitrario en un número dado de clases, pero la partición obtenida puede tener o no sentido.

Los algoritmos de agrupamiento se clasifican en dos categorías principales (Kaufman & Rousseeuw, 1990; Caruso et al., 2018; Ezugwu et al., 2022; Ikotun et al., 2023):

a. Métodos jerárquicos: Organizan las observaciones en una estructura de árbol. Las hojas del árbol corresponden a las observaciones y los nodos agrupan subconjuntos de observaciones relacionadas. Existe una jerarquía en los subconjuntos asociados a las ramas del árbol, y hay dos amplias familias de métodos jerárquicos:

> a.1 Métodos aglomerativos que, parten de una situación inicial de n grupos (uno por cada observación) y fusionan los más parecidos s hasta que todas las observaciones pertenecen al mismo grupo.
>
> a.2 Métodos de división que, en cambio, comienzan con un solo grupo y luego los grupos se dividen de forma recursiva hasta que se obtienen n grupos individuales.

b. Métodos no jerárquicos o Particionales: se caracterizan por la identificación de puntos de agregación, denominados centroides, alrededor de los cuales se crean los grupos. Cada observación del conjunto de datos se asigna al grupo con el centroide más cercano. Los algoritmos de partición producen los agrupamientos en un enfoque heurístico mientras se optimiza una función de criterio definida globalmente en todos los objetos del conjunto o localmente dentro de un subconjunto de objetos (Ahmad & Khan, 2019). Este tipo de algoritmos de agrupamiento requieren la especificación de diferentes valores n suministrados en diferentes ejecuciones para obtener la mejor configuración para producir los agrupamientos óptimos, puesto que optimizar una función de criterio en un conjunto de objetos utilizando una búsqueda combinatoria de todos los valores posibles para obtener el valor óptimo es computacionalmente prohibitivo.

Como se puede apreciar, la similitud entre observaciones desempeña un papel fundamental en el análisis de conglomerados. Esta constituye la base para la definición de un grupo, y hay muchas opciones posibles para cuantificarla, dependiendo tanto de la naturaleza de las variables analizadas como de los objetivos del estudio (Caruso et al., 2018; Ezugwu et al., 2022). Por tanto, como indican Wierzchoń & Kłopotek (2018), dado que el propósito del análisis de conglomerados es dividir un conjunto de objetos en grupos homogéneos, su aplicación práctica requiere responder a dos preguntas básicas: cómo definir la similitud entre los objetos, y de qué manera debe hacer uno uso de la similitud así definida en el proceso de agrupación.

De forma general, para poder cuantificar asociaciones entre pares de objetos, hay una noción de medida de similitud $s: X \times X \to R$ o de disimilitud (Wierzchoń & Kłopotek, 2018). Las dos medidas son, en principio, duales, es decir, cuanto menor es el valor de disimilitud, más similares son los dos objetos comparados. Un ejemplo particular de disimilitud es la distancia, entendida como métrica usual, es decir, una función $d: X \times X \to R_+ \cup \{0\}$ que cumple tres condiciones: (a) $d(x, y) = 0$ si y sólo si $x = y$ "igualdad", (b) $d(x, y) = d(y, x)$ "simetría", (c) $d(x, y) \leq d(x, z) + d(z, y)$ "desigualdad triangular", para $x, y, z \in X$ arbitrarios. Cuando solo se satisfacen las condiciones (b) y (c), entonces d es llamado pseudodistancia.

Así pues, la similitud y la disimilitud son medidas de proximidad. En este marco, una matriz de similitud almacenará una colección de proximidades disponibles para todos los pares de objetos. Si el conjunto X está compuesto por n observaciones, la matriz de similitud o disimilitud será simétrica y de orden $n \times n$. La disimilitud entre x y y representada por $d(x, y)$ será un número no negativo, cuando x e y sean muy similares estará cerca de 0, y a medida que difieran crecerá. Y como se vio en la propiedad (a) de la métrica usual, la diferencia entre una observación con ella misma es 0, por lo que la diagonal principal de dicha matriz estará conformada por elementos nulos. Es importante destacar que, por lo general, la cuantificación de la similitud/disimilitud se hace en función de las variables o atributos que describen a las observaciones.

Los enfoques para calcular las diferencias entre las observaciones son muchos, y además varían de acuerdo con el tipo de atributos que describen a las observaciones. En algunos trabajos (Caruso et al., 2018; Wierzchoń & Kłopotek, 2018; Abu Alfeilat et al., 2019; Ezugwu et al., 2022) se definen las medidas más utilizadas para calcular la disimilitud de objetos descritos por atributos nominales, ordinales, numéricos y de tipos mixtos. La distancia euclidiana ha sido ampliamente generalizada y se han desarrollado variantes como la Distancia de Minkowski, Distancia de Mahalanobis, Divergencia de Bregman, Distancia del coseno y Distancia de potencia, los cuales se detallan en Wierzchoń & Kłopotek (2018). Algunos estudios (Cha, 2007; Abu Alfeilat et al., 2019) ofrecen interesantes y exhaustivas revisiones sobre varias medidas de similitud/disimilitud.

Teniendo claro el objetivo del análisis clúster y comprendiendo la importancia de los aspectos referidos anteriormente para su aplicación, se hace interesante entender el proceso para realizarlo. Un esquema general para la agrupación basada en

disimilitudes se puede apreciar a partir del esquema definido por Hansen & Jaumard (1997) al indicar que los algoritmos de análisis de conglomerados se basan en estadísticas, matemáticas e informática, y que la mayoría de los métodos de análisis de conglomerados se basan en diferencias (o similitudes o proximidades). Tal enfoque es el siguiente, que a menudo es aplicado en la actualidad.

(a) Muestra. Seleccionar una muestra $X = \{x_1, x_2, x_3, \dots, x_n\}$ de n observaciones entre las cuales se encontrarán las agrupaciones.

(b) Datos. Observar o medir p características o atributos en cada una de las observaciones de X. Esto produce una matriz de datos M de orden $n \times p$.

(c) Disimilitudes. Calcular a partir de la matriz M una matriz D de disimilitudes $d(x_i, x_j)$ entre observaciones. Tales diferencias, normalmente, satisfacen las propiedades $d(x_i, x_j) \geq 0$, $d(x_i, x_i) = 0, d(x_i, x_j) = d(x_j, x_i)$ para $i, j = 1,2,3, \dots, n$. No necesitan satisfacer la desigualdad triangular, es decir; ser distancias.

(d) Restricciones. Elegir el tipo de agrupamiento deseado (jerárquico, particional, basado en posiciones, etc.), así como cualquier restricción adicional sobre los clústeres, como límites en el peso, en la cardinalidad o requisitos de conectividad.

(e) Criterio. Seleccionar un criterio (o posiblemente dos criterios) para expresar homogeneidad y/o separación de los clústeres en la agrupación que se va a encontrar.

(f) Algoritmo. Escoger o diseñar un algoritmo para el problema definido en (d), (e). A su vez identificar o implementar el software necesario para su ejecución.

(g) Computación. Aplicar el algoritmo elegido a la matriz de disimilitudes D, con el fin de obtener clústeres o agrupaciones de acuerdo con el enfoque previamente establecido.

(h) Interpretación. Aplicar pruebas formales o informales para seleccionar las mejores agrupaciones entre las obtenidas en (g). Describir los clústeres por sus listas de observaciones y estadísticas descriptivas. Y finalmente, proceder a una interpretación sustantiva de los resultados.

2.3. Análisis Clúster de Series Temporales mediante Reducción de la Dimensionalidad

Como se ha señalado anteriormente, la agrupación en clústeres, como técnica de la minería de datos, permite agrupar observaciones con alta similitud en conjuntos homogéneos, sin necesidad de disponer de información previa sobre la estructura de los grupos. Esto facilita que la similitud entre objetos u observaciones de grupos diferentes sea mínima. Este enfoque resulta especialmente útil para el análisis exploratorio, pues al permitir agrupaciones de datos sin etiquetar, posibilita la identificación de estructuras subyacentes en el conjunto de datos.

El avance en la capacidad de procesamiento y almacenamiento ha incrementado la posibilidad de conservar datos a lo largo del tiempo, lo cual permite analizar la evolución de sus características. En este contexto, es posible distinguir entre datos estáticos y datos dinámicos. Se consideran datos estáticos aquellos cuyas características permanecen constantes o experimentan cambios insignificantes en el tiempo, mientras que los datos dinámicos presentan variaciones significativas en sus atributos a lo largo del tiempo.

La mayor parte de los trabajos de agrupamiento son realizados sobre datos estáticos, para estos hay cinco clases de métodos de agrupamiento (Han et al., 2011): métodos de partición, métodos jerárquicos, métodos basados en densidad, métodos basados en cuadrículas y métodos basados en modelos. Según Rani & Sikka (2012) específicamente, los métodos de partición, jerárquicos y los basados en modelos, han sido utilizados directamente o con modificaciones para agrupar series temporales que, como se vio en su definición, se clasifican como datos dinámicos. Un método de partición constituye k particiones a partir de los datos sin etiquetar de un conjunto de n tuplas, de modo que cada partición representa un clúster que contiene al menos una observación, $k \leq n$. Se dice que la partición es nítida si toda observación pertenece exclusivamente a un clúster, y es borrosa si se permite que un objeto esté en más de un clúster con un grado de pertenencia diferente. Los algoritmos *k-medias* y *k-medoides* son métodos heurísticos conocidos de particiones nítidas y los algoritmos *difusos de c-medias* y *difuso de c-medoides* son métodos de particiones difusas. El método jerárquico fue definido anteriormente, y por su lado, los métodos basados en modelos asumen un modelo para cada uno de los grupos e intentan ajustar mejor los datos al modelo asumido. Dentro de esta categoría, se distinguen principalmente los enfoques estadísticos y los basados en redes neuronales.

Aunque una serie temporal está conformada por un número elevado de puntos de datos, esta puede verse como un único objeto (Kumar & Nagabhushan, 2006). Así pues, para adelantar la tarea de agrupación de un conjunto de series temporales, es preciso tratar cada serie temporal como un solo objeto, independientemente del enfoque que sea empleado. Según Aghabozorgi et al. (2015) la importancia y la necesidad de agrupar objetos tan complejos como son las series temporales (ya que son objetos de objetos realizados cronológicamente) se justifica por los siguientes objetivos superpuestos:

1) La agrupación en clúster de series temporales es una solución común que se realiza para descubrir patrones en los conjuntos de datos de series temporales y extraer la información valiosa que estos contienen,

2) Facilita la estructuración de conjuntos de datos de gran tamaño mediante la agrupación de series temporales similares, ya sea mediante la agregación de datos en grupos no superpuestos o mediante una taxonomía como una jerarquía de conceptos abstractos,

3) El agrupamiento de series temporales es el enfoque más empleado como técnica exploratoria y como subrutina en algoritmos de minería de datos más complejos,

como el descubrimiento de reglas, indexación, clasificación y detección de anomalías, y

4) La visualización de clústeres de series temporales puede ayudar a comprender rápidamente la estructura de datos, clústeres, anomalías y otras regularidades en conjuntos de datos.

De acuerdo con la definición formal sobre el problema agrupamiento de series temporales (Aghabozorgi et al., 2015; Ergüner Özkoç, 2021; Holder et al., 2024), se define la agrupación de series temporales de la siguiente manera.

Agrupación de series temporales: Dodo un conjunto $T = \{T_1, T_2, \dots, T_n\}$ en el que los objetos T_i son series temporales, una partición no supervisada de T en $P = \{P_1, P_2, \dots, P_k\}$, siendo $k \leq n$, de tal manera que las series T_i similares entre sí se agrupen en función de una determinada medida de similitud. Cada grupo P_i representa un clúster de series homogeneas, cumpliendose que $T = \bigcup_{i=1}^{k} P_i$ y $P_i \cap P_j = \emptyset,\ \forall i \neq j$.

La mayoría de las técnicas que se han recomendado para el agrupamiento de series temporales han adoptado alguno de los siguientes enfoques (Warren Liao, 2005; Aghabozorgi et al., 2015; Ergüner Özkoç, 2021):

1) Adaptación de los algoritmos de agrupación en clústeres convencionales existentes, diseñados para datos estáticos, para que sean compatibles con el tipo de datos de series temporales. Normalmente se modifica la medida de distancia.

2) Convertir los datos de series temporales en objetos simples o estáticos para que funcionen como entrada de los algoritmos de agrupamiento convencionales.

3) Uso de múltiples resoluciones de series temporales como entrada de un enfoque de múltiples etapas.

De acuerdo con Aghabozorgi et al. (2015), los principales desafíos asociados al agrupamiento de series temporales son los siguientes:

1.- El volumen de datos puede ser excesivamente grande, superando con frecuencia la capacidad de memoria y requiriendo almacenamiento en disco, lo que ralentiza el proceso.

2.- La alta dimensionalidad dificulta el manejo eficiente de los datos y complica la ejecución de muchos algoritmos.

3.- Las series temporales son naturalmente ruidosas e incluyen valores atípicos y desplazamientos, lo que hace que la selección de una medida de similitud adecuada sea una tarea crítica y compleja.

El proceso de encontrar dicha similitud también se conoce como Coincidencia de Secuencia Completa, de ahí, la importancia de los *métodos de representación* y las *medidas de similitud*, dos aspectos claves para lograr la eficacia y la eficiencia en la gestión de grupos de series temporales, ya que la alta dimensión de los datos de series temporales eleva los costos de procesamiento y almacenamiento, por lo tanto,

no es muy conveniente tratar directamente con dichos datos en su formato sin procesar. Así pues, se hace interesante desarrollar técnicas de representación que puedan reducir la dimensionalidad de las series temporales, sin dejar de preservar las características fundamentales de un conjunto de datos en particular.

Métodos de reducción de la dimensionalidad de series temporales: Estos métodos, también conocidos como métodos de representación de series temporales, representan las series temporales sin procesar en otro espacio de menor dimensión, ya sea mediante la transformación de la serie temporal inicial en una de menor dimensión o mediante la extracción de características. El objetivo principal es simplificar la estructura de los datos, reduciendo su tamaño sin perder información esencial.

Dada una serie temporal T con longitud de dimensión n, se dice que un modelo T' con longitud de dimensión reducida $k < n$ es una representación de T, además T' debe preservar las características fundamentales de T. Si T_i y T_j son similares en el espacio original, entonces sus representaciones T_i' y T_j' también deberán ser similares en el espacio de transformación.

Considerando los aspectos de tamaño y de multidimensionalidad que caracterizan las series temporales y que dificultan el proceso de agrupamiento de estas, es relevante que los datos de series temporales se representen sin ralentizar el tiempo de ejecución del algoritmo y sin una pérdida de datos significativa. En Ergüner Özkoç (2021) se enumeran algunos requisitos para cualquier método de representación de datos de series temporales, los cuales son: Reducir significativamente el tamaño (dimensionalidad de los datos), Mantener las características de forma local y global de la serie de tiempo, Costo computacional aceptable, Nivel razonable de reconstrucción a partir de la representación reducida, Insensibilidad al ruido o manejo de ruido implícito.

Según la taxonomía propuesta por Aghabozorgi et al. (2015) y Ergüner Özkoç (2021), los métodos de representación de series temporales pueden clasificarse en cuatro categorías principales:

1) *Representaciones adaptables a los datos:* se pueden emplear en todos los conjuntos de datos de series temporales y buscan minimizar el error de reconstrucción global utilizando segmentos de longitud variable. Estas técnicas se han aplicado en diferentes enfoques como: Interpolación de polinomios por partes (PPI), Regresión de polinomios por partes (PPR), Aproximación lineal por partes (PLA), Aproximación constante por partes (PCA), Aproximación constante por partes adaptable (APCA), Descomposición de valores singulares (SVD), Lenguaje natural, Lenguaje natural simbólico (NLG), Aproximación agregada simbólica (SAX) e iSAX.

2) *Representaciones adaptables sin datos:* estos enfoques son representaciones adecuadas para series temporales con segmentación de tamaño fijo o de igual longitud, y son aplicaciones de métodos como: Wavelets; HAAR, DAUBECHIES, Coeiflets, Symlets, Transformada Wavelet Discreta (DWT) y

Polinomios espectrales de Chebyshev, DFT espectral, Mapeos Aleatorios, Aproximación Agregada por Piezas (PAA) y Aproximación lineal por partes indexable (IPLA).

3) *Representaciones Basadas en modelos:* al igual que en los dos enfoques anteriores, en este se puede definir la relación de compresión en función de la aplicación en cuestión. Estos enfoques representan una serie temporal de forma estocástica, como los modelos de Markov y el modelo de Markov oculto (HMM), los modelos estadísticos, los mapas de bits de series de tiempo y los modelos autorregresivos de medias móviles (ARMA).

4) *Representaciones dictadas por los datos:* en este enfoque la relación de compresión se define automáticamente en función de series temporales sin procesar, Un ejemplo es el método Clipped, que transforma las series según valores umbrales extraídos directamente de los datos originales.

Medidas de similitud/disimilitud en la agrupación de series temporales: La función que se utiliza para medir la similitud entre dos elementos que se comparan es un componente clave de la agrupación en clústeres (Warren Liao, 2005; Ergüner Özkoç, 2021), y debe considerarse teniendo en cuenta la forma de los datos, puesto que los objetos que se comparan pueden tener varias formas, incluidos valores sin procesar de longitud igual o desigual, vectores de pares de características y valores, matrices de transición, etc. Sin embargo, aunque la medida de similitud/disimilitud sea un componente tan esencial para agrupar series temporales, la similitud o la distancia en la agrupación de series temporales no se basa en la coincidencia exacta como en los métodos de agrupación convencionales, sino que se calcula aproximadamente, es decir; la similitud entre series temporales no se obtiene a través de un cálculo exacto, sino que se estima mediante métodos apropiados (Aghabozorgi et al., 2015; Ergüner Özkoç, 2021). Así, si la distancia estimada es grande, entonces la similitud entre las series temporales es menor y viceversa.

Dado que una serie temporal univariante es la forma más simple de este tipo de datos, una forma sencilla de calcular la distancia entre un par de series temporales es considerando cada una de las series como univariada y luego calcular la medición de la distancia en todos los puntos de tiempo. La similitud entre dos series de tiempo de igual tamaño es la longitud de la ruta que conecta un par de puntos (Ghysels et al., 2006).

De acuerdo con Ergüner Özkoç (2021), una definición formal de la distancia entre series temporales es la siguiente. Sean T_i y T_j dos ST, la distancia $dis(T_i, T_j)$ es una función que mide la similitud entre las dos series, tomando como entrada el par (T_i, T_j) y teniendo en cuenta los puntos de tiempos o las características de las series temporales.

Generalmente, los estudios y procesos para la definición de una medida de distancia que permita comparar series de tiempo deben atender a los problemas originados por las propiedades o características más comunes de este tipo de datos temporales, los cuales son: el ruido, la escala de amplitud, la escala longitudinal, la desviación

lineal, la desviación temporal, la traslación de desplazamiento y las discontinuidades. Según Ergüner Özkoç (2021) y Aghabozorgi et al. (2015), de acuerdo con el objetivo de agrupación de series de tiempo, normalmente hay tres enfoques para medir la similitud entre los datos de series de tiempo, los cuales se detallan a continuación.

- *Series temporales similares en el tiempo:* Este enfoque evalúa la similitud punto a punto, considerando las diferencias entre valores en instantes temporales coincidentes. Es apropiado para series temporales sincronizadas, y se utilizan medidas como la distancia euclidiana y la distancia basada en correlación. No obstante, este tipo de medición de la similitud es costoso para ST sin procesar, por lo que es necesario aplicar algún proceso de transformación previa, como la transformada de Fourier, transformada wavelet, o aproximación agregada por partes, con el fin de reducir la dimensionalidad y facilitar la comparación (Bagnall & Janacek, 2005; Ratanamahatana et al., 2005).

- *Series temporales similares en forma:* La medición de la similitud basada en la forma de series temporales permite obtener grupos de series temporales con patrones de cambio similares independientemente de los puntos temporales, sin importar cuantas veces ocurra el patrón. Para medir la disimilitud, en este enfoque, se han empleado métodos elásticos (Agrawal et al., 1993; Aref et al., 2004) así como Dynamic time Warping (Chu et al., 2002).

- *Series temporales similares en el cambio:* Este enfoque, también conocido como de similitud estructural, permite identificar grupos de series temporales que presentan una estructura de autocorrelación similar. Por lo general, en el proceso, se utilizan métodos de modelado como los modelos ocultos de Márkov o procesos ARMA, y luego se mide la similitud en los parámetros del modelo ajustado a las series de tiempo (Smyth, 1997; Kalpakis et al., 2001; Xiong & Yeung, 2002). Esta medida de similitud no es adecuada para series de tiempo cortas ya que ser requiere una base suficiente para ajustar con fiabilidad los modelos.

Como se ha observado, los métodos de representación, las estimaciones de distancias y los mismos algoritmos de agrupamientos se complementan y son determinantes en el proceso de agrupación de las series temporales. La forma en la que estos componentes se definan y se integren da lugar a distintos enfoques de agrupamiento. De acuerdo con la duración de la serie temporal, los enfoques para su agrupamiento podrían clasificarse en un nivel de forma o en un nivel de estructura (Aghabozorgi et al., 2015). Por lo general, el primero se utiliza para medir la similitud en agrupaciones de series temporales de corta duración y evalúa similitudes locales, mientras que, el segundo mide la similitud que se basa en la estructura global y de alto nivel, y se utiliza para datos de series de tiempo de larga duración.

Tal y como se indicó en la introducción, principalmente hay tres categorías o enfoques para agrupar series temporales, según el tipo de información que utilicen: datos brutos, indirectamente con características extraídas de los datos brutos o indirectamente con modelos construidos a partir de datos brutos. A diferencia del enfoque basado en datos brutos, que se clasifica en un nivel de forma, los enfoques basados en características y basados en modelos se clasifican en un nivel de estructura global, y, por tanto, emplean métodos de representación que permiten reducir la dimensionalidad.

En los enfoques basados en características, las series temporales sin procesar se transforman en vectores de características de menor dimensión. A partir de estos vectores, generalmente de igual longitud para todas las series, se aplican algoritmos de agrupamiento convencionales. En este contexto, la distancia Euclidiana es la medida más comúnmente empleada para comparar los vectores resultantes (Hautamaki et al., 2008).

En los enfoques basados en modelos, cada serie temporal sin procesar se transforma en parámetros de un modelo (un modelo paramétrico para cada serie de tiempo) y luego se elige un algoritmo de agrupamiento, por lo general convencional, con una respectiva distancia de modelo adecuada y se aplica a los parámetros del modelo extraído (Warren Liao, 2005).

Hasta aquí, se han introducido los puntos de interés que se tienen en cuenta para agrupar las series temporales, con especial énfasis en la reducción de la dimensionalidad. A continuación, se presentan los algoritmos de clustering usados en los procesos de agrupamiento, así como los índices de validación considerados para optimizar las agrupaciones obtenidas. Los enfoques de reducción de la dimensionalidad junto con las técnicas de preprocesamiento empleadas se detallan más adelante en los capítulos correspondientes, al igual que las distancias empleadas y las herramientas y estadísticas utilizadas para analizar los patrones del clima y sus cambios.

2.4. Algoritmos de clustering empleados

En este apartado se incluye una descripción de los algoritmos de clustering empleados durante los procesos de agrupamientos. Es importante señalar que no todos ellos fueron utilizados en la definición final de los agrupamientos presentados en los distintos capítulos, ya que la selección se basó en los resultados arrojados por los índices de validación.

K-means

El algoritmo K-means (MacQueen, 1967) es uno de los métodos de aprendizaje automático no supervisados más sencillos y populares. Este algoritmo separa el conjunto de datos inicial en k clústeres diferentes, donde cada observación se asocia al clúster con la media μ_k más cercana.

De los algoritmos K-means, el algoritmo estándar es el de Hartigan & Wong (1979), que define la variación total dentro del clúster como la suma de las distancias euclidianas al cuadrado entre los elementos y el centroide correspondiente. Aunque

este algoritmo tiene una asociación con las medias, la idea básica de K-means consiste en definir clústeres de forma que se minimice la variación total intra-cluster (conocida como variación total dentro de clúster). La función $F(C_k)$ (que representa la suma de errores al cuadrado dentro de los conglomerados) se minimiza:

$$F(C_k) = \sum_{k=1}^{k} \sum_{i=1}^{n_k} \left(\left\| x_i^{(k)} - \mu_k \right\| \right)^2$$

donde x_i designa una observación perteneciente al clúster k, μ_k es el valor medio de los puntos asignados al clúster k y n_k es el total de elementos incluidos en el clúster k.

Con el fin de que la compacidad (es decir, la bondad), medida por la suma cuadrática total dentro del clúster, de la agrupación sea lo más pequeña posible, este algoritmo asigna cada x_i a un clúster determinado, de modo que la distancia suma de cuadrados de la observación a sus centros de clúster asignados μ_k sea mínima.

El algoritmo K-means opera de forma iterativa mediante dos pasos principales hasta que se alcanza la convergencia:

Asignación: Tras asignar de manera aleatoria cada observación a un clúster, primero se calcula la media de cada agrupación.

Actualización: Luego se reasignan las observaciones al clúster más cercano minimizando la distancia a las medias de los clústeres. El algoritmo se detiene cuando no hay más cambios en las asignaciones.

K-medoids

K-medoids es un algoritmo de partición que tiene un enfoque de clustering relacionado con el algoritmo de análisis de agrupamiento iterativo K-means. En una partición de datos obtenida mediante K-medoids cada clúster se representa por un punto seleccionado de la propia muestra (el más próximo realmente al centro) denominado *medoid* que, es el que minimiza la disimilitud media entre él y todos los demás miembros del clúster, y esto es justo lo que lo diferencia de K-means, ya que, como se vio anteriormente, en K-means el centro de un clúster dado se obtiene como el valor promedio de los elementos pertenecientes al clúster, lo que lo hace más sensible al ruido y a los valores atípicos en comparación con K-medoids. Por tanto, K-medoids es una alternativa robusta al clustering K-means. Esto significa que, en comparación con K-means, el algoritmo es menos sensible al ruido y a los valores atípicos, ya que utiliza medoides como centros de clúster en lugar de medias.

Entre los métodos de clustering K-medoids, el más común es el algoritmo de Partición Alrededor de Medoids (PAM) propuesto por Kaufman & Rousseeuw (Kaufman & Rousseeuw, 1990) al introducir por primera vez el algoritmo K-medoids. Este procedimiento selecciona k objetos representativos o medoids entre las observaciones del conjunto de datos, para luego asignar cada objeto al clúster donde se minimice la distancia, es decir; a su medoid más cercano. A continuación, se actualizan los medoids; si alguno de los objetos del conglomerado disminuye el

coeficiente medio de disimilitud, debe ser seleccionado como nuevo medoid. Si hay actualización de medoids, se repite el procedimiento de asignación de objetos al medoid más cercano, en caso contrario se finaliza el algoritmo.

Clustering jerárquico

Los algoritmos jerárquicos son métodos de clustering que, a diferencia de los algoritmos de partición, no requieren especificar previamente el número de clústeres para agrupar objetos basándose en su similitud.

El resultado de la agrupación jerárquica produce un dendrograma que representa la agrupación anidada de patrones y niveles de similitud en los que cambian las agrupaciones (Jain et al., 1999). El dendrograma es una jerarquía multinivel en la que los clústeres de un nivel se unen para formar los clústeres de los niveles siguientes, permitiendo decidir el nivel en el que cortar el árbol para generar grupos adecuados de un conjunto de datos.

El clustering jerárquico aglomerativo (HA, por sus siglas en inglés), también conocido como AGNES (Agglomerative Nesting), es el tipo más común de clustering jerárquico que se utiliza para agrupar objetos basándose en la similitud (Murtagh & Legendre, 2014), y se realiza en dos etapas. En la primera, se define una medida de similitud en un conjunto de observaciones y se considera cada una de las observaciones del conjunto como un grupo único. Se unen las observaciones más cercanas (más similares), eliminando los aglomerados previos a la unión. Se redefinen las diferencias entre clústeres con respecto a las actualizaciones (clústeres recién creados). Las actualizaciones se realizan de acuerdo con la cardinalidad del conjunto de observaciones iniciales, así, si n es la cardinalidad del conjunto total de observaciones, entonces HA se completa en $n-1$ pasos aglomerativos. En este estudio, empleamos el HA con el método de Ward (Murtagh & Legendre, 2014) que minimiza la varianza total intragrupo.

Mapas Auto organizados de Kohonen

Los Mapas Autoorganizados de Kohonen (SOM, por sus siglas en inglés Self-Organising Maps) son un algoritmo de agrupamiento basado en modelos que asume que existen órdenes en los objetos de entrada o estructuras topológicas, que pueden realizar el mapeo de reducción dimensional desde el espacio de entrada ($n-dimensional$) al plano de salida ($2-D$) (Kohonen & Oja, 1996). Por tanto, SOM es una red neuronal que consta solo de una capa de entrada y una capa oculta, por lo que dicho mapeo tiene una fuerte conexión teórica con el funcionamiento real del cerebro y es capaz de mantener las características topológicas de los datos. En el entrenamiento de esta red neuronal, se adopta el método de “aprendizaje competitivo”, y cada nodo en la capa oculta representa una clase que debe agregarse. Cada objeto de entrada encuentra un nodo en la capa oculta que mejor se adapta a él, que se llama su nodo activo o "neurona ganadora". Después se realiza la actualización de los parámetros del nodo activo. Estos parámetros se actualizan mediante el método de descenso de gradiente aleatorio. Los puntos cercanos al nodo activo, según su distancia al mismo, actualizan sus propios valores. En esta tesis se ha empleado una de las variantes de SOM que considera el caso de matrices de

disimilitud (Olteanu & Villa-Vialaneix, 2015), de modo que nuestras series se puedan agrupar formando estructuras u órdenes topológicos similares en una clase.

Hierarchical K-means Clustering

El algoritmo Jerárquico K-means (hkmeans) es un método híbrido que combina de forma recursiva el algoritmo Jerárquico con el algoritmo K-means (Lee et al., 2010). El procedimiento consta de tres pasos:

1.- hkmeans calcula la agrupación jerárquica y corta el árbol o dendrograma en k agrupaciones.

2.- Se calcula el centro o la media de cada grupo, y

3.-Se aplica K-means utilizando como centros iniciales los centroides obtenidos en el paso anterior (Lee et al., 2010; Kassambara, 2017).

Aunque K-means es muy popular, presenta algunas limitaciones, como la especificación del número de clústeres por adelantado y la selección de los centroides iniciales al azar, ya que la solución final del agrupamiento es sensible a esa selección aleatoria inicial de los centroides. Esto hace que hkmeans sea interesante, ya que solventa estas debilidades utilizando los resultados del agrupamiento jerárquico para la inicialización, lo que reduce la sensibilidad a las condiciones iniciales y puede mejorar los resultados finales del agrupamiento K-means.

Fuzzy Clustering

El clustering difuso es considerado un clustering blando, ya que en este algoritmo cada elemento tiene una probabilidad de pertenecer a los diferentes clústeres, lo que significa que cada elemento tiene una serie de coeficientes de pertenencia que corresponden al grado de estar en un clúster determinado. Dichos coeficientes o grados de pertenencia de un elemento a un clúster determinado son valores numéricos que varían entre 0 y 1. Así los elementos cercanos al centro de un clúster pueden estar en tal clúster en mayor grado que los puntos en el borde del clúster.

El primero en proponer representar los clústeres como conjuntos difusos fue Ruspini (1969). Posteriormente Bezdek (1981) desarrolló el algoritmo Fuzzy C-Means (FCM), basado en el algoritmo clásico C-means. En este método, el centroide de cada clúster es calculado como la media ponderada de todos los puntos según su grado de pertenencia al mismo.

Con la idea de que un elemento puede pertenecer parcialmente a diferentes clústeres, la partición suave se define formalmente de la siguiente manera.

Si x_i es un elemento perteneciente al conjunto de datos X, se dice que una partición $P = \{P_1, P_2, \dots, P_c\}$ es una partición suave de X si se cumple que:

$$(1)\ \forall x_i \in X \quad \forall P_j \in P \quad 0 \leq \mu_{P_j}(x_i) \leq 1$$

$$(2)\ \forall x_i \in X \quad \exists P_j \in P \quad tal\ que \quad \mu_{P_j}(x_i) > 0$$

Donde $\mu_{P_j}(x_i)$ es el grado de pertenencia de x_i al clúster P_j.

Si la suma de los grados de pertenencia de un elemento específico en todos los clústeres es igual a 1, entonces se tiene un tipo de partición suave especial.

$$\sum_j \mu_{P_j}(x_i) = 1 \qquad \forall x_i \in X$$

Si la partición suave cumple dicha condición adicional, entonces es llamada *partición suave restringida*, y para ello, el FCM extiende la función objetivo J del clásico C-means incorporando los grados de pertenencia difusos de cada elemento en cada clúster e introduciendo un parámetro adicional m como peso exponencial en la función de pertenencia, por lo que la función objetivo J_m del FCM es:

$$J_m(P,V) = \sum_{i=1}^{k} \sum_{xk \in X} \left(\mu_{P_i}(xk)\right)^m \|xk - v_i\|^2$$

Siendo P una partición difusa de X formada por $P_1, P_2, \dots, P_c$. V es el vector de los centros v_i de cada clúster a ser identificados. El peso m determina el grado en el cual los miembros parciales de un clúster afectan el resultado.

En general, FCM pretende generar una partición adecuada mediante la búsqueda de los prototipos v_i y de las funciones de pertenencia μ_{P_i} que minimicen J_m, dos condiciones que definen el teorema que sirve de fundamento del FCM:

$$(3)\ \mu_{P_i}(x) = \frac{1}{\sum_{j=1}^{k} \left[\frac{\|x - v_i\|^2}{\|x - v_j\|^2}\right]^{\frac{1}{m-1}}} \qquad 1 \le i \le k, x \in X$$

$$(4)\ v_i = \frac{\sum_{x \in X} \left(\mu_{P_i}(x)\right)^m}{\sum_{x \in X}^{n} \left(\mu_{P_i}(x)\right)^m} \qquad 1 \le i \le k$$

FCM efectúa dos pasos tras determinar los valores de k y de m, y el criterio de parada. Primero calcula las funciones de pertenencia mediante (3), y luego actualiza los prototipos empleando (4), estos pasos se repiten iterativamente hasta alcanzar el criterio de parada.

Los algoritmos de clustering anteriores son los que se han tenido en cuenta en esta tesis para agrupar las series temporales en base a los vectores de características definidos en cada capítulo posterior, donde se verá cuáles de estos algoritmos se incluyen en cada caso.

2.5. Índices de validación de clustering empleados

Tras aplicar los algoritmos de clustering descritos anteriormente considerando diferentes valores de k, en los diferentes capítulos se verá la aplicación de algunos de los siguientes índices seleccionados para optimizar las agrupaciones finales.

Estos índices se aplican empleando un paquete de R elaborado por Desgraupes (2017), y que se detallan a continuación.

Índice de Dunn

El índice de Dunn (Dunn, 1974) es una métrica empleada para evaluar la calidad de las agrupaciones permitiendo identificar los conglomerados que son compactos y están bien separados. Por lo cual, tiene en cuenta las distancias inter-cluster, así como las distancias intra-cluster.

De manera formal, el índice de Dunn se define como la relación entre la distancia mínima inter-cluster y la distancia máxima intra-cluster:

Si d_{min} es la distancia mínima entre puntos de conglomerados diferentes, sabiendo que la distancia entre conglomerados diferentes C_i y C_j (inter-cluster) se mide por la distancia entre sus puntos más cercanos, y d_{max} la mayor distancia dentro de un conglomerado, considerando que la distancia dentro de un clúster C_k (inter-cluster), por lo general es medida como la distancia máxima entre dos puntos cualesquiera dentro del mismo clúster, entonces el índice de Dunn se denota como:

$$I_{Dunn} = \frac{d_{min}}{d_{max}}$$

Cuanto más altos sean los valores del índice de Dunn, indican mejores resultados de agrupación (Dunn, 1974; Desgraupes, 2017).

Índice de Silhouette

El coeficiente de Silhouette (Rousseeuw, 1987) es uno de los métodos intrínsecos que se emplean para evaluar la calidad de la agrupación, toda vez que, al evaluar la agrupación, permite examinar lo bien que se encuentra cada elemento o punto de datos dentro de su clúster considerando tanto la cohesión dentro del clúster como la separación entre clústeres. Por tanto, este índice permite cuantificar la similitud de un objeto con su propio clúster en comparación con otros clústeres.

De acuerdo con su definición formal, este índice es esencialmente un cociente cuyo numerador está determinado por la diferencia entre la separación y la cohesión del grupo, y cuyo denominador es el máximo de estos dos aspectos (Rousseeuw, 1987; Desgraupes, 2017), por tanto, para su cálculo se debe calcular la cohesión del grupo $a(i)$ y la separación del grupo $b(i)$.

Si M_i es la distancia media de un punto i a cada clúster, la cohesión del grupo $a(i)$ se define como la distancia media del punto M_i a los demás puntos del clúster al que pertenece. Así, si $M_i \in C_k$, entonces:

$$a(i) = \frac{1}{n_k - 1} \sum_{\substack{i' \in I_k \\ i' \neq i}} d(M_i, M_{i'})$$

De otro lado, si $d(M_i, C_{k'})$ es la distancia media de M_i a los puntos de cada uno de los otros conglomerados $C_{k'}$, entonces, la separación del grupo $b(i)$ es la menor de

estas distancias medias, o más bien, la distancia de M_i a todos los demás puntos de datos en el grupo más cercano:

$$b(i) = \min_{k' \neq k} d(M_i, C_{k'}), \qquad con \quad d(M_i, C_{k'}) = \frac{1}{n_{k'}} \sum_{i' \in I_{k'}} d(M_i, M_{i'})$$

Así, tenemos que,

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

donde $s(i)$ es el coeficiente de Silhouette para el punto i, cuyo valor varía entre -1 y 1. Un valor del coeficiente de Silhouette cercano a $+1$ significan que un punto i está lejos de las agrupaciones vecinas, si es 0 significa que i está en el límite de decisión entre dos agrupaciones vecinas o muy cerca de él, mientras que los valores negativos indican que los puntos i podrían haber sido asignados a grupos equivocados (Rousseeuw, 1987).

Índice C

Considerando las distancias entre los pares de puntos, se define N_W como el número total de pares de puntos distintos que pertenecen a un mismo clúster y N_T como el número total de pares de puntos distintos en el conjunto de datos. El índice C (Hubert & Schultz, 1976) se define a partir de la siguiente fórmula:

$$C = \frac{S_W - S_{min}}{S_{max} - S_{min}}$$

donde S_W: es la suma de las distancias del total de pares de puntos distintos (N_W) dentro de cada clúster.

S_{min}: es la suma de las N_W distancias más pequeñas entre todos los pares de puntos en el conjunto de datos completo.

S_{max}: es la suma de las N_W distancias más grandes entre todos los pares de puntos en el conjunto de datos completo.

Cuanto menor sea el valor de este índice, mejor será la calidad del agrupamiento (Hubert & Schultz, 1976; Desgraupes, 2017).

Índice Gamma de Baker-Hubert

En el contexto de la agrupación, el índice Gamma de Baker-Hubert es una adaptación del índice Γ de correlación entre dos vectores de datos del mismo tamaño (Baker & Hubert, 1975; Desgraupes, 2017), y se define como:

$$\Gamma = \frac{S^+ - S^-}{S^+ + S^-}$$

donde S^+ y S^- definen el número de pares concordantes y el número de pares discordantes, respectivamente, en relación con los dos vectores de datos. Dos vectores son concordantes si los valores se clasifican en el mismo orden en ambos vectores. S^+ es el número de veces que una distancia entre dos puntos que pertenecen al mismo clúster es estrictamente menor que la distancia entre dos

puntos que no pertenecen al mismo clúster, y S^- representa el número de veces que ocurre la situación contraria, es decir, que la distancia entre dos puntos que pertenecen al mismo clúster es estrictamente mayor que la distancia entre dos puntos que no pertenecen al mismo clúster.

El valor de C está entre -1 y 1. Se considera una mejor agrupación cuando el valor sea mayor.

Índice McClain-Rao

Este índice fue introducido por McClain & Rao (1975) para probar la calidad del agrupamiento de un conjunto de objetos. De acuerdo con Desgraupes (2017) para una partición con K clúster, este índice se define a partir de la suma de las distancias dentro de un grupo (S_w) como se mostró anteriormente en la definición del C-index.

$$S_W = \sum_{(i,j)\in N_w} d(M_i, M_j) = \sum_{k=1}^{K} \sum_{\substack{\mathrm{i,j}\in N_w \\ i\neq j}} d(M_i, M_j)$$

N_w denota el número total de distancias entre pares de puntos pertenecientes a un mismo clúster y M_r es el vector que representa al $r-$ é$simo$ objeto.

Y sea S_B la suma de las distancias entre grupos:

$$S_B = \sum_{(i,j)\in N_B} d(M_i, M_j) = \sum_{k\neq k'} \sum_{\substack{\mathrm{i}\in N_k, \mathrm{j}\in N_{k'} \\ i\neq j}} d(M_i, M_j)$$

siendo $N_B = N(N-1)/2 - N_W$ el número total de distancias entre pares de puntos que no pertenecen al mismo grupo. N es el número de observaciones.

El índice de McClain-Rao se define como el cociente entre las distancias medias dentro del grupo y entre los grupos:

$$C = \frac{S_W/N_W}{S_B/N_B} = \frac{N_B}{N_W}\frac{S_W}{S_B}$$

Cuanto menor sea el valor de este índice, mejor será la calidad de la agrupación.

Índice de Ray-Turi

Ray & Turi (1999) presentaron su índice como una medida de validez simple basada en las medidas de distancia intra-cluster e inter-cluster que permite determinar automáticamente el número de clústeres. Este índice se define como un cociente:

El numerador, corresponde a la media de las distancias al cuadrado de todos los puntos con respecto al baricentro del grupo al que pertenecen:

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{\mathrm{i}\in I_k} \left\| M_i^{\{k\}} - G^{\{k\}} \right\|^2 = \frac{1}{N}\sum_{k=1}^{K} WGSS^{\{k\}} = \frac{1}{N} WGSS$$

$WGSS^{\{k\}}$ representa la dispersión dentro de los grupos [(Desgraupes, 2017), sección 1.1.2].

El denominador, corresponde al mínimo de las distancias $\Delta_{kk'}$ al cuadrado entre todos los baricentros de los grupos:

$$\min_{k\neq k'} \Delta_{kk'}^2 = \min_{k\neq k'} d\left(G^{\{k\}}, G^{\{k'\}}\right)^2 = \min_{k\neq k'} \left\| G^{\{k\}} - G^{\{k'\}} \right\|^2$$

Por lo que el índice de Ray-Turi, cuyo valor requiere ser minimizado para una mejor agrupación, se puede escribir así:

$$C = \frac{1}{N} \frac{WGSS}{\min_{k\neq k'} \Delta_{kk'}^2}$$

Índice de Xie Beni

El índice de Xie Beni (Xie & Beni, 1991) es un índice que, aunque generalmente es una medida de validez de clustering difuso, también es aplicable al agrupamiento nítido (Desgraupes, 2017). Mediante este índice es posible evaluar la compacidad y separación de los conglomerados sin hacer suposiciones sobre el número de conglomerados inherente a los datos.

Matemáticamente, el índice de Xie Beni I_{X-B} se define como la siguiente relación entre compacidad y separación.

$$I_{X-B} = \frac{1}{n} \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m \left\| x_j - v_i \right\|^2}{\min_{k\neq i} \|v_k - v_i\|^2}$$

Como se puede ver, I_{X-B} es un cociente en el que el numerador se corresponde con la compacidad de la partición definida por la suma de la distancia al cuadrado dentro del clúster, equivalente a la función objetivo $J_m(U,V)$, y el denominador que corresponde a la separación, está representado por la distancia al cuadrado mínima de los centros de los clústeres.

En la relación, x_j es el objeto j del conjunto de datos, v_i y v_k son los centros correspondientes de los clústeres i y k, u_{ij} es el valor de membresía de x_j al clúster i, y m es el exponente difuso.

En el caso de agrupamiento nítido, el error cuadrático medio equivalente a $J_m(U,V)$ simplemente es reemplazado por la media de las distancias al cuadrado de todos los puntos con respecto al baricentro del clúster al que pertenecen ($\frac{1}{N}WGSS$, indicado anteriormente en el índice de Ray-Turi) (Desgraupes, 2017).

Un conglomerado es más compacto cuanto menor sea el numerador, mientras que cuanto mayor es el denominador, más separado esta un conglomerado. Por tanto, cuanto menor sea I_{X-B}, mejor será la partición (Xie & Beni, 1991; Desgraupes, 2017).

Capítulo 3

Capítulo 3

Agrupación de series temporales univariantes

Resumen: *En este capítulo se presenta un estudio de regionalización climática realizado con el fin de identificar patrones de cambio en la temperatura máxima en la Península Ibérica. El procedimiento de regionalización climática propuesto se basa en la agrupación de series temporales univariantes de temperatura máxima haciendo uso del SSA de manera secuencial como técnica de preprocesamiento inicial durante el proceso de agrupamiento. Las componentes extraídas de las series temporales se modelan mediante modelos de regresión, permitiendo la construcción de los vectores de características para el agrupamiento.*

3.1. Objetivos

Identificar patrones de cambio regionales de la temperatura máxima en la Península Ibérica entre 1931 y 2009.

Proponer un procedimiento para agrupar series temporales univariadas de temperatura máxima, representándolas por vectores de características que capturen sus componentes de tendencia, estacionalidad y de ruido. Este enfoque emplea el SSA de manera secuencial para la descomposición y el modelado, permitiendo reducir la dimensionalidad y el proceso de agrupamiento.

3.2. Metodología

El principal método utilizado en este estudio fue el SSA para la descomposición de las series temporales de temperatura máxima. En primer lugar, SSA se aplicó de manera secuencial para extraer los componentes $\mathbb{T}_{trend}$, $\mathbb{T}_{seasonal}$ y $\mathbb{T}_{residual}$ de cada serie temporal. En la segunda fase, cada uno de los tres componentes extraídos para cada serie fue modelado mediante un modelo, de forma que $\mathbb{T}_{trend} = \mu - \beta t$, $\mathbb{T}_{seasonal} = c_1 \sin(2\pi t/T) + c_2 \cos(2\pi t/T)$ y $AR(T)$ es un autorregresivo para $\mathbb{T}_{residual}$ como describe en la página 47, permitiendo estimar los coeficientes que capturarían su comportamiento y que permitirían construir los vectores de características $C_i = [\mu_i, \beta_i, c_{i1}, c_{i2}, \varphi_{i1}, \varphi_{i2}, \varphi_{i3}, \varphi_{is}]$ que, representarían las series temporales de temperatura máxima en un espacio de dimensión reducido y que servirían de entrada para el agrupamiento.

Una vez representadas las series temporales mediante sus vectores de características C_i, se emplean tres de los algoritmos de agrupamientos referenciados en el capítulo anterior, sección 2.4, haciendo uso de la distancia euclidiana y tras comparar los resultados de los agrupamientos obtenidos con base en tres de los índices de validación definidos en 2.5, se obtuvieron agrupaciones finales en base al algoritmo *hkmeans*.

Posteriormente, se definieron los prototipos representativos de clustering $\mathbb{T}_{gi}$, en base al promedio de las series temporales originales de cada clúster, y finalmente se analizaron las características esenciales de cada clúster obtenido en base a los componentes y estimaciones obtenidas.

3.3. Resumen de resultados

Tras estudiar las similitudes de 1776 series temporales de temperatura máxima distribuidas en la Península Ibérica en una red de 25x25km^2, en base al enfoque propuesto, con el algoritmo hibrido *hkmeans* se identificaron 3 clústeres de series temporales diferenciados.

Cada clúster representa una zona delimitada dentro de la Península Ibérica. Este enfoque permitió identificar patrones de cambio de temperatura asociados al gradiente climático del área de estudio, mostrando diferencias en los niveles de temperaturas regionales y en sus cambios asociados. Las zonas del centro y el norte mostraron aumentos que variaron desde los 0.14°C hasta los 0.20°C, mientras que

las zonas del sur solo mostraron un pequeño descenso en sus niveles de temperaturas.

3.4. Conclusiones

Este enfoque mostró que la combinación de técnicas como SSA con el análisis clustering permitió identificar eficazmente patrones de cambio diferenciados de temperatura máxima en la Península Ibérica.

Aunque este estudio se centra específicamente en el comportamiento de la temperatura, su metodología ejemplifica cómo pueden aplicarse técnicas avanzadas de análisis de series temporales para mejorar la comprensión de los patrones climáticos regionales. Queda demostrado el potencial de utilizar componentes de tendencia, estacionalidad y de ruido para agrupar series temporales, proporcionando valiosos conocimientos sobre la dinámica climática regional.

3.5. Publicación JCR

De acuerdo con la estructura de esta tesis, como cuerpo de este capítulo se incluye la publicación del presente estudio, identificable en las bases de datos científicas de acuerdo con lo siguiente.

Estado: Publicado

Año de Publicación: 2023

Nombre de la Revista: Environmental and Ecological Statistics

Editorial: Springer

Indicadores de Calidad: Factor de Impacto año 2023: 3.0; Rango: Q1 – Primer Decil, 16/168 de la categoría *STATISTICS & PROBABILITY*; CiteScore: 5.9 (Scopus), SNIP 0.867

DOI: https://doi.org/10.1007/s10651-023-00572-9

Environmental and Ecological Statistics (2023) 30:421–442
https://doi.org/10.1007/s10651-023-00572-9

Time series clustering using trend, seasonal and autoregressive components to identify maximum temperature patterns in the Iberian Peninsula

Arnobio Palacios Gutiérrez[1,2] · **Jose Luis Valencia Delfa**[1] · **María Villeta López**[1]

Received: 11 March 2023 / Revised: 21 June 2023 / Accepted: 23 June 2023 /
Published online: 15 July 2023

Abstract
Time series (TS) clustering is a crucial area of data mining that can be used to identify interesting patterns. This study introduces a novel approach to obtain clusters of TS by representing them with feature vectors that define the trend, seasonality and noise components of each series in order to identify areas of the Iberian Peninsula (IP) that follow the same pattern of change in regards to maximum temperature during 1931–2009. This representation allows for dimensionality reduction, and is obtained based on singular spectrum analysis decomposition in a sequential manner, which is a well-developed methodology of TS analysis and forecasting with applications ranging from the decomposition and filtering of nonparametric TS to parameter estimation and forecasting. In this approach, the trend, seasonality and residual components of each TS corresponding to a specific area in the Iberian region are extracted using the proposed SSA methodology. Afterwards, the feature vectors of the TS are obtained by modelling the extracted components and estimating their parameters. Finally, a clustering algorithm is applied to group the TS into clusters, which are defined according to the centroids. This methodology is applied to a climate database with reasonable results that align with the defined characteristics, enabling a spatial exploration of the IP. The results identified three differentiated zones that can be used to describe how the maximum temperature varied: in the northern and central zones, an increase in temperature was noted over time, whereas in the southern zone, a slight decrease was noted. Moreover, different seasonal variations were observed across the zones.

Keywords Clustering · Iberian Peninsula · Maximum temperature time series · Singular spectrum analysis · Time series feature vectors

Handling Editor: Luiz Duczmal.

Extended author information available on the last page of the article

1 Introduction

Climate change is an increasingly noticeable global problem that has a significant impact on society as a whole and on ecosystems. As a result, research on climate change has become increasingly important, especially studies relating to temperature, and has been expanding. As stated in the report of the Sixth Assessment of the Intergovernmental Panel on Climate Change (IPCC), the global mean surface temperature (GMST) increased by 1.1 °C between 2001–2021 and 1850–1900, accelerating its rate after the 1970s (IPCC, 2021). Moreover, in the very near future (2025–2050), the GMST may warm up as much as 0.25 °C per decade, according to the climate model predictions by Samset et al. (2020) and Tebaldi et al. (2021).

Although these numbers may seem low, the changes and effects are really remarkable, as global warming manifests prolonged droughts, heat waves and forest fires. For instance, over the last fifteen years (2003, 2010, 2015, 2018), Europe has experienced extreme heat waves (Kuglitsch et al. 2010; Russo et al. 2015; Molina et al. 2020). According to Calheiros et al. (2021), weather and climate conditions such as high temperatures, moderate annual precipitation and prolonged dry spells are the cause of a large number of forest fires and burnt areas around the globe, particularly in areas of the Iberian Peninsula (IP). According to climate predictions, these conditions are expected to continue for the foreseeable future, with some projections predicting that more intense, prolonged, and frequent extreme heat events will occur in Europe in the twenty first century, with a higher impact on the IP and Mediterranean regions (IPCC 2014; King and Karoly 2017; Dosio et al. 2018; Vicedo-Cabrera et al. 2018).

Analysis of the temperature changes experienced by the IP, as well as the projections that have presented concerning this issue, will be a more manageable process if studies that were conducted to observe how these changes or variations in temperature have occurred in the IP are included. This will certainly be the case if these changes are defined by zones or sub-regions, since the temperature in the IP presents spatial variations strongly influenced by distance from the sea and complex orography, promoting a marked climate gradient from north to south (Lorenzo and Alvarez 2022). On the other hand, considering analyses that take into account temperature extremes will be more beneficial, since most studies in the climate field have focused on mean climate trend analysis, which does not tell us about unusual changes (Gebremichael et al. 2022).

The extreme temperature changes experienced in a geographical area or subregion can be understood by obtaining time series (TS) clusters of these temperatures and defining them in points or areas distributed over a geographical area. Since TS clustering is used to identify interesting patterns in TS data sets, finding TS clusters can be valuable in different domains, such as responding to anomalies or novelties, detecting discordance, integrating applications in dynamic change recognition in TS, and predicting, recommending, and discovering patterns (Aghabozorgi et al. 2015). Different studies on TS clustering (Warren Liao 2005; Rani and Sikka 2012; Aghabozorgi et al. 2015; Ergüner Özkoç 2021) agree that there are three main categories or approaches to TS clustering, depending on

whether one is working directly with raw data, indirectly with features or characteristics extracted from the raw data, or indirectly with models built from raw data. Some studies of different domains have used TS clustering to define patterns or find matching TS (Keogh and Smyth 1997; Huhtala et al. 1999; Wang and Wang 2000; Möller-Levet et al. 2003; Huiting et al. 2006; Guo et al. 2008; Lee et al. 2010, 2018; Li and Wei 2020; Shi et al. 2021).

This study is framed within TS clustering based on the approach of extracting features from data, and proposes a procedure to cluster TS by trend, seasonality, and main autocorrelations to ensure that patterns of change in maximum temperature (TMAX) can be identified for each zone in the IP during the period of 1931–2009. The novelty of this methodology is due to the decomposition of TS using singular spectrum analysis (SSA), a well-developed TS analysis and forecasting methodology whose applications have a wide scope ranging from nonparametric TS decomposition and filtering to parameter estimation and forecasting (Golyandina and Korobeynikov 2014). First, in this decomposition process, three components associated with the trend, seasonality and residual of the initial TS are reconstructed, allowing the parameters that describe these components to be extracted. Second, the representation of each TS is obtained from a feature vector generated on the basis of the calculated parameters, which allows the TS to be clustered using unsupervised learning algorithms, such as k-means (Hartigan and Wong 1979), k-medoids (Park and Jun 2009), hierarchical agglomerative (HA; Lukasová 1979) and Kohonen self-organising maps (SOM; Kohonen and Oja 1996), which are well-known and representative conventional algorithms that use the Euclidean distance. Finally, in the experiment on a climatic database, after comparing the clusters obtained with the different methods, a hybrid approach that combines HA and k-means, called hkmeans, was selected as a clustering algorithm to identify TS that are similar and follow a pattern. A set of 1776 points from a grid of 25×25 km^2 elaborated through spatial interpolation kriging by the "Servicio de Desarrollos Climatológicos" of the Meteorological Spanish State Agency was used. This grid includes points distributed throughout Spain, Portugal and the closest areas of the Atlantic Ocean and the Mediterranean Sea; for each point, a monthly TS of TMAX from January 1931 to 2009 is considered. The way in which the TS are represented here allows the clustering process to be optimised, since one of the advantages of implementing the SSA decomposition is the ability to eliminate the noise of the series (one of its main applications), together with the study of the spectral profile. In addition, costs and speed are improved with the reduction of the dimensionality. Therefore, SSA is used not only to decompose a series into a number of components to determine its spectral profile or to estimate its parameters, but also as a basis for recognizing characteristics and patterns of TS ensembles. The clustering results of this study made it possible to identify three differentiated zones: zone 1 is situated in the north of the IP, in areas with the lowest TMAX and a higher proportion of increase compared to the other areas; zone 2 is located more to the south, in areas with the highest TMAX, as well as a slight decline over an extended period; and zone 3 is stationed more towards the centre, in areas with intermediate TMAX, showing an increase over time. In addition, it was observed that the identified zones show different seasonal variations.

The remainder of this document is organised as follows. Section 2 describes the TS decomposition method, which uses SSA in a sequential manner. Section 3 proposes the new method for defining the trend, seasonality and autoregression patterns of TS, including extracting TS trend, seasonality and residual components and clustering and defining patterns. Section 4 presents the results of the method. Section 5 presents the main conclusions of the paper.

2 Sequential SSA decomposition method

In the following section, the theory of the sequential SSA decomposition method for extracting TS components is briefly presented.

SSA is a technique that is known for its application in TS analysis and prediction and has recently been used to analyse digital images and other objects that are not necessarily flat or rectangular and may contain gaps. SSA is a very unique and attractive methodology for solving a wide range of problems in different areas related to TS and digital images, as it naturally combines parametric and nonparametric techniques (Golyandina et al. 2018).

This technique is based on the singular value decomposition (SVD) of a specific matrix obtained from a TS and aims to decompose original TS into a small number of interpretable components, such as a trend that is smooth and slowly varying, with oscillatory components that are periodic, pure quasiperiodic or amplitude-modulated, and noise without any pattern or structure (Golyandina et al. 2001; Zhigljavsky 2011; Xiao et al. 2014; Golyandina and Korobeynikov 2014).

As stated in Golyandina et al. (2001), SSA consists of four steps: embedding, SVD, grouping and diagonal averaging. In some references, steps 1 and 2 of the generic SSA scheme are combined in the *decomposition stage*, whereas steps 3 and 4 are combined in the *reconstruction stage*. Several additive components of the original TS are obtained through SSA. In the following, the SSA method is presented formally.

Input: $\mathbb{T} = \left(t_1, t_2, \ldots, t_N\right)$, the initial TS, which is a one-dimensional N-order TS.

Result: A decomposition of $\mathbb{T}$ into a sum of identifiable components, $\mathbb{T} = \widetilde{\mathbb{T}}_1 + \widetilde{\mathbb{T}}_2 + \cdots + \widetilde{\mathbb{T}}_m$.

2.1 Step 1: Embedding

The so-called *trajectory matrix* is obtained through the equation $X = \mathcal{T}(\mathbb{T})$, where $\mathcal{T}$ is a linear map that transforms the TS $\mathbb{T}$ into a matrix of order $L \times K$, and where L is an integer that is called the *window length*, $1 < L < N$, and $K = N - L + 1$.

The set of all possible path matrices can be denoted as Hankel matrices, $\mathcal{M}_{L,K}^{(\mathcal{H})}$, where all elements along the diagonal are equal. If N and L are fixed, then there is a biunivocal correspondence between the path matrices and the TS.

The trajectory matrix $\boldsymbol{X}$ constructed from lagged vectors, which are generated from the TS $\mathbb{T}$, can be represented in the following way:

$$\mathcal{T}(\mathbb{T}) = \begin{pmatrix} t_1 & t_2 & t_3 & & t_K \\ t_2 & t_3 & t_4 & \cdots & t_{K+1} \\ t_3 & t_4 & t_5 & & t_{K+2} \\ \vdots & & & \ddots & \vdots \\ t_L & t_{L+1} & t_{L+2} & \cdots & t_N \end{pmatrix}. \tag{1}$$

2.2 Step 2: Decomposition of *X* into the sum of the rank 1 matrices

The result obtained in this step is the following decomposition:

$$\boldsymbol{X} = \sum_i \boldsymbol{X}_i, \boldsymbol{X}_i = \boldsymbol{\sigma}_i \boldsymbol{U}_i \boldsymbol{V}_i^{\boldsymbol{T}}, \tag{2}$$

where $U_i \in R^L$ and $V_i \in R^K$ are vectors, such that $\|U_i\| = 1$ and $\|V_i\| = 1$ for all i and σ_i denotes nonnegative numbers.

If such a decomposition is performed using conventional SVD, the corresponding SSA method is "Basic SSA." In addition, the SVD of the matrix $\boldsymbol{X}$ is calculated via the eigenvalues and eigenvectors of the matrix $S = \boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}$ of size $L \times L$. Here, $\lambda_1, \ldots, \lambda_d$ denotes the eigenvalues of the matrix S listed in decreasing order of magnitude ($\lambda_1 \geq \cdots \geq \lambda_d \geq 0$), whereas $U_i, \ldots, U_d$ denotes the orthonormal system of the eigenvectors of the matrix S corresponding to these eigenvalues, considering that $d = L$. Since $V_i = \boldsymbol{X}^{\boldsymbol{T}} U_i / \sqrt{\lambda_i}, (i = 1, \ldots, d)$ are factor vectors, $\boldsymbol{X}_i = \sqrt{\lambda_i} U_i V_i^T$ are elementary matrices of rank 1. Thus, the SVD of the trajectory matrix can be written as the following:

$$\boldsymbol{X} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_d. \tag{3}$$

The collection $(\sqrt{\lambda_i}, U_i, V_i^T)$ is called an SVD eigenvector of order i and consists of the singular value (equal to $\sqrt{\lambda \sigma_{ii}}$), an eigenvector U_i (the left singular vector) and a factor vector V_i (the right singular vector).

2.3 Step 3: Grouping

The input of this step consists of expansion (2) and specification of how to cluster the components of (2). The index set {1, 2,…,d} must be segmented into m disjoint subsets. $I_1, I_2, \ldots, I_m$, where $I = \{i_1, i_2 \ldots, i_p\} \subset \{1, 2, \ldots, d\}$ as a subset of indices. The resulting matrix $\boldsymbol{X}_I$ corresponding to group I is defined as the following:

$$\boldsymbol{X}_I = \boldsymbol{X}_{i1} + \boldsymbol{X}_{i2} + \cdots + \boldsymbol{X}_{ip}. \tag{4}$$

Thus, if a partition is specified in m disjoint subsets of the index set $\{1, 2, \ldots, d\}$, then, by expansion (2), the result of the grouping step leads to the following decomposition:

$$\boldsymbol{X} = \boldsymbol{X}_{I1} + \boldsymbol{X}_{I2} + \cdots + \boldsymbol{X}_{Im}. \tag{5}$$

The above procedure for choosing the sets $I_1, I_2, \ldots, I_m$ is called the *eigentriple grouping* procedure. The grouping of expansion (2), where $I_j = \{j\}$, is called *elementary*.

2.4 Step 4: Reconstruction

In this step, each matrix $\boldsymbol{X}_{Ik}$ from lumped decomposition (5) is transferred into the form of the input object $\mathbb{T}$, which is a TS of length N. To do this, each matrix $\boldsymbol{X}_{Ik}$ is hankelised and, by means of one-to-one correspondence between Hankel matrices and TS, is transformed into a new series of length N. Thus, applying diagonal averaging to $\boldsymbol{X}_{Ik}$ produces a reconstructed series $\widetilde{\mathbb{T}}_k$ of order N (for more details, see Sect. 1.1.2.6 of Golyandina et al. 2018).

Consequently, the resulting decomposition of the input object $\mathbb{T}$ is the following:

$$\mathbb{T} = \widetilde{\mathbb{T}}_1 + \widetilde{\mathbb{T}}_2 + \cdots + \widetilde{\mathbb{T}}_m. \tag{6}$$

If the grouping is elementary, the reconstructed objects $\widetilde{\mathbb{T}}_k$ in (6) are called *elementary components*.

The SSA parameters, i.e., the length of the window L and the way in which $\boldsymbol{X}_{ik}$ matrices are grouped, are very important for the outcome of the decomposition and depend on the properties of the initial TS and the objective of the analysis. One aspect that helps in choosing these parameters is the notion of separability. The separability of two TS, $\mathbb{T}_N^{(1)}$ and $\mathbb{T}_N^{(2)}$, means the possibility of extracting $\mathbb{T}_N^{(1)}$ from an observed sum $\mathbb{T}_N^{(1)} + \mathbb{T}_N^{(2)}$. According to Golyandina et al. (2001), SSA can approximately separate signals, noise, sinusoidal waves with different frequencies, trends and seasonality, etc.

It is possible to obtain recommendations for the choice of window length based on the (approximate) separability conditions. For example, the value of L should be large enough ($L \approx N/2$), and if extracting an existing periodic component of a TS with one or several known periods is needed, it is advisable to choose a window length that is proportional to the highest period. In Golyandina (2010), the choice of SSA parameters is discussed.

SSA can be performed sequentially, which is recommended when the TS structure is complex (Golyandina et al. 2012). ***Sequential SSA*** consists of two stages: in the first stage, the extraction of the TS trend with a small L is performed, and in the second stage, the periodic components of the residue are detected and extracted with $L \approx N/2$.

3 Trend, seasonality and autoregression SSA-based TS pattern identification

In TS data mining, feature extraction is one of the dimensionality reduction procedures. Features extracted from series concisely represent the relevant features of each TS as a finite set of inputs for a clustering algorithm that can discern the similarities and differences of TS (Wang and Hyndman 2006). In this paper, features from complete series, rather than subsequences, are extracted to look for complete TS that have similar patterns (i.e., similar trend, stationarity and autoregression patterns).

3.1 Pattern identification algorithm

The algorithm used to identify trend, seasonality and autoregression patterns in TS that is proposed in this study can be summarised in the following steps:

1 Perform sequential SSA to extract the three components of the initial series that are associated with trend, seasonality and residual.
2 Model the extracted series in such a way that their associated characteristics can be extracted.
   2.1 Trend component: from $\mathbb{T}_{trend} = \mu - \beta t$, estimate μ and β using a linear regression, where $\mathbb{T}_{trend}$ is the extracted trend series and t is time.
   2.2 Seasonality component: from $\mathbb{T}_{seasonal} = c_1 \sin(2\pi t/w) + c_2 \cos(2\pi t/w)$, estimate c_1 and c_2 using linear regression, where w is the period, $\mathbb{T}_{seasonal}$ is the extracted seasonal component and t is time.
   2.3 Residual component: from $\mathbb{T}_{residual}$, obtain an $AR(T)$ and calculate the autocorrelations $\varphi_1, \varphi_2, \varphi_3$ and φ_w, where w is the period and $\mathbb{T}_{residual}$ is the extracted residual component.

   For each initial series, a feature vector is constructed by considering the estimated parameters.

3 Use a conventional clustering algorithm to obtain a similar TS.
4 Average the initial series of each group to obtain their representative patterns based on the defined characteristics.

The new algorithm is explained step by step below.

3.2 Extracting the trend, seasonality and residual TS

The series for trend and seasonality and those associated with the residual of the original TS are extracted using a sequential SSA-based decomposition approach from the TS trend, seasonality and autoregression pattern identification algorithm proposed in this paper. In sequential SSA decomposition, TS are decomposed into a small number of components (i.e., trend, seasonality and residual) in two stages (Golyandina et al. 2012). When the trend shape is complex, it is impossible to completely decompose the TS at once, resulting in the decomposition process being performed sequentially (Golyandina and Korobeynikov 2014). Thus, sequential SSA is applied to each initial TS to ensure that, in the first stage, the component associated with the trend is extracted by choosing an L value that is as small as possible, (i.e., divisible by the identified period w), to make the TS containing a periodic component smooth. In the second stage of sequential SSA, by considering the maximum L value ($L \approx N/2$ and it is divisible by w) for greater separability, the seasonality of the residual generated in the first stage can be extracted, generating a new TS for the residual part.

As noted in Sect. 2, the four steps of the SSA are highly connected, with each step immediately linked to the previous one, making each of them fundamental to the representation process proposed here, which starts with the extraction of the series of $\mathbb{T}_{trend}$, $\mathbb{T}_{seasonal}$ and $\mathbb{T}_{residual}$ (all of order N) described in this section. Thus, a correct identification of w and a correct definition of L are essential.

This process generates three TS for each initial TS, allowing them to be modelled and the parameters of interest to be calculated.

3.3 Constructing a feature vector for each TS

After modelling the extracted TS according to steps in 3.1.2 of the proposed TS trend, seasonality and autoregression pattern identification algorithm and to the estimated parameters, a feature vector for each TS is constructed.

Following the decomposition of TS $\mathbb{T}_i$ by means of sequential SSA to obtain the reconstruction of the trend, seasonal and residual components, which are modelled based on steps 3.1.2.1. to 3.1.2.3. of the proposed algorithm, the feature vector or representation of $\mathbb{T}_i$, denoted as $C_i = \left[\mu_i, \beta_i, c_{i1}, c_{i2}, \varphi_{i1}, \varphi_{i2}, \varphi_{i3}, \varphi_{is}\right]$, can be constructed.

3.4 Obtaining similar TS

A variety of methods have been developed to assess whether two TS are similar or in the same group. In this paper, the Euclidean distance is employed as a similarity criterion using the unsupervised learning algorithm hkmeans to group feature vectors that translate from the trend, stationarity and residual TS.

Given a dataset of TS with n points $\left\{\mathbb{T}_1, \mathbb{T}_2, \ldots, \mathbb{T}_n\right\}$, where each $\mathbb{T}_i$ is represented by a feature vector $C_i = \left[\mu_i, \beta_i, c_{i1}, c_{i2}, \varphi_{i1}, \varphi_{i2}, \varphi_{i3}, \varphi_{is}\right]$, the unsupervised hkmeans partitioning process $P = \left\{P_1, P_2, \ldots, P_k\right\}$, where the C_i vectors are clustered according to the Euclidean distance as a similarity measure and are called characteristic-based or representation-based TS clustering. P_i is called a cluster, where $D = \bigcup_{i=1}^{k} P_i$ and $P_i \cap P_j = \emptyset, \forall i \neq j$.

In the clustering process, each reconstructed C_i vector has been normalised by centring to a mean of 0 and scaling to a standard deviation of 1 (Bro and Smilde 2003). Afterwards, a similarity matrix is obtained by taking into account these vectors and considering the Euclidean distance as the similarity measure. Finally, the similarity matrix is used as input to the clustering algorithms. Consequently, the clustering process is optimised, noise is eliminated, and computational costs are reduced.

3.5 Representative patterns of each group

After clustering with a cluster algorithm, the original TS of each group is utilised to calculate a new TS $\mathbb{T}_{gi}$ by averaging the TS of each group g_i, with each $\mathbb{T}_{gi}$ serving as

a representative prototype of group g_i. Subsequently, steps in 3.1.2 of the proposed algorithms are applied to define the trend, seasonality and autoregression patterns.

4 Empirical study

4.1 Data preparation

A set of 1776 points from a grid in an IP of 25×25 km^2 elaborated through spatial interpolation kriging conducted by the "Servicio de Desarrollos Climatológicos" of the Meteorological Spanish State Agency was used. According to Fig. 1, this grid includes points distributed in Spain, Portugal, Southern France, North Africa and the closest areas of the Atlantic Ocean and the Mediterranean Sea, and considers a monthly TS of TMAX from January 1931 to December 2009 for each point, with 948 observations each.

4.2 TS analysis

As shown in Fig. 2, the TS of the monthly mean TMAX at the points located in such a grid of the IP are clearly very seasonal, with temperature peaks occurring in the summer months and temperature drops occurring in the winter months. Moreover, a variable trend is observed throughout the period, as shown in Fig. 6.

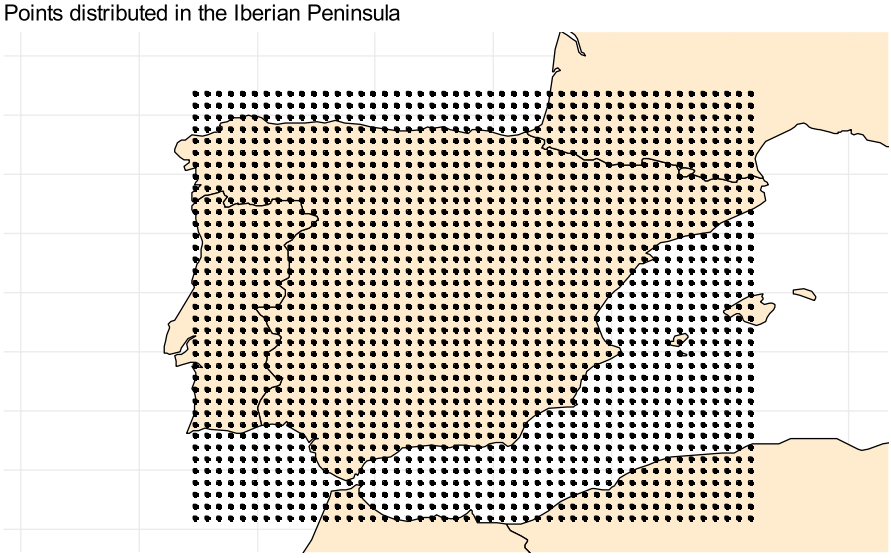


Fig. 1 Spatial distribution of the points on the IP

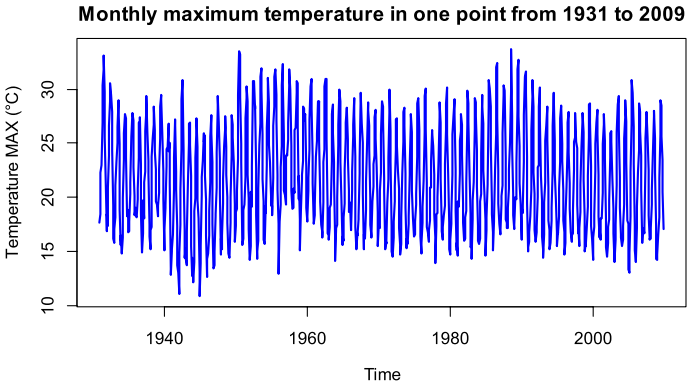


Fig. 2 Average monthly TMAX at a point in IP from January 1931 to December 2009

4.3 Decomposition and reconstruction of SSA

Sequential SSA was applied as a data preprocessing tool due to its capacity for component separability, which facilitates dimensionality reduction, TS representation and result interpretability.

Figure 6 shows that the trend shape of the TS studied is complex, indicating that the decomposition is performed sequentially. First, the trend is extracted. For such a changing trend shape, its extraction is similar to smoothing (Golyandina and Korobeynikov 2014), starting with choosing a window length $L = 12$ to smooth the TS containing a periodic component, as would be done in the moving average procedure. The window length must be the minimum possible length and must be divisible by the period, which in this case is 12, as seen in the periodogram in Fig. 3, suggesting that the seasonality consists of sine waves with the indicated period. Since it is a monthly series, the horizontal axis of this graph varies from 0 to 6, and the peak at point 0 is related to a long-term trend, whereas the peaks at points 1 and 2 are related to the annual and semi-annual cycles, respectively. Generally, the same was noted in all TS in this study.

After performing the first decomposition, it is confirmed that the first eigentriple corresponds to the trend, whereas the other eigentriples contain high-frequency components, meaning that they are not related to the trend. This can be seen in Figs. 4 and 5, which show the shape of the six main eigenvectors and the result of the reconstruction carried out by each of the six eigentriples. Additionally, it can be seen that the principal eigenvector has practically constant coordinates; thus, it corresponds to pure smoothing by the Bartlett filter (Golyandina et al. 2012). Notably, the first reconstructed principal eigenvector in Fig. 5 produces the same exact reconstructed trend shown in Fig. 6.

After extracting the trend, the seasonality is extracted from the residual generated in the first decomposition stage. Figure 7 shows the periodogram of this residual, demonstrating that there is seasonality consisting of sinusoidal waves with periods of 12 and 6, and that it is not possible to observe the peak at the point 0 shown in Fig. 3 since the trend was removed.

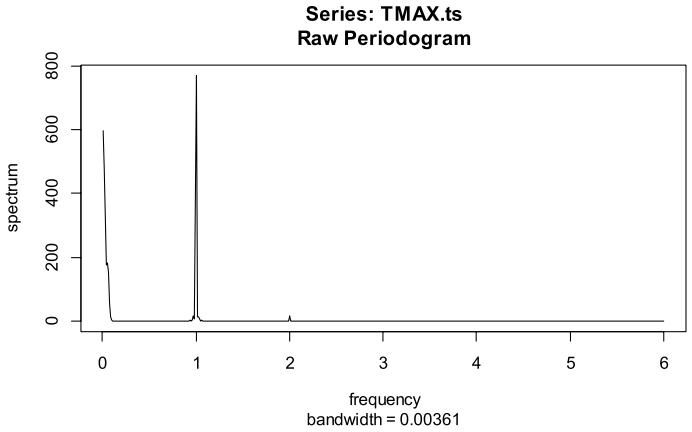
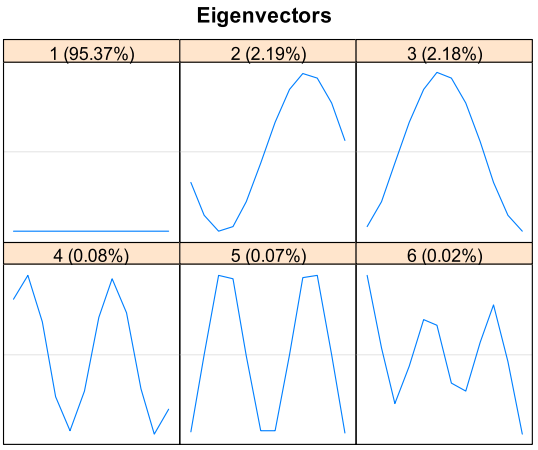


Fig. 3 Periodogram of the initial TS

Fig. 4 Eigenvectors in the 1st stage (L = 12) of the initial TS



Since the length of the TS is $N = 948$, to guarantee better separability, the window length $L = 468$ is taken as the maximum window length, ensuring that $L \leq N/2$ and L is divisible by 12.

To properly identify the searched sinusoidal waves, eigenvalue plots, eigenvector scatter plots and the correlation matrix of the elementary components ($\boldsymbol{W} - matrix$) are used. Figure 8 shows several steps produced by approximately equal eigenvalues, each of which generates a pair of eigenvectors corresponding to a sine wave. This is confirmed in Fig. 9, which shows 2 almost regular polygons ET1–2 and ET3–4 corresponding to periods of 12 and 6, which occur due to seasonality and are clearly explained by the periodogram in Fig. 7.

The components are organised according to the order of the periodogram values at these frequencies (Golyandina and Korobeynikov 2014). The correlation matrix in Fig. 10 demonstrates that the indicated component pairs show high within-pair

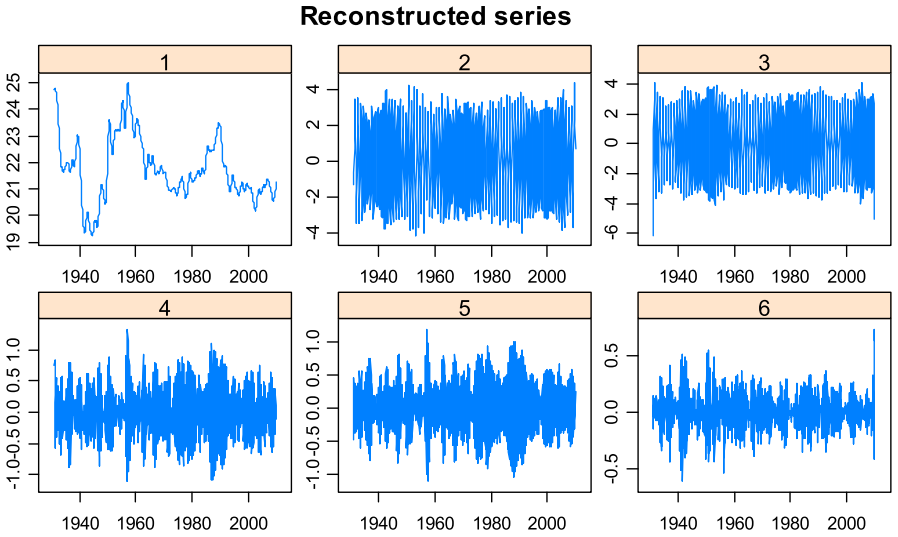


Fig. 5 Elemental reconstructed series in the 1st stage (L = 12) of the initial TS

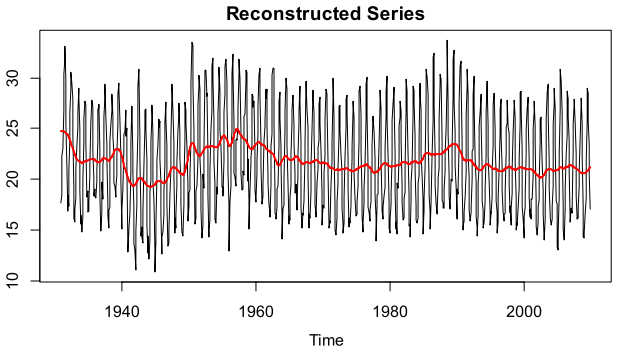


Fig. 6 Initial series and estimated trend in the 1st stage (L = 12)

correlations, but between them, the correlation is almost zero. Some pairs of eigentriples that satisfy the same properties are observed, but they are referred to as noise because identifying the period to which they correspond is not interpretable for monthly data.

Thus, from the previous identification, the seasonality is extracted by reconstructing the ET1–4 clustering result. As shown in Fig. 11, the original TS corresponds to the residual of the 1st stage, TS F1 is the extracted seasonality and TS Residuals is the final noise or residual generated in the 2nd stage. The noise residuals obtained are heterogeneous (Golyandina and Korobeynikov 2014).

The resulting sequential SSA decomposition is shown in Fig. 12.

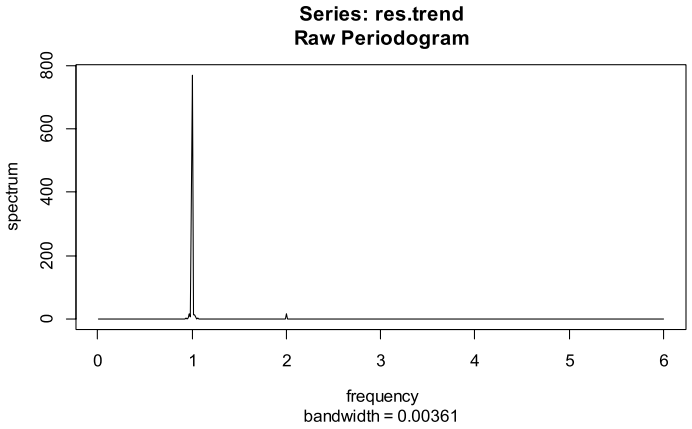


Fig. 7 2nd stage periodogram of the 1st stage residual

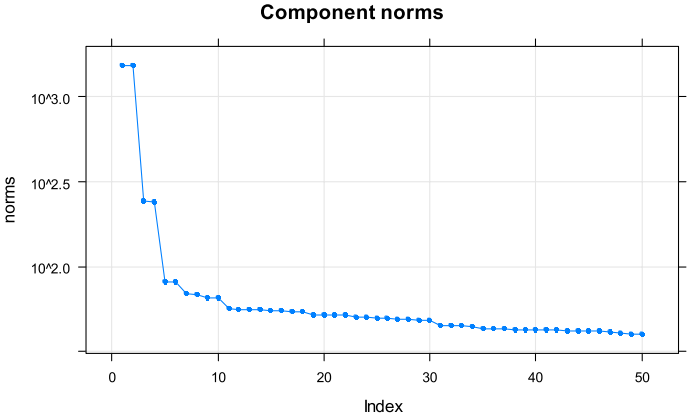


Fig. 8 2nd stage: Eigenvalues (L = 468)

4.4 Parameter estimation for the representation

Considering the trend, seasonality and noise TS extracted through sequential SSA, the parameters that correspond to each one are estimated, as indicated in steps 3.1.2 of the proposed algorithm. The parameters estimated by modelling each component of the initial TS are shown in Table 1.

The procedure is applied to the distinct TS considered in the dataset, and a feature vector is constructed for each TS.

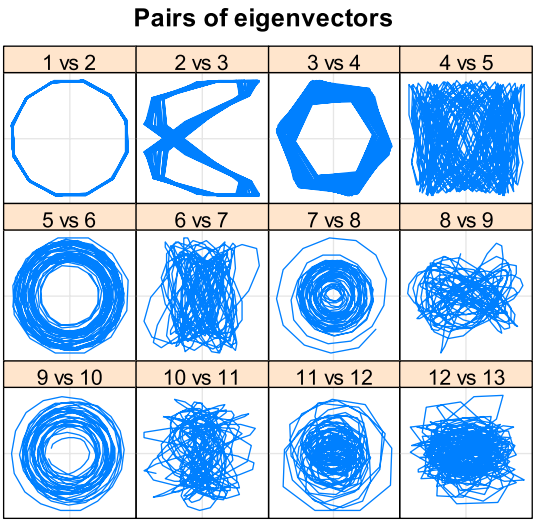


Fig. 9 2nd stage: Scatter plots for pairs of eigenvectors (L = 468)

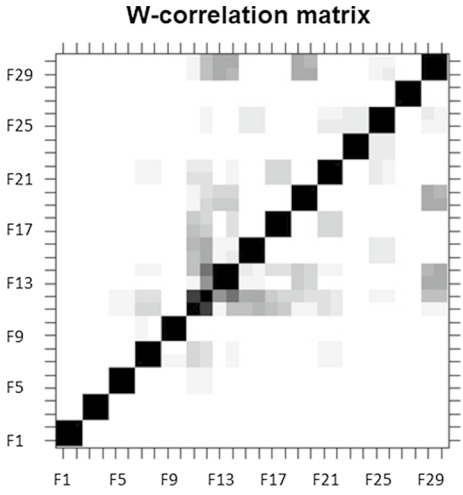


Fig. 10 2nd stage: W-correlation matrix (L = 468)

4.5 TS clustering

When clustering methods are applied to any dataset (with random structures or not), the data is divided into groups. Thus, the clustering tendency of the new dataset obtained with the feature vectors of each TS is evaluated. The Hopkins statistic test evaluated the spatial randomness of the data by measuring the probability that the dataset was generated by uniform data distribution, enabling it to be used to evaluate the clustering tendency of a dataset (Cross and Jain 1982; Banerjee and Davé 2004). If the value of such a statistic is close to 0.5, the data is considered random, but if it is close to 1.0, it indicates that the dataset contains

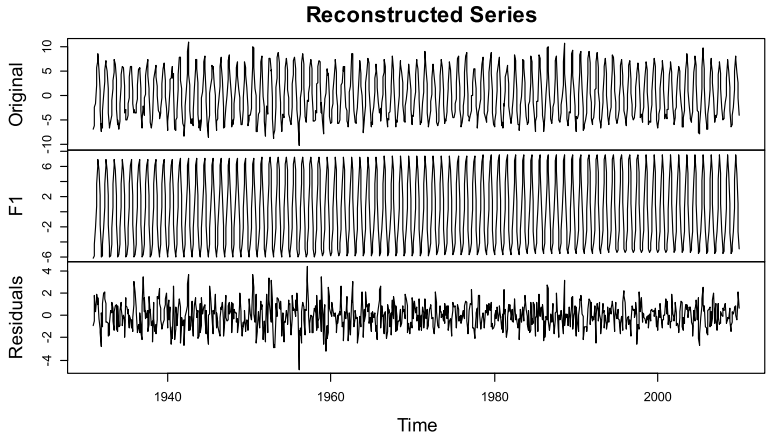


Fig. 11 2nd stage: TS Residual of 1st stage and extracted seasonal component (L = 468)

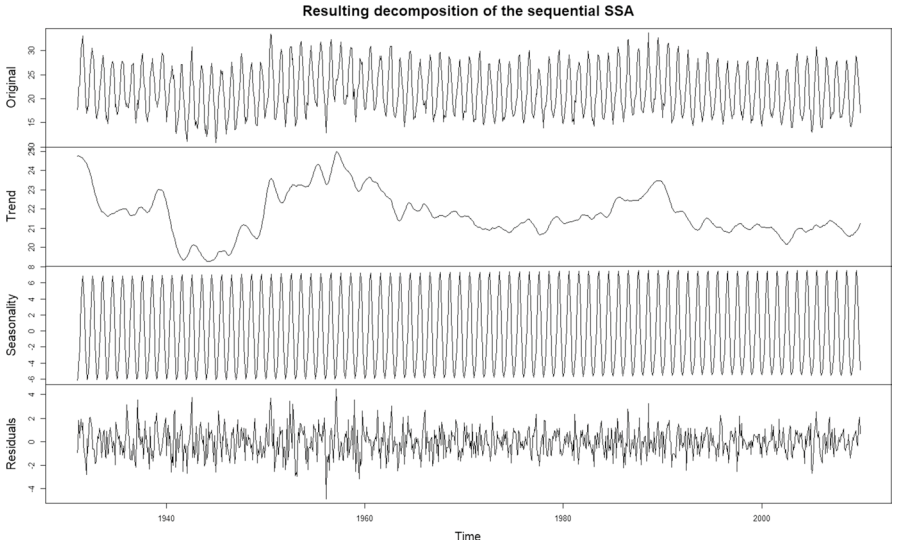


Fig. 12 Original TS and its trend-periodic residual decomposition

very well-defined clustered data. For the dataset, a Hopkins statistic value of 0.8550996 was obtained, which indicates that the dataset has natural clusters.

Since the dataset shows a clustering tendency, this study proceeds with clustering. Initially, four clustering algorithms, i.e., k-means, k-medoids, HA and SOM, are compared to the dataset. Internal validation measures of clustering are used.

Internal clustering measures the goodness of a clustering structure without taking into account external information that is not present in the data (Kremer et al. 2011; Song and Zhang 2008; Brun et al. 2007). This approach is usually based on compactness and separation criteria because clustering aims to arrange similar objects

Table 1 Parameters associated the trend, seasonality and noise components *with the SSA-extracted components*: they will be used to obtain the feature vectors for the clustering

Component	Parameters	
Trend	μ	β
	22.2145675	−0.0010091

Component	Parameters	
Seasonality	c_1	c_2
	−4.8043661	−4.3305937

Component	Parameters			
Noise	φ_1	φ_2	φ_3	φ_{12}
	0.1817	−0.150	−0.0731	−0.0342

within a single cluster and distinct objects in different clusters (Kraus et al. 2011; Zhao and Karypis 2002). Three such internal validation indices are selected: the silhouette index (Rousseeuw 1987), the Dunn index (Dunn 1974) and the connectivity index (Liu et al. 2013). These indices suggest that among the applied algorithms, the best algorithm for clustering data into two groups is HA. However, although there are no objective criteria for choosing the number of clusters, when the popular silhouette (Rousseeuw 1987), elbow (Thorndike 1953) and GAP (Tibshirani et al. 2001) methods are applied, they estimate that the optimal number of clusters is 3. Looking at the information extracted from the selected internal validation indices with respect to three clusters, the methods to consider include HA and k-means clustering, as noted in Table 2. The connectivity index corresponds to the extent to which elements are placed in the same cluster as their nearest neighbours in the data space and should be minimised. In contrast, the Dunn index should be maximised, since, according to its calculation, a dataset has compact and well-separated clusters if the diameter of the clusters is small and the distance between clusters is large. The silhouette index should also be maximised because it measures how well an observation is clustered, in addition to estimating the average distance between clusters.

Considering what is suggested by these indices, in order to improve the clustering results, a hybrid method, called hierarchical k-means (hkmeans), was selected to cluster the TS according to their extracted features. The hybrid hkmeans method combines the two algorithms suggested by the indices in three steps. In the first step, it calculates the hierarchical clustering results and cuts the tree into k clusters. In the second step, it calculates the centre or mean of each group. In the third step, it calculates the k-means using the set of cluster centres defined in the previous step as the centre of the initial clusters (Kassambara 2017; Lee et al. 2010). Although k-means is one of the most popular clustering algorithms, it has some limitations; for example, the number of clusters needs to be specified in advance, the initial centroids need to be selected randomly, and the final clustering solution is sensitive to that initial random selection of centroids. However, for hkmeans, the selection of the

Table 2 Internal measures of cluster validation: the connectivity, Dunn and Silhouette indices of the clustering results of Hierarchical, k-means, PAM and SOM for the IP, which measure the compactness and stability of clustering

Clustering methods	Validation measures	Value for cluster sizes				
		2	*3*	4	5	6
Hierarchical	Connectivity	*2.9290*	**5.8579**	11.6159	57.8544	143.8742
	Dunn	*0.5797*	**0.2311**	0.1918	0.0650	0.0571
	Silhouette	*0.5592*	0.2808	0.1714	0.1702	0.2620
k-Means	Connectivity	180.9877	218.1381	344.0829	342.5528	503.3266
	Dunn	0.0244	0.0263	0.0178	0.0257	0.0257
	Silhouette	0.1925	**0.2955**	0.2439	0.2451	0.1880
PAM		197.5643	303.2500	351.5698	388.8127	522.7710
	Dunn	0.0249	0.0229	0.0234	0.0294	0.0294
	Silhouette	0.2719	0.2462	0.2326	0.2390	0.1904
SOM	Connectivity	194.5016	222.0254	361.2802	496.3302	531.1806
	Dunn	0.0269	0.0210	0.0178	0.0178	0.0217
	Silhouette	0.2729	0.2908	0.2524	0.1909	0.1997

*Values in italics suggest two clusters with HA, and values in bold indicate that if three clusters are considered HA or k-means should be used. On this basis hk means with three clusters has been considered

initial centroids for k-means is defined using the hierarchical clustering result, which reduces these limitations and improves the final results of k-means clustering.

Here, hkmeans is applied to the set formed by the feature vectors of the TS. In the first step, an HA algorithm with Ward's linkage method is considered. The final clusters are shown in Fig. 13.

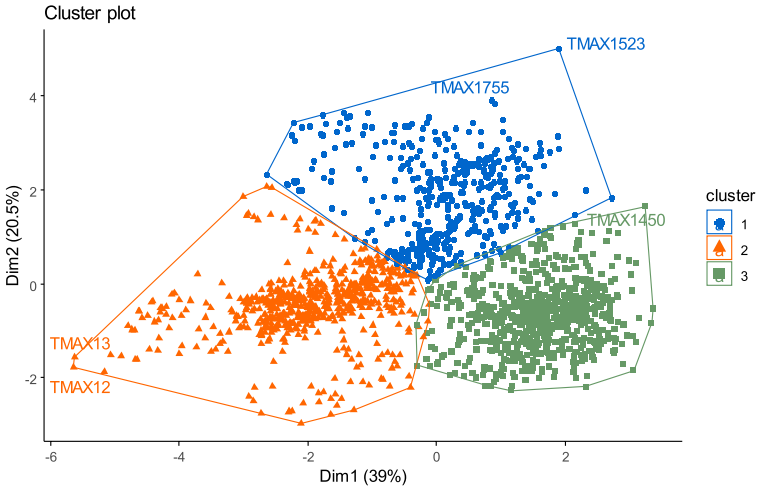


Fig. 13 Distribution of the final groupings of the TS

To retrieve information on the groups obtained, the centroids of each group are accessed, which are the averages of each characteristic used to define the characteristic vectors that represent each cluster. Thus, Cluster 1 is identified as containing points in IP with lower temperatures than the points in other clusters, although the point values of this cluster vary greatly over time, increasing over time. The seasonal component describing these TS typically corresponds to a cyclical variation, and is described by $c_1 = -4.62810$ and $c_2 = -5.20363$. In addition, the main autocorrelation of the residual component is negative. On the other hand, Cluster 2 contains points in IP where the average TMAX reaches the highest values, but the values decrease slightly over time, with a seasonal component whose amplitude of cyclical variation is defined by $c_1 = -4.96816$ and $c_2 = -5.13317$ on average. In addition, the main autocorrelation of the residual component is positive and has a higher absolute value than that of the other groups. Cluster 3 is characterised by points in IP where the average TMAX is at an intermediate level with respect to the temperatures of the other clusters and experiences a positive change over time. The seasonal component generally shows a cyclical variation described by $c_1 = -6.07224$ and $c_2 = -8.18860$, and its main autocorrelation in the residual component, which is negative, is the lowest absolute value compared to the other clusters. Figure 14 illustrates the distribution of the points in IP according to the clusters obtained (the grid is composed of longitudes and latitudes in UTM coordinates).

As can be seen in the map in Fig. 14, Cluster 1 includes areas of northern Spain, northern Portugal and southern France, as well as the Mediterranean region. In Cluster 2, most of the points are distributed in the Mediterranean Ocean, with the remaining points in southern Spain, northern Africa and the Atlantic Ocean. In Cluster 3, the points are primarily distributed in the Spanish and Portuguese territories.

When analysing the resulting patterns using the series obtained from the centroids of each cluster, the three zones in the IP are clearly differentiated: in the north and central zones, an increase in temperature was noted over time, whereas in the

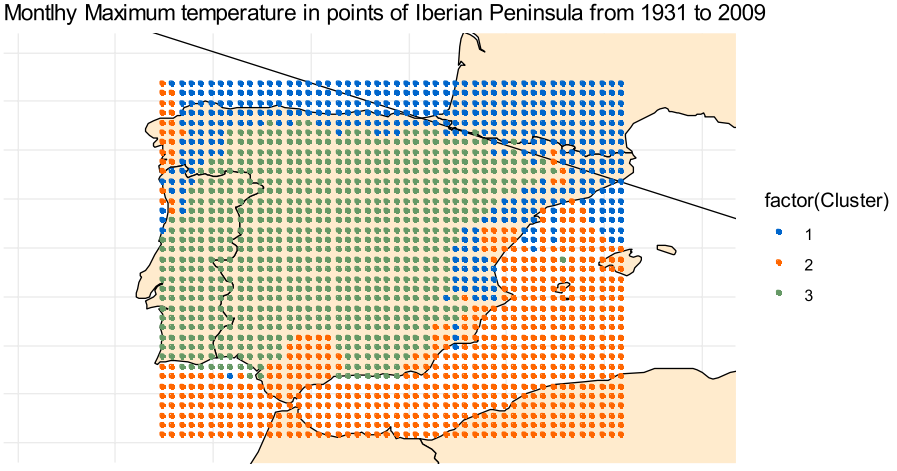


Fig. 14 Distribution of points in IP according to geographical location and clusters

south, a slight decrease was observed. The north zone of the IP, where the areas with the lowest TMAX are found, experienced a 0.2034 °C increase in its TMAX per decade between 1931 and 2009, whereas the central zone showed an increase of 0.135 °C per decade. In contrast, the south zone, where the areas with the highest TMAX are located, showed a slight decline.

Moreover, different seasonal variations were noted for each zone: the north zone shows its largest variations in winter months, whereas the central zone shows its variations in spring and autumn months. The south zone does not show any marked differences in monthly variation.

5 Conclusion

This paper proposes a novel method for clustering TS data that considers their trend, seasonality, and residual components. This approach involves representing TS data as feature vectors that are constructed by extracting the trend, seasonality and noise components using SSA decomposition. The results demonstrate reasonable groupings based on the defined features.

The proposed procedure can be applied to discover patterns in TS datasets, extract valuable information, and perform exploratory analysis on large TS datasets to support modelling efforts. The experiments allowed for spatial exploration and description of the variations of TMAX in the IP from 1931 to 2009 based on different zones defined by their trend and monthly variation.

Furthermore, this method could be used to test TS datasets with varying lengths or seasonal periods, as it is not restricted to TS data with uniform characteristics. Future research could explore the applicability of this method in multivariate TS analysis, as SSA can also be utilised for decomposition of such types of series.

Acknowledgements We express our gratitude to "Servicio de Desarrollos Climatológicos" of the Meteorological Spanish State Agency for providing the data used in the study, and to the Colombian Ministry of Science and the Technological University of Chocó for supporting the doctoral formation of Arnobio Palacios. The research is also supported by a Grant from Agencia Estatal de Investigación (PID2019-106433GB-I00/AEI/10.13039/501100011033), Spain.

Author contributions Conceptualization: AP, JLV; Methodology: AP; Software: AP; Validation: AP, JLV, MV; Formal analysis and investigation: AP, MV; Data curation: AP, JLV; Writing—original draft preparation: AP; Writing—review and editing: AP, JLV, MV; Supervision: JLV, MV.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest We declare that no financial interests or personal relationships influenced the work reported in this article.

References

Aghabozorgi S, Seyed Shirkhorshidi A, Ying Wah T (2015) Time-series clustering—a decade review. Inf Syst 53:16–38. https://doi.org/10.1016/j.is.2015.04.007

Banerjee A, Davé RN (2004) Validating clusters using the Hopkins statistic. In: IEEE international conference on fuzzy systems, 2004, vol 1, pp 149–153. https://doi.org/10.1109/FUZZY.2004.1375706

Bro R, Smilde AK (2003) Centering and scaling in component analysis. J Chemom 17(1):16–33. https://doi.org/10.1002/cem.773

Brun M, Sima C, Hua J, Lowey J, Carroll B, Suh E, Dougherty ER (2007) Model-based evaluation of clustering validation measures. Pattern Recognit 40(3):807–824. https://doi.org/10.1016/j.patcog.2006.06.026

Calheiros T, Pereira MG, Nunes JP (2021) Assessing impacts of future climate change on extreme fire weather and pyro-regions in Iberian Peninsula. Sci Total Environ 754:142233. https://doi.org/10.1016/j.scitotenv.2020.142233

Cross GR, Jain AK (1982) Measurement of clustering tendency. IFAC Proc 15(1):315–320. https://doi.org/10.1016/s1474-6670(17)63365-2

Dosio A, Mentaschi L, Fischer EM, Wyser K (2018) Extreme heat waves under 1.5 °C and 2 °C global warming. Environ Res Lett 13(5):054006. https://doi.org/10.1088/1748-9326/aab827

Dunn JC (1974) Well-separated clusters and optimal fuzzy partitions. J Cybern 4(1):95–104

Ergüner Özkoç E (2021) Clustering of time-series data. In: Birant D (ed) Data mining—methods, applications and systems. https://doi.org/10.5772/intechopen.84490

Gebremichael HB, Raba GA, Beketie KT, Feyisa GL, Siyoum T (2022) Changes in daily rainfall and temperature extremes of upper Awash Basin, Ethiopia. Sci Afr 16:e01173. https://doi.org/10.1016/j.sciaf.2022.e01173

Golyandina N (2010) On the choice of parameters in Singular Spectrum Analysis and related subspace-based methods. Stat Interface 3(3):259–279

Golyandina N, Korobeynikov A (2014) Basic singular spectrum analysis and forecasting with R. Comput Stat Data Anal 71:934–954. https://doi.org/10.1016/j.csda.2013.04.009

Golyandina N, Nekrutkin V, Zhigljavsky AA (2001) Analysis of time series structure: SSA and related techniques. CRC Press, Boca Raton

Golyandina N, Pepelyshev A, Steland A (2012) New approaches to nonparametric density estimation and selection of smoothing parameters. Comput Stat Data Anal 56:2206–2218. https://doi.org/10.1016/j.csda.2011.12.019

Golyandina N, Korobeynikov A, Zhigljavsky A (2018) Singular spectrum analysis with R. Springer, Berlin

Guo C, Jia H, Zhang N (2008) Time series clustering based on ICA for stock data analysis. In: 2008 International conference on wireless communications, networking and mobile computing, WiCOM 2008, 2008, pp 1–4. https://doi.org/10.1109/WiCom.2008.2534

Hartigan JA, Wong MA (1979) Algorithm AS 136: a K-means clustering algorithm. Appl Stat 28(1):100. https://doi.org/10.2307/2346830

Huhtala Y, Karkkainen J, Toivonen HTT (1999) Mining for similarities in aligned time series using wavelets. In: Data mining and knowledge discovery: theory, tools, and technology, vol 3695. https://doi.org/10.1117/12.339977

Huiting L, Zhiwei N, Jianyang L (2006) Time series similar pattern matching based on empirical mode decomposition. In: Proceedings—ISDA 2006: sixth international conference on intelligent systems design and applications, 2006, vol 1(050460402), pp 644–648. https://doi.org/10.1109/ISDA.2006.273

IPCC (2014) Climate change 2014: mitigation of climate change. Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, IPCC. Cambridge University Press. https://scholar.google.com/scholar_lookup?title=

Climate%20change%202014&publication_year=2014&author=IPCC&author=K.P.%20R&author=A.M.%20L

IPCC: Masson-Delmotte V, Zhai P, Chen Y, Goldfarb L, Gomis MI, Matthews JBR, Berger S, Huang M, Yelekçi O, Yu R, Zhou B, Lonnoy E, Maycock TK, Waterfield T, Leitzell K, Caud N (2021) In: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (ed) Climate Change 2021: the physical science basis. IPCC. www.ipcc.ch

Kassambara A (2017) Multivariate analysis I: practical guide to cluster analysis in R. In: Unsupervised machine learning. Taylor & Francis Group, New York, p 188

Keogh E, Smyth P (1997) A probabilistic approach to fast pattern matching in time series databases. In: Proceedings of the 3rd international conference of knowledge discovery and data mining, M(1994), 1997, pp 52–57. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+Probabilistic+Approach+to+Fast+Pattern+Matching+in+Time+Series+Databases#0

King AD, Karoly DJ (2017) Climate extremes in Europe at 1.5 and 2 degrees of global warming. Environ Res Lett 12(11):114031. https://doi.org/10.1088/1748-9326/aa8e2c

Kohonen T, Oja E (1996) Engineering applications of the self-organizing map. https://doi.org/10.1109/5.537105

Kraus JM, Müssel C, Palm G, Kestler HA (2011) Multi-objective selection for collecting cluster alternatives. Comput Stat 26(2):341–353. https://doi.org/10.1007/s00180-011-0244-6

Kremer H, Kranen P, Jansen T, Seidl T, Bifet A, Holmes G, Pfahringer B (2011) An effective evaluation measure for clustering on evolving data streams. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, 2011, pp 868–876. https://doi.org/10.1145/2020408.2020555

Kuglitsch FG, Toreti A, Xoplaki E, Della-Marta PM, Zerefos CS, Türkeş M, Luterbacher J (2010) Heat wave changes in the eastern Mediterranean since 1960. Geophys Res Lett. https://doi.org/10.1029/2009GL041841

Lee AJT, Lin MC, Kao RT, Chen KT (2010) An effective clustering approach to stock market prediction. In: PACIS 2010—14th Pacific Asia conference on information systems, 2010, pp 345–354

Lee Y, Na J, Lee WB (2018) Robust design of ambient-air vaporizer based on time-series clustering. Comput Chem Eng 118:236–247. https://doi.org/10.1016/j.compchemeng.2018.08.026

Li H, Wei M (2020) Fuzzy clustering based on feature weights for multivariate time series. Knowl Based Syst 197:105907. https://doi.org/10.1016/j.knosys.2020.105907

Liu Y, Li Z, Xiong H, Gao X, Wu J, Wu S (2013) Understanding and enhancement of internal clustering validation measures. IEEE Trans Cybern 43(3):982–994. https://doi.org/10.1109/TSMCB.2012.2220543

Lorenzo MN, Alvarez I (2022) Future changes of hot extremes in Spain: towards warmer conditions. Nat Hazards 113(1):383–402. https://doi.org/10.1007/s11069-022-05306-x

Lukasová A (1979) Hierarchical agglomerative clustering procedure. Pattern Recognit 11(5–6):365–381. https://doi.org/10.1016/0031-3203(79)90049-9

Molina MO, Sánchez E, Gutiérrez C (2020) Future heat waves over the Mediterranean from an Euro-CORDEX regional climate model ensemble. Sci Rep 10(1):8801. https://doi.org/10.1038/s41598-020-65663-0

Möller-Levet CS, Klawonn F, Cho K, Wolkenhauer O (2003) Fuzzy clustering of short time-series and unevenly distributed sampling points. Adv Intell Data Anal 2810:330–340

Park HS, Jun CH (2009) A simple and fast algorithm for K-medoids clustering. Expert Syst Appl 36(2):3336–3341. https://doi.org/10.1016/J.ESWA.2008.01.039

Rani S, Sikka G (2012) Recent techniques of clustering of time series data: a survey. Int J Comput Appl 52(15):1–9. https://doi.org/10.5120/8282-1278

Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20(C):53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Russo S, Sillmann J, Fischer EM (2015) Top ten European heatwaves since 1950 and their occurrence in the coming decades. Environ Res Lett 10(12):124003. https://doi.org/10.1088/1748-9326/10/12/124003

Samset BH, Fuglestvedt JS, Lund MT (2020) Delayed emergence of a global temperature response after emission mitigation. Nat Commun 11(1):3261. https://doi.org/10.1038/s41467-020-17001-1

Shi Y, Li B, Du G, Dai W (2021) Clustering framework based on multi-scale analysis of intraday financial time series. Physica A. https://doi.org/10.1016/j.physa.2020.125728

Song M, Zhang L (2008) Comparison of cluster representations from partial second- to full fourth-order cross moments for data stream clustering. In: Proceedings—IEEE international conference on data mining, ICDM, 2008, pp 560–569. https://doi.org/10.1109/ICDM.2008.143

Tebaldi C, Debeire K, Eyring V, Fischer E, Fyfe J, Friedlingstein P, Knutti R, Lowe J, O'Neill B, Sanderson B, van Vuuren D, Riahi K, Meinshausen M, Nicholls Z, Tokarska KB, Hurtt G, Kriegler E, Lamarque J-F, Meehl G et al (2021) Climate model projections from the Scenario Model Intercomparison Project (ScenarioMIP) of CMIP6. Earth Syst Dyn 12(1):253–293. https://doi.org/10.5194/esd-12-253-2021

Thorndike RL (1953) Who belongs in the family? Psychometrika 18(4):267–276

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc B 63(2):411–423. https://doi.org/10.1111/1467-9868.00293

Vicedo-Cabrera AM, Guo Y, Sera F, Huber V, Schleussner C-F, Mitchell D, Tong S, de Coelho MSZS, Saldiva PHN, Lavigne E, Correa PM, Ortega NV, Kan H, Osorio S, Kyselý J, Urban A, Jaakkola JJK, Ryti NRI, Pascal M et al (2018) Temperature-related mortality impacts under and beyond Paris Agreement climate change scenarios. Clim Change 150(3–4):391–402. https://doi.org/10.1007/s10584-018-2274-3

Wang X, Hyndman R (2006) Characteristic-based clustering for time series data. Data Min Knowl Discov 13:335–364. https://doi.org/10.1007/s10618-005-0039-x

Wang C, Wang XS (2000) Supporting content-based searches on time series via approximation. In: Proceedings of the international conference on scientific and statistical database management, SSDBM, 2000, pp 69–81. https://doi.org/10.1109/ssdm.2000.869779

Warren Liao T (2005) Clustering of time series data—a survey. Pattern Recognit 38(11):1857–1874. https://doi.org/10.1016/j.patcog.2005.01.025

Xiao Y, Liu JJ, Hu Y, Wang Y, Lai KK, Wang S (2014) A neuro-fuzzy combination model based on singular spectrum analysis for air transport demand forecasting. J Air Transp Manag 39:1–11. https://doi.org/10.1016/j.jairtraman.2014.03.004

Zhao Y, Karypis G (2002) Evaluation of hierarchical clustering algorithms for document datasets. In: International conference on information and knowledge management, proceedings, 2002, pp 515–524. https://doi.org/10.1145/584792.584877

Zhigljavsky A (2011) Singular spectrum analysis for time series. In: International encyclopedia of statistical science. Springer, Berlin. https://doi.org/10.1007/978-3-642-04898-2_521

Authors and Affiliations

Arnobio Palacios Gutiérrez[1,2] · Jose Luis Valencia Delfa[1] · María Villeta López[1]

✉ Arnobio Palacios Gutiérrez
arnobiop@ucm.es

Jose Luis Valencia Delfa
joseval@ucm.es

María Villeta López
mvilleta@ucm.es

[1] Faculty of Statistical Studies, Complutense University of Madrid, 28040 Madrid, Spain

[2] Group Valoración y Aprovechamiento de la Biodiversidad, Technological University of Chocó, Quibdó, Chocó, Colombia

Capítulo 4

Capítulo 4

Agrupación de series temporales multivariantes

Resumen: *Este capítulo presenta un método de regionalización climática que tiene en cuenta las interacciones de varias escalas en los procesos del clima para la identificación de patrones de temperaturas extremas y de precipitaciones en España. El procedimiento, aquí propuesto, permite agrupar series temporales multivariantes bajo un enfoque de análisis multiescala, destacando la importancia del orden temporal en la construcción de los vectores de características.*

4.1. Objetivos

Definir regiones climáticas diferenciadas en España de acuerdo con patrones de cambio característicos que expliquen la evolución de las temperaturas y precipitaciones extremas durante 1951 y 2021.

Generar un procedimiento de agrupación de series temporales que además de considerar múltiples variables climáticas, permita la integración de diferentes escalas temporales en las que se puedan representar los procesos climáticos, y que dé relevancia al orden temporal en la reducción de la dimensionalidad para el agrupamiento.

4.2. Metodología

El presente enfoque se inspira en técnicas del ámbito financiero y se basa en la adaptación de una metodología de análisis multiescala específicamente para los estudios climáticos. Por lo que aquí, es fundamental considerar la definición de escala temporal, ya que es clave para el desarrollo del enfoque. Tal como se estableció en el capítulo 1, esta se refiere a la unidad de tiempo característica utilizada para representar los datos de una serie temporal, permitiendo definir tanto la resolución como el período en el que se analizan las variaciones de los datos.

El procedimiento consiste en reducir la dimensionalidad de series temporales multivariantes de temperaturas y precipitaciones extremas para agruparlas y luego analizar los patrones identificables. La reducción de la dimensionalidad, aquí empleada, convierte cada serie temporal en una escala de tiempo más grande, seleccionado valores específicos en el orden de ocurrencia por cada segmento. Estas secuencias se unen cronológicamente hasta conformar nuevas series reducidas para cada variable. A continuación, se emplea una distancia multivariante definida que permite obtener matrices de similitudes entre las series temporales en base a las nuevas series. Luego se utilizan diferentes algoritmos de clustering para agrupar las series teniendo en cuenta las matrices de similitudes. Posteriormente, se comparan los resultados en base a índices de validación lo que permite seleccionar las agrupaciones finales. Por último, se obtienen los prototipos de clustering y se analizan con base en lo siguiente.

Análisis de tendencias: Para evaluar las tendencias de las temperaturas extremas y de la precipitación para cada zona se ha empleado la prueba de Man-Kendall (Mann, 1945; Kendall, 1975), que es una prueba no paramétrica que se utiliza habitualmente para identificar tendencias monótonas en series de datos medio ambientales o climáticos. Esta, es una prueba robusta, puesto que no se ve afectada por la discontinuidad de los datos o los valores atípicos (Jaagus, 2006), de modo que no se requiere que la serie de datos se distribuya normalmente. No obstante, esta prueba si se ve afectada por las autocorrelaciones de los datos, que deben eliminarse para evitar errores de tipo I (Wang et al., 2015), por lo que antes de aplicar la prueba a series con autocorrelaciones significativas, debe aplicarse el proceso de preblanqueamiento a la serie (Hamed, 2009). El preblanqueamiento es un procedimiento que se realiza antes de aplicar un test de tendencia, con el objetivo

de eliminar la influencia de la autocorrelación en los dato, y para su aplicación, primero se identificó si los valores de cada serie estaban correlacionados con valores previos, luego se ajustó un modelo ARIMA que capturara la dependencia temporal y finalmente tras la extracción de los residuos se obtuvo una nueva serie sin autocorrelación, sobre la cual se aplicó el test de tendencia. Además de ello, se empleó el estimador de pendiente de Sen (Sen, 1968) para describir la magnitud del cambio de las temperaturas extremas y de las precipitaciones.

Análisis de los eventos de sequias: Para estudiar los eventos de sequias en cada zona identificada, durante el periodo de estudio, se calculó el SPI (Mckee et al., 1993) a una escala temporal de 12 empleando un kernel de tipo rectangular, con desplazamiento 0, y con ajuste de los datos a una distribución Gamma. Para identificar los eventos de sequía definimos un umbral de detección de -1 de acuerdo con los criterios establecidos por Mckee et al. (1993) para determinar el principio y el fin de un evento de sequía, a partir de lo cual, se calcularon tres métricas de sequía (magnitud, intensidad y duración), las cuales se analizaron en el periodo de tiempo de forma mensual y por década. En general, la *duración* corresponde al número de meses consecutivos en los que el SPI está por debajo del umbral establecido; la *intensidad* es la suma de los valores absolutos del SPI durante el período de sequía, y la *magnitud de la sequía* se obtiene al multiplicar la duración por la intensidad acumulada, indicando la severidad del evento.

4.3. Resumen de resultados

Se tuvieron en cuenta 16,156 series temporales multivariantes de temperaturas extremas y precipitaciones distribuidas en toda España en una red de 5x5km^2. La agrupación mediante análisis multiescala permitió revelar patrones significativos de temperaturas y precipitaciones extremas en toda España. Téngase en cuenta que, al hablar de precipitaciones extremas, nos referimos a eventos inusuales asociados a la precipitación, como los valores de la precipitación máxima diaria anual o las sequías. A nivel regional, España se dividió en doce zonas climáticas distintas, permitiendo identificar que la zona comprendida por Andalucía oriental y el interior del sur de Murcia presentaron el mayor crecimiento de las temperaturas extremas máximas durante 1951 y 2021. Al mismo tiempo se identificaron algunas zonas, principalmente en el sur de España y en las regiones de montaña media, que vienen experimentando un aumento significativo de la duración de las sequías.

En general, se observó una tendencia significativa al calentamiento en toda España, evidenciada a partir de temperaturas extremas. Este comportamiento sugiere la presencia de cambios en eventos extremos, como olas de calor o frío intenso. La tasa de aumento de las temperaturas máximas anuales extremas varía según las zonas, alcanzando hasta 0.49°C por década, mientras que las temperaturas mínimas han registrado aumentos que oscilan entre 0,10 °C y 0,22 °C por decenio. En cuanto a precipitaciones se identificó un menor número de días lluviosos y a la vez un aumento de los valores máximos diarios anuales. De igual forma, se encontró que la duración y magnitud de las sequías han aumentado aproximadamente un 10% por década.

4.4. Conclusiones

El procedimiento de agrupamiento mediante análisis multiescala de series temporales proporcionó valiosos conocimientos sobre la dinámica espacial y temporal de los fenómenos climáticos extremos en España. Una fortaleza clave del enfoque reside en su capacidad para incorporar múltiples escalas temporales, permitiendo la identificación de tendencias graduales a largo plazo y cambios climáticos abruptos a corto plazo.

4.5. Publicación JCR

De acuerdo con la estructura de esta tesis, como cuerpo de este capítulo se incluye la publicación del presente estudio, identificable en las bases de datos científicas de acuerdo con la siguiente referencia.

Estado: Publicado

Año de Publicación: 2025

Nombre de la Revista: Natural Hazards

Editorial: Springer

Indicadores de Calidad: Factor de Impacto año 2023: 3.3; Rango: T1 – Primer Tercil, 64/254 de la categoría *GEOSCIENCES, MULTIDISCIPLINARY*; CiteScore: 6.6 (Scopus), SNIP 1.121

DOI: https://doi.org/10.1007/s11069-024-07082-2

Natural Hazards (2025) 121:7991–8009
https://doi.org/10.1007/s11069-024-07082-2

ORIGINAL PAPER

Identification of extreme temperature and precipitation patterns in Spain based on multiscale analysis of time series

Arnobio Palacios-Gutiérrez[1,2] · **Jose Luis Valencia-Delfa**[1] · **María Villeta**[1]

Received: 17 January 2024 / Accepted: 12 December 2024 / Published online: 21 January 2025

Abstract
Climate change is a matter of global interest. Great part of the research focuses on extreme climate events. This study investigates the patterns of change in extreme climate events in Spain, from 1951 to 2021. A total of 16,156 multivariate time series, based on monthly extreme data of maximum and minimum temperatures as well as precipitation, corresponding to 5×5-km^2 small areas, were analysed. The analysis procedure used reduces the dimensionality of the time series, clustering them and identifying changing patterns for each cluster. To carry out the research, a new multiscale analysis methodology is proposed, based on the financial field (Shi et al. in Phys A Stat Mech Appl, 567:71932008, 2021) and adapted to the climatic field: (1) considering the asymmetry of the precipitation series, (2) extending to the multidimensional case to obtain a better representation of the climatic reality, and (3) adapting the definition of distance between univariate series to multivariate series. The clustering results from applying the novel methodology, divide Spain into twelve zones. Among the results, it is worth highlighting that: extreme maximum temperatures have grown more in the cluster that covers Eastern Andalusia and the interior of Southern Murcia; there are three clusters corresponding mainly to Southern Spain and Middle Mountain that have significantly increased the duration of their droughts. Climate series show cyclical patterns and trends over multiple time scales, making the new methodology very useful. Furthermore, this approach allows detecting possible differences between the evolution of microclimates of small neighbouring territorial areas.

Keywords Climate events · Multiscale analysis · Clustering · Multivariate time series · SPI index · Spain

✉ Arnobio Palacios-Gutiérrez
arnobiop@ucm.es

Jose Luis Valencia-Delfa
joseval@ucm.es

María Villeta
mvilleta@ucm.es

[1] Faculty of Statistical Studies, Complutense University of Madrid, Avenida Puerta de Hierro, n° 1, 28040 Madrid, Spain

[2] Group Valoración y Aprovechamiento de la Biodiversidad, Technological University of Chocó, Quibdó, Chocó, Colombia

1 Introduction

Climate change represents global challenge, and its impact is becoming increasingly evident. The significant growth in the emission of greenhouse gases into the atmosphere, such as CO2, has led to an elevation in atmospheric temperatures. As outlined in the Sixth Assessment Report by the Intergovernmental Panel on Climate Change (IPCC 2021), the growth rate of Global Mean Surface Temperature (GMST) has accelerated, with a recorded rise of 1.1 °C in the period 2001–2021. Moreover, changes in precipitation patterns have also become apparent. Since the year 2000, there has been a 29% increase in the number and duration of droughts, as reported by the United Nations Convention to Combat Desertification (UNCCD 2022).

Fluctuations in temperature and precipitation can have large impacts on both society and ecosystems by modifying extreme weather events and hydrological patterns (Javadinejad et al. 2019a; Kuriqi et al. 2020; Młyński et al. 2021; Ebi et al. 2021; Abbass et al. 2022; Ostad-Ali-Askari et al. 2022). Therefore, investigating variations in these fundamental climatic variables is particularly important. Analyses of climatic indices obtained from various regions worldwide highlight the importance of studying the evolution of extreme temperatures and precipitation within the field of climate change (Javadinejad et al. 2019b; Fatahi Nafchi et al. 2021; Newell et al. 2021).

Most articles about extreme climate events analyse a single climate variable: temperature or precipitation (Wang et al. 2021). Furthermore, in the studies in which both variables are used, the research is mostly focused on a specific homogeneous region (Gebremichael et al. 2022). Unlike these studies, this research focuses on the identification of both types of extreme weather patterns across a country, Spain, which has great geographical diversity. Spain is a country located on the Iberian Peninsula in southwestern Europe, with diverse climate patterns across different regions (González-Hidalgo et al. 2021, 2024). Therefore, to identify regional climate patterns in Spain, a climate regionalization procedure has been carried out in this study.

One of the most widely used climate regionalization multivariate methods is Cluster Analysis (Li et al. 2022). The physical processes influencing climate variables, often operate under a wide range of time scales (Tessier et al. 1996). Many of clustering climate regionalization studies based on Multiscale Analysis of time series, turn around concepts such as Wavelet Transform and Multiscale Entropy (Agarwal et al. 2016; Roushangar et al. 2018; Li et al. 2022). To identify regional areas of Spain that follow the same patterns of change with regards to extreme maximum temperature (T_{max}), minimum temperature (T_{min}) and precipitation ($PRCP$) along 1951–2021, the present study introduces a novel approach to dimensionally reduction of multivariate time series before clustering them.

The new approach of Multiscale Analysis proposed in this paper, is based on the financial time series field (Shi et al. 2021) and adapted to the climatic field. This novel approach is easier to apply than methods mentioned before, it is valid to simultaneously analyse temperatures and precipitation, as well as to be applied to an unusually large number of climatological stations (16,156 in the present study). The adaptation is justified by the following reasons. Since the distributions of both temperatures and precipitations are not symmetrical (all the temperature and precipitation series in the study database, the mean exceeds the median), the median estimator better reflects the behaviour of these climatic variables than the mean estimator. Therefore, the mean, included in the method of Shi et al. (2021), has been replaced by the median. Furthermore, to obtain clusters that better represent the climatic reality, this paper carries out a multivariate study, with 3 observations (maximum,

minimum and median) for each of 3 climatic variables per temporal scale, instead of a univariate study (in the case of the financial method). Because of using multivariate analysis, it has also been necessary to adapt the definition of the distance between univariate time series to multivariate time series. Moreover, to characterize the obtained clusters, various climatic indices have been used to cover diverse aspects of climatology.

Comparing multiple data windows throughout this novel approach of Multiscale Analysis, it can be revealed whether climate variability is amplified or smoothed across different time scales. This approach also allows detecting moments when temperature or precipitation patterns change abruptly. Furthermore, in the present investigation, the time series are associated with very small areas 5×5-km^2, which is quite unusual. This fact shows that the new methodology carried out, allows many microclimates analysing jointly (16,156 in this case), and establishing possible differences between microclimates of small neighbouring areas.

In the following, Sect. 2 offers a description of the study area, and the methodologies employed. Section 3 presents the findings of the application of new clustering approach and describes the observed patterns of change in T_{max}, T_{min} and $PRCP$ for obtained regions in Spain. Finally, Sect. 4 summarizes main conclusions achieved in the present study.

2 Data and method

2.1 Study area and database

The study area is Spain, located in southwestern Europe, bordered by the Atlantic Ocean, the Mediterranean Sea, Portugal, and France (Fig. 1). The cooler temperatures of the Atlantic Ocean influence the northern and western regions of the country, while the Mediterranean Sea predominantly affects the south and east, where it moderates temperatures. Geography of Spain creates a defined north-to-south climatic gradient. Additionally, diurnal and

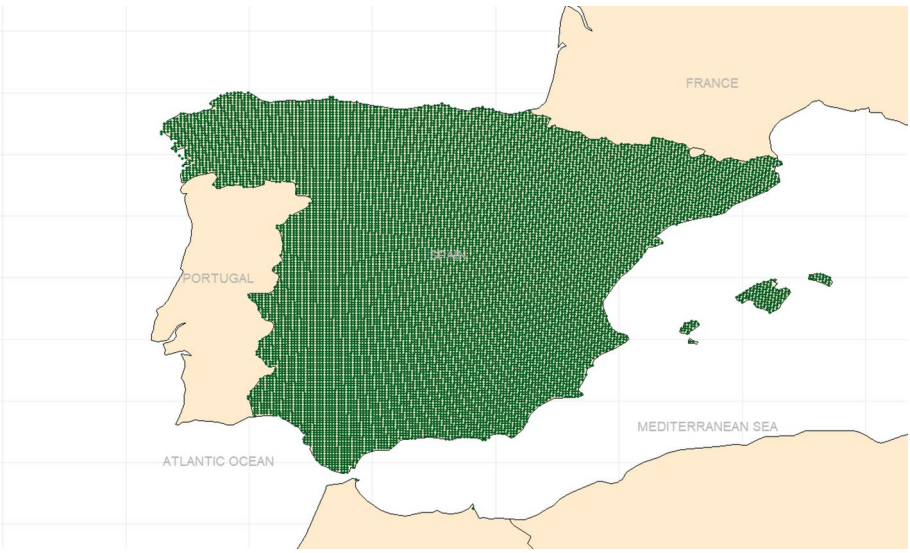


Fig. 1 Spatial distribution of the grid points in Spain

seasonal temperature gradients are observed, especially between coastal and inland areas (Rodrigo 2023).

Multivariate time series of temperature and precipitation (T_{max},T_{min} and *PRCP*) corresponding to the years 1951–2021 were used. The Climatological Development Service of the Spanish Meteorological Agency (AEMET) provided these series. The data were obtained through spatial interpolation using kriging on a 5×5-km^2 grid that covers the Spanish territory (Fig. 1). The data can be accessed at the following website: (https://www.aemet.es/es/serviciosclimaticos/cambio_climat/datos_diarios?w=2&w2=0).

For each month, T_{max} and T_{min} were calculated as the averages of the daily maximum and minimum values, respectively. For *PRCP*, the monthly average of daily precipitation was used, ensuring that months with different numbers of days are comparable. In total, the dataset includes 16,156 multivariate time series, with 852 observations per climate variable (one per month). These series cover all points of the grid in mainland Spain, the Balearic Islands, and the North African cities of Ceuta and Melilla, encompassing the entire Spanish territory except for the Canary Islands.

The time series exhibit strong seasonality. T_{max} and T_{min} temperatures reach their peak values in summer and decrease during winter, while *PRCP* shows its lowest values in summer and varies throughout the rest of the year. Figure 2 provides an example of the temporal evolution of the series at a single grid point.

2.2 Methodology

The main objective of this research is to develop a new territorial clustering method that allows identifying patterns of change in the variables T_{max}, T_{min} and *PRCP*, which differ across each generated group. These variations are highly relevant in climate analysis, as they capture extreme events such as heatwaves, cold spells, droughts, or periods of heavy precipitation.

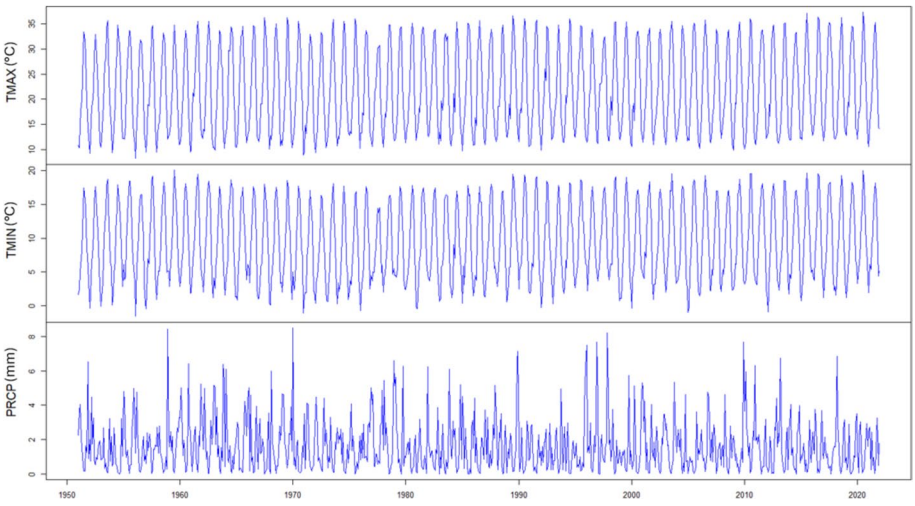


Fig. 2 Average maximum and minimum temperatures and monthly average of daily precipitation from 1951 to 2021 at a grid point of Spain

The methodological approach is based on the clustering of multivariate time series. To achieve this, we adapt the dimensionality reduction technique and similarity measure proposed by Shi et al. (2021), which effectively captures the multiscale effect. However, we employ a multivariate approach, as their method is univariate, and we incorporate the median instead of the mean in one of the parameters summarizing the information from the series in each time window. This adjustment addresses the common asymmetry present in climatic variables. This approach is useful because climate series exhibit cyclical patterns and trends across multiple time scales, often displaying local fluctuations, seasonal cycles, and long-term trends. This methodology allows for the identification of extreme values over multi-year periods and simplifies the clustering process.

2.2.1 Dimensionality reduction of T_{max}, T_{min} and *PRCP* series

To reduce dimensionality, the original monthly time series (T_{max}, T_{min} and $PRCP$) are converted to a larger time scale covering several years. Specifically, for the triennial scale, the maximum, minimum, and median values are taken every 36 months for each variable (T_{max}, T_{min} and $PRCP$) maintaining the order in which they appear. For example, if the first value is the median, followed by the maximum and then the minimum, the sequence will be median, maximum, minimum. This configuration may vary among variables for the same period. These combinations are then organized chronologically, generating new sequences of T_{max}, T_{min} and $PRCP$ for each point in the network. The chosen period must be a multiple of 12 (if the data are monthly) and can span several years. It is recommended to use windows of at least 36 months to capture longer trends than annual seasonal cycles, thus reflecting fluctuations associated with cyclical climatic phenomena or interannual variability, such as El Niño or La Niña. Comparing 3-year and 6-year windows can reveal whether variability is amplified or smoothed across different scales.

2.2.2 Distance between multivariate time series (T_{max}, T_{min}, *PRCP*)

For each scale S_e and each time series $y_j(T_{max}, T_{min}, PRCP)$, the Euclidean distance between two grid points z_i and z_r is calculated according to Eq. (1).

$$d_{z_i,z_r}(y_j, S_e) = \sqrt{\sum_{t=1}^{n_{scale}} \left(y_{z_i,j,t(s_e)} - y_{z_r,j,t(s_e)}\right)^2} \tag{1}$$

where t is the sequence of values of the series, which will depend on the scale. Since different scales are used that will have a variable number of summands, it is advisable to standardize them so that them all fall within the same range. A distance range from 0 to 100 has been established.

$$\widetilde{d}_{z_i,z_r}(y_j, S_e) = 100 \frac{d_{z_i,z_r}(y_j, S_e)}{\underset{w,x}{Max}\ \left(d_{z_w,z_x}(y_j, S_e)\right)} \tag{2}$$

The multivariate distance for a scale S_e is obtained by summing the weighted distances of each univariate series (Eq. 3):

$$D_{z_i,z_r}(S_e) = \sqrt{\sum_{j=1}^{3} w_j \left(\widetilde{d}_{z_i,z_r}(y_j, S_e)\right)^2} \tag{3}$$

When selecting the weights w_j we consider that T_{max} and T_{min} are more correlated with each other than with *PRCP;* thus, we assign $w_1 = w_2 = 0.3$ and $w_3 = 0.4$. The standardized univariate distances are squared to increase the weight of significant differences in any of the variables. Finally, we generate the multiscale distance matrix DM_{z_i,z_r} by summing the weighted distances of each scale (Eq. (4)).

$$DM_{z_i,z_r} = \sum_{e} w_{(e)} D_{z_i,z_r}(S_e) \tag{4}$$

2.2.3 Time series clustering

With the calculated multiscale multivariate distance between each pair of grid points, we apply three methodologically different clustering algorithms:

- Partitioning Around Medoids (K-medoids, PAM) (Kaufman and Rousseeuw 1990).
- Hierarchical Clustering using the Ward method to form groups of observations (HA) (Murtagh and Legendre 2014).
- Self-Organizing Maps (SOM) (Kohonen and Oja 1996). This is a competitive learning neural network.

To determine the optimal number of clusters for each algorithm, we use four internal validation indices (Desgraupes 2017):

- The C Index. This index compares the distances within clusters with the smallest and largest distances between points in the network. It ranges from 0 to 1, with values close to 0 indicating compact and well-separated clusters (Hubert and Schultz 1976; Desgraupes 2017).
- Baker-Hubert Gamma Index. This is an adaptation of the correlation index Γ between two data vectors of the same size; thus, values closer to 1 are preferred (Baker and Hubert 1975; Desgraupes 2017).
- McClain-Rao Index. This index was introduced by McClain and Rao (1975) and compares the average distances within clusters against the average distances between clusters. Logically, it should be as small as possible.
- Ray-Turi Index (Ray and Turi 1999). This index compares the mean squared distances of all points to the centroid of the group to which they belong, with the square of the minimum distance between centroids. The goal is also to minimize this index.

2.2.4 Analysis of change patterns by clusters

In each cluster, the evolution of the three time series T_{max}, T_{min} and *PRCP* is studied, evaluating trends using the non-parametric Mann–Kendall test (Mann 1945; Kendall 1975), which detects monotonic trends. This test is robust (Jaagus 2006); however, autocorrelations must be removed to avoid Type I errors (Wang et al. 2015). Pre-whitening is applied

to series with significant autocorrelations (Hamed 2009). For series with significant trends, the Theil-Sen slope estimator is used to determine the magnitude of changes (Sen 1968).

To analyse droughts, the Standardized Precipitation Index (SPI) (Mckee et al. 1993) is calculated on a 12-month scale, with a threshold of -1, following the criteria set by Mckee et al. (1993). A rectangular kernel distribution is employed, with a shift of 0 and data fitting to a Gamma distribution. Three drought metrics are analysed: magnitude, intensity, and duration, both monthly and by decade.

3 Results and discussion

This study provides a more detailed climate regionalization of Spain compared to previous works. For instance, Parracho et al. (2016) applied a k-means clustering analysis to divide the Iberian Peninsula into four regions based on precipitation regimes, while Rodrigo (2023) identified four regions based on winter indices and method of Ward. Our study extends the analysis to 16,156 series, a significantly higher number of series than previous studies, allowing for a more nuanced regional classification. This detailed regionalization has practical applications for land use planning, water policy and infrastructure projects by enabling strategies tailored to local climatic conditions. The methodology employed aligns with recent publications (Pereira et al. 2021) and emphasizes the importance of a fine-grained approach to capture current climate dynamics.

This section presents how climate clusters were obtained, describes the characteristics of each cluster, and analyses climate change patterns within the clusters.

3.1 Generation of climate clusters

Climate zones were generated by calculating distances between multivariate time series, as described in Sects. 2.2.1 and 2.2.2. The approach involved reducing dimensionality and calculating multivariate and multiscale distances for grid points across Spain. Two timescales were used (36 and 72 months), weighted equally to ensure balanced temporal analysis.

Subsequently, three clustering algorithms were applied: Hierarchical Agglomerative (HA), Partitioning Around Medoids (PAM), and Self-Organizing Maps (SOM), detailed in Sect. 2. A range of 3 to 14 clusters was considered, and the optimal number was determined using internal validation indices (Sect. 2.2.3). Figure 3 shows the variation of these indices with the number of clusters. The optimal clustering is reached when the index value is minimized, except for the Gamma index. The HA algorithm performed best, leading to a 12-cluster model.

Figure 4 shows how the clusters have been formed. The cut-off point for clustering is marked with a dashed line.

Geographical visualization of the clusters was achieved by color-coding grid points according to the assigned cluster. Figure 5 displays this spatial distribution, with the names of autonomous communities overlaid for clarity.

Table 1 summarizes key climate variables for each of the 12 clusters from 1951 to 2021, including the Annual Maximum Daily Temperature (AMxDT), Annual Average Monthly Maximum Temperature (AAMxMT), Annual Minimum Daily Temperature (AMnDT), Annual Average Monthly Minimum Temperature (AAMnMT), Annual Average Precipitation (AAP), and Annual Maximum Daily Precipitation (AAMdP).

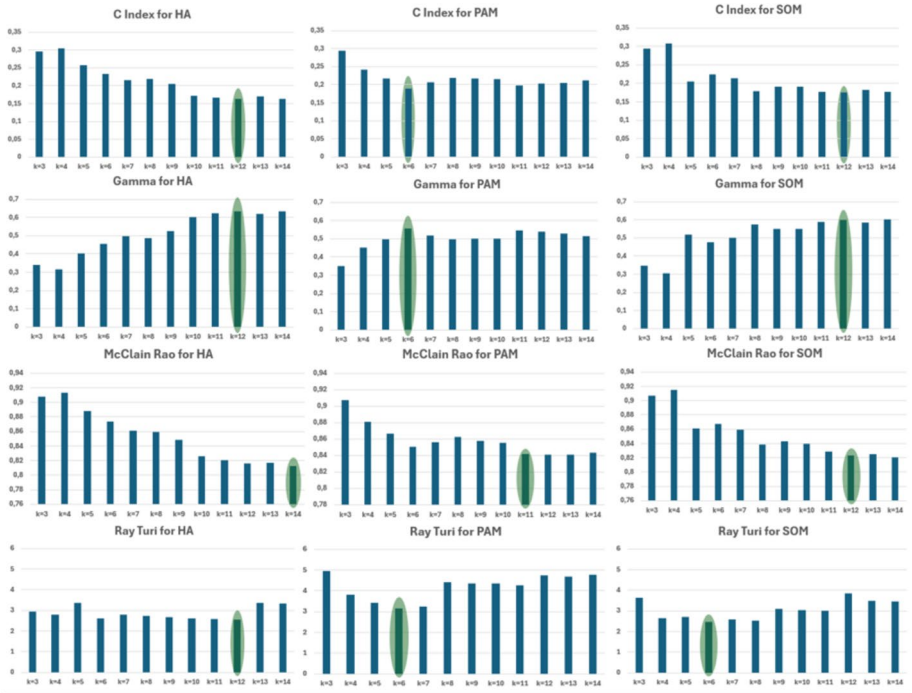


Fig. 3 Internal validation indices for different numbers of clusters (3–14)

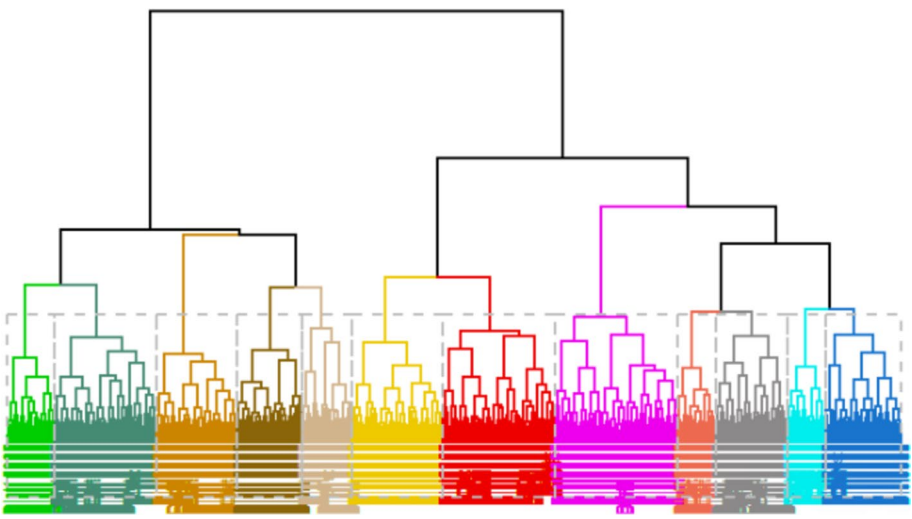


Fig. 4 Cluster Dendrogram

Considering Figs. 5 and 6 (which show annual averages of T_{max}, T_{min} and $PRCP$), as well as Table 1, the general characteristics of the 12 clusters can be identified. The clusters show clear geographical differences, based on altitude, latitude, longitude, and proximity

Fig. 5 Cluster-Based Geographic distribution of network points in Spain

Table 1 Annual Maximum Daily Temperature Average (AMxDT), Annual Average of Monthly Maximum Temperatures (AAMxMT), Annual Minimum Daily Temperature Average (AMnDT), Annual Average of Monthly Minimum Temperatures (AAMnMT), Average rain Annually (AAP) and Average Annual Maximum daily Precipitation (AAMdP), for each cluster in the period 1951–2021

Cluster	1	2	3	4	5	6	7	8	9	10	11	12
AMxDT	36.7	39.8	36.6	33.6	30.8	37.8	40.2	35.5	36.0	38.1	32.9	33.5
AAMxMT	21.7	23.2	20.6	16.2	14.1	20.0	22.2	19.2	18.0	20.4	17.1	16.1
AAMnDT	− 1.8	− 1.2	− 3.2	− 9.5	− 9.7	− 6.5	− 2.7	− 5.5	− 7.7	− 5.2	− 3.4	− 6.3
AAMnMT	10.6	10.8	8.9	4.2	3.3	7.2	9.8	7.8	5.2	8.4	7.2	6.2
AAP	450.8	647.4	503.7	682.6	1117.9	503.8	607.1	590.3	517.3	405.7	1397.3	1220.5
AAMdP	49.0	44.5	37.1	34.6	50.9	29.8	37.1	46.1	28.1	32.9	50.0	48.9

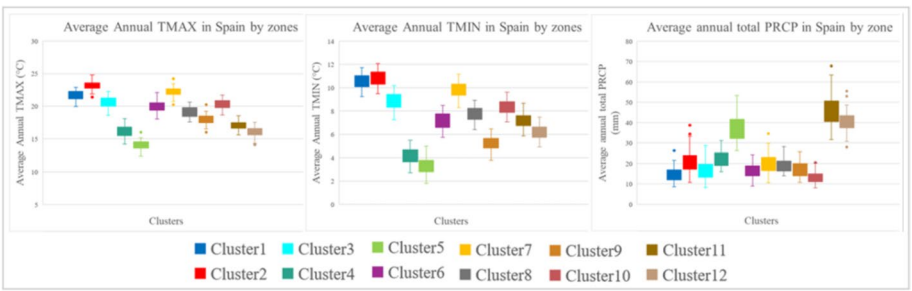


Fig. 6 Annual distributions of T_{max}, T_{min} and *PRCP* by cluster

to the sea. The following descriptions outline the clusters geographically according to their predominant regions, although they do not strictly follow administrative boundaries:

Cluster 1: Includes Melilla, the central and southern Mediterranean strip, and the Balearic Islands. This cluster is characterized by mild winters, hot summers, and high maximum daily precipitation, but relatively low annual precipitation.

Cluster 2: Covers Western Andalusia and southern Extremadura, characterized by very hot summers and mild winters, with higher precipitation compared to neighbouring clusters.

Cluster 3: Encompasses much of Eastern Andalusia, excluding the Sierra Nevada region and parts of the eastern coast. It has mild temperatures, with slightly higher precipitation than Cluster 1, but lower than Cluster 2.

Cluster 4: Extends across the Iberian System, parts of Castilla y León, Sierra de Cazorla, and other mountainous areas. It is characterized by cold winters and mild summers.

Cluster 5: Covers high mountain areas such as the Pyrenees, Sierra Nevada, the Central System, and the Cantabrian Range. It experiences the lowest temperatures in Spain, with high annual and extreme daily precipitation.

Cluster 6: Includes Castilla-La Mancha, Madrid, and southern Castilla y León. It has moderate temperatures compared to adjacent clusters, with milder winters than the northern regions and cooler summers than the eastern and southern areas. The annual rainfall is low.

Cluster 7: Encompasses Extremadura and the western part of Castilla-La Mancha. This area has the hottest summers on the peninsula, with very high temperatures.

Cluster 8: Comprises the northern Mediterranean region and parts of Aragón. It shows lower extreme daily precipitation compared to Cluster 1, with milder summers and colder winters.

Cluster 9: Located in the northwestern plateau, this cluster has cold winters and the lowest daily precipitation extremes.

Cluster 10: Covers the interior of Aragón, characterized by low altitude. It is the driest cluster, with very hot summers and low maximum daily precipitation.

Cluster 11: Includes Galicia and Asturias, characterized by the highest annual and daily precipitation.

Cluster 12: Situated along the eastern Cantabrian coast. Like Cluster 11, it is very rainy, but with warmer summers and colder winters.

This regionalization clearly separates Clusters 11 and 12 due to their high precipitation from the rest of Spain. The patterns follow logical gradients, with temperature generally increasing from north to south and precipitation decreasing from west to east.

3.2 Analysis of climate change patterns by cluster

To understand the climatic trends in each cluster, we analysed the centroid time series for T_{max}, T_{min} and $PRCP$. The Mann–Kendall test was applied to the data, with a prewhitening process to account for autocorrelation, and the slopes were calculated using Theil-Sen estimator. Table 2 presents the results, with a zero indicating that trend is not statistically significant, with the level of significance set at 0.05.

A significant increase in both maximum and minimum temperatures can be observed across most clusters. On average, daily maximum temperature extremes have risen between 0.14 °C and 0.49 °C per decade, with Cluster 3 exhibiting the highest increase, followed

Table 2 Linear rates of change per decade

Cluster	AMxDT	AAMxMT	AAMnDT	AAMnMT	AAP	AAMdP
1	0.36	0.17	0	0.14	0	2.27
2	0.33	0.15	0	0.14	−17.38	1.84
3	0.49	0.26	0	0.12	−25.19	0
4	0.31	0.25	0.24	0.20	0	0.88
5	0.24	0.14	0.36	0.22	−18.46	0.99
6	0.30	0.21	0	0.21	−13.90	0.62
7	0.27	0.17	0	0.12	−17.72	0
8	0.35	0.24	0	0.18	0	1.80
9	0.29	0.20	0	0.10	0	0
10	0.14	0.19	0	0.18	0	1.04
11	0.25	0.18	0.17	0.20	0	0.99
12	0.34	0.20	0.30	0.19	0	1.88

by Cluster 1 and Cluster 8. Curiously, Cluster 10, which has one of the highest maximum temperatures, has experimented the smallest absolute increase.

Annual average maximum temperatures have risen by 0.14 °C to 0.26 °C per decade, mainly driven by summer warming. Cluster 3 (Eastern Andalusia) and Cluster 4 (Low Mountain) displayed the largest increases. In the case of Cluster 3, this is mainly due to rising summer temperatures, whereas in Cluster 4 the increase is more related to winter temperatures. It is worth mentioning that Cluster 3 is warming at a faster rate than its neighbouring geographic Cluster 2. High mountain areas (Cluster 5) showed the smallest temperature increase, both in absolute and relative terms, with the rise concentrated mainly in winter temperatures.

Regarding extreme minimum temperatures, no significant increase in extremely cold days was detected. In fact, significant changes have been recorded in areas with historically low minimum temperatures, such as Cluster 4 and Cluster 5, as well as in wetter regions (Cluster 11 and Cluster 12).

Annual average minimum temperatures have also risen, although at a slower rate than maximum temperatures in absolute terms. However, the relative increase is slightly more pronounced, particularly in colder regions. These results are consistent with IPCC (2021) findings and studies by Lorenzo and Álvarez (2022), which highlight that extreme cold events are becoming less frequent, while extreme heat episodes are more common, especially in the Pyrenees and central regions, where the largest temperature increases have been recorded.

Our results differ somewhat from those of González-Hidalgo et al. (2021) and Peña-Angulo et al. (2021), who reported a greater increase in minimum temperatures compared to maximum temperatures. In absolute terms, we find that warm extremes and average maximum temperatures have risen more than minimum temperatures, except in high mountain areas. The lower trends reported by Peña-Angulo et al. (2021) may arise from their longer study period (100 years), which could smooth out the trends because before 1951 warming phases were weaker.

Annual precipitation has significantly decreased, ranging from 11 to 25 mm per decade. This reduction is most pronounced in mountainous regions (Cluster 4 and Cluster 5) and inland areas, including Andalusia, Extremadura, and Castilla-La Mancha (clusters 2,

3, 6, and 7). These results are in line with the work of González-Hidalgo et al. (2023) and Senent-Aparicio et al. (2023), who also reported a general reduction in precipitation. Our results show larger decreases (1.3–2.5 mm per year) than those reported by González-Hidalgo et al. (2024), probably due to our focus on specific clusters rather than the whole of Spain.

In contrast to the overall decline in total annual precipitation, maximum daily rainfall has shown significant increases. The only exception is Cluster 9 (Northwest of the Plateau), where neither precipitation variable exhibits significant changes. In the remaining clusters, either extreme rainfall events have become more frequent, or average precipitation has declined significatively, suggesting a trend toward more concentrated rainfall on specific days. This pattern is particularly evident in regions with heavy rainfall, such as Cluster 1 and Cluster 12, where extreme daily precipitation increases are more pronounced.

Meseguer-Ruiz et al. (2021) also found that the smallest reductions in precipitation occurred in the eastern and southeastern Mediterranean regions of Spain. In our analysis, no significant decrease is observed in Cluster 1, which can be associated with the increase in torrential rainfall events recorded in September and November, consistent with the positive trend in extreme rainfall intensity detected in this region.

3.3 Drought patterns across cluster

Figure 7 presents the Standardized Precipitation Index (SPI) at a 12-month scale for each cluster over the 1951–2021 period. This index measures deviations in accumulated

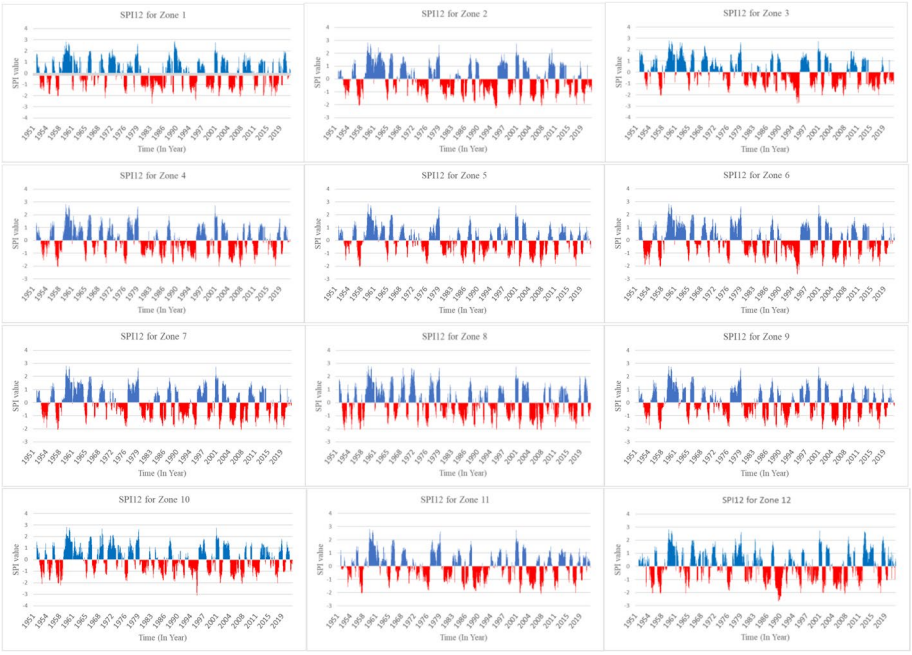


Fig. 7 Standardized Precipitation Index (SPI) at a 12-month scale for each cluster (1951–2021)

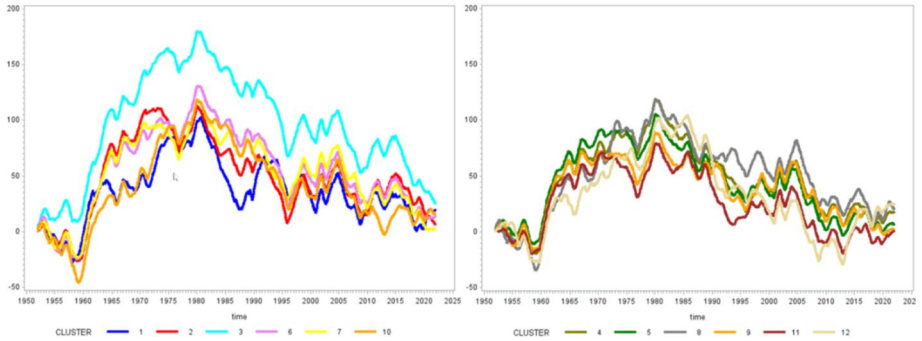


Fig. 8 Cumulative SPI Values by Cluster

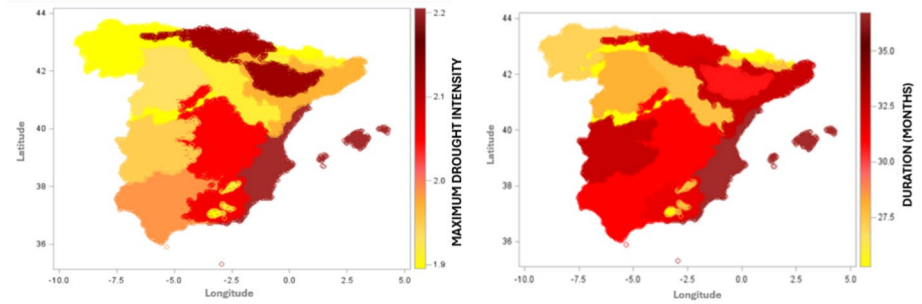


Fig. 9 Average maximum drought intensity and average number of months with drought per decade

precipitation from the historical average, allowing us to identify prolonged droughts and periods of excess moisture.

Negative SPI values (shown in red) signify drought conditions or precipitation deficits. A drought event starts when the SPI falls below -1 and ends when it returns to 0. For each drought event, we calculated three main metrics: duration (in months), intensity (most negative SPI value), and magnitude (area under the SPI curve). Conversely, positive SPI values (shown in blue) indicate wetter-than-average conditions.

Significant droughts were observed in the 1980s, 1990s, and between 2004 and 2006, with increased fluctuations in the last two decades. Drought intensity showed a temporary moderation between 1997 and 2007. After 1980, Cluster 3 (southeastern Spain) experienced more severe droughts, as highlighted in the cumulative SPI trends (Fig. 8). The wettest periods generally occurred from 1960 to 1980, while drought events became more prevalent afterward. All clusters recorded at least one prolonged drought lasting over 20 months, especially during November 2004 to August 2006. Among them, Cluster 3 had the most intense droughts in 1995–1996, while Cluster 12 (eastern Cantabrian coast) suffered its longest drought from January 1989 to November 1990. Cluster 10 (inland Aragón) showed the highest drought intensity in 1996.

Figure 9 displays the average maximum drought intensity and the number of drought-affected months per decade. Cluster 1 (southern Mediterranean), Cluster 10 (inland Aragón), and Cluster 11 (eastern Cantabrian coast) recorded the highest drought intensities. In terms of duration, Cluster 1 had the longest average drought periods, reaching

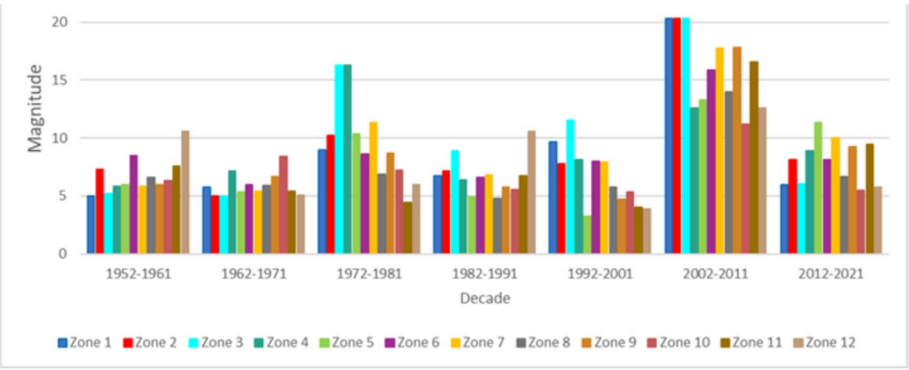


Fig. 10 Average magnitude of drought per decade by region in Spain from 1951 to 2021

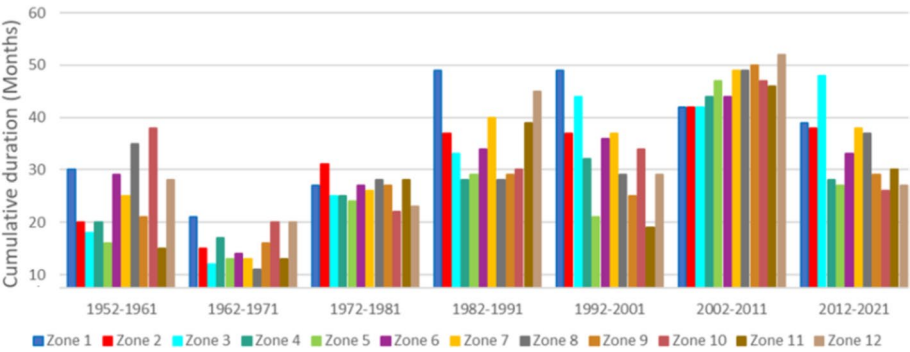


Fig. 11 Average drought duration per decade by region in Spain, 1951–2021

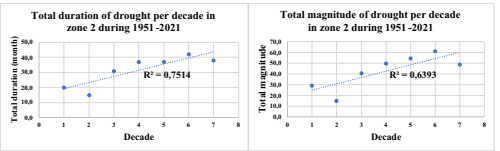
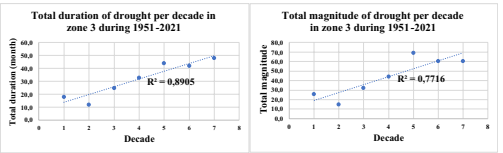
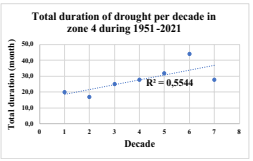
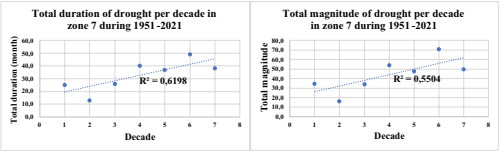
36 months per decade, closely followed by Cluster 7 and Cluster 8. In contrast, mountainous and northwestern regions experimented shorter and less intense droughts.

Figure 10 shows the average drought event magnitude by decade and cluster, while Fig. 11 presents the average duration of events for the same areas. While no significant trend was detected in drought intensity, both average magnitude and duration have increased, particularly in the 2002–2011 decade.

Table 3 presents statistically significant increases in drought event magnitude and total duration per decade, with p-values below 0.05 and R^2 values exceeding 0.5. Clusters 2, 3, and 7 (southern and central-eastern Spain) stood out, especially Cluster 3, which showed an increase of six drought months per decade. Cluster 2 and Cluster 7 exhibited increases of more than four months per decade, while Cluster 4 showed a more moderate increase of three months per decade (equivalent to nine additional drought days per year). In Cluster 4, the total magnitude of drought per decade was not statistically significant; that is why its graphic representation is not included in Table 3.

In comparison with earlier studies, our outcomes reflect similar patterns as Domínguez-Castro et al. (2019), who reported longer and more intense droughts in central and southwestern Spain. However, our results indicate more moderate regional differences. Southeastern Spain (Cluster 3) had the highest drought magnitudes, consistent with González-Hidalgo et al. (2023). In contrast to the overall trend, flash droughts, which

Table 3 Pattern change of the total magnitude and duration of drought events per decade

Cluster	Significance level	Slope	Linear regression plots for total magnitude and total duration versus decade
2	Magnitude p value = 0.048 Duration p value = 0.041	Magnitude 6.53885 Duration 4.82143	
3	Magnitude p value = 0.0104 Duration p value = 0.0016	Magnitude 8.16986 Duration 5.92867	
4	Magnitude p value = 0.0913 Duration p value = 0.0390	Magnitude 4.00547 Duration 3.0714	
7	Magnitude p value = 0.0487 Duration p value = 0.03	Magnitude 6.07135 Duration 4.42857	

Noguera et al. (2021) reported to be widespread throughout Spain, increased notably in Andalusia and Extremadura. These areas also show steep rises in maximum temperatures and changing precipitation patterns, possibly intensifying drought severity through increased evapotranspiration, as suggested by Stagge et al. (2017) and Wang and Yuan (2018).

4 Conclusions

This study introduces a novel method for regionalizing climate variability in Spain, using an innovative distance measure between multivariate climate time series. The method employs dimensionality reduction techniques based on algorithm of Shi (Shi et al. 2021), modified to better align with the multivariate nature of climate data. A key strength of the approach lies in its ability to incorporate multiple temporal scales, enabling the identification of gradual long-term trends and abrupt short-term climatic shifts.

Various clustering algorithms were employed to provide valuable information on the spatiotemporal patterns of climate change in Spain. Twelve different climatic zones were identified, each characterized by similar climatic conditions and trends. The patterns of climatic zones were consistent with the results of previous studies, summarized in Pereira

et al. (2021), but providing a more granular regional analysis, highlighting differential rates of climate change across Spain. Key findings include:

- A significant warming trend was observed across all zones of Spain, with the most pronounced increases occurring during the summer months. The rate of increase in extreme maximum annual temperatures varies by zones, reaching up to 0.49 °C per decade, while the average maximum annual temperatures were rising at a rate up to 0.25 °C per decade.
- A reduction in extreme cold events was detected, mainly in mountainous zones and Atlantic wetlands.
- Minimum temperatures have shown increases ranging from 0.10 °C to 0.22 °C per decade, with the most notable changes occurring in mountainous zones.
- Precipitation has decreased in central and southern zones at a rate lower than 5% per decade. However, intense rainfall events have increased by 2% to 5% per decade across most zones, resulting in fewer rainy days but higher maximum precipitation values.
- The duration and magnitude of droughts have increased by approximately 10% per decade, particularly in Extremadura, Andalusia, and mid-mountain zones.

The findings emphasize the need for region-specific climate adaptation strategies, especially in zones suffering the most pronounced temperature increases and changes in rainfall intensity. The proposed methodology is flexible and can be extended to other climate variables and geographical regions, other observation windows, and other drought indices, offering a powerful tool for environmental studies.

Acknowledgements We express our gratitude to Servicio de Desarrollos Climatológicos of the Meteorological Spanish State Agency for generating and distributing the data we have used in this study, and to the Colombian Ministry of Science and the Technological University of Chocó for supporting the doctoral formation of Arnobio Palacios. The research is also supported by a grant from Agencia Estatal de Investigación (PID2019-106433GB-I00 / AEI / https://doi.org/10.13039/501100011033), Spain.

Author contributions Conceptualization: Arnobio Palacios, Jose Luis Valencia, María Villeta; Methodology: Arnobio Palacios, José Luis Valencia; Data acquisition and resources: José Luis Valencia; Data processing: Arnobio Palacios, Jose Luis Valencia; Validation: Arnobio Palacios, Jose Luis Valencia, María Villeta; Formal analysis and investigation: Arnobio Palacios, Jose Luis Valencia, María Villeta; Writing—original draft preparation: Arnobio Palacios; Writing—review and editing: Jose Luis Valencia, María Villeta; Funding acquisition: María Villeta; Supervision: Jose Luis Valencia, María Villeta.

Funding The authors indicate that there are no relevant financial interests to disclose for this research.

Declarations

Conflict of interest We declare that no financial interests or personal relationships influenced the work reported in this article.

References

Abbass K, Qasim MZ, Song H, Murshed M, Mahmood H, Younis I (2022) A review of the global climate change impacts, adaptation, and sustainable mitigation measures. Environ Sci Pollut Res 29(28):42539–42559. https://doi.org/10.1007/s11356-022-19718-6

Agarwal A, Maheswaran R, Sehgal V, Khosa R, Sivakumar B, Bernhofer C (2016) Hydrologic regionalization using wavelet-based multiscale entropy method. J Hydrol 538:22–32. https://doi.org/10.1016/j.jhydrol.2016.03.023

Baker FB, Hubert LJ (1975). Measuring the power of hierarchical cluster analysis. In: Source: journal of the American statistical association, vol 70, issue 349

Desgraupes B (2017) clusterCrit: Clustering Indices. https://CRAN.R-project.org/package=clusterCrit

Domínguez-Castro F, Vicente-Serrano SM, Tomás-Burguera M, Peña-Gallardo M, Beguería S, El Kenawy A, Luna Y, Morata A (2019) High spatial resolution climatology of drought events for Spain: 1961–2014. Int J Climatol 39(13):5046–5062. https://doi.org/10.1002/joc.6126

Ebi KL, Capon A, Berry P, Broderick C, de Dear R, Havenith G, Honda Y, Kovats RS, Ma W, Malik A, Morris NB, Nybo L, Seneviratne SI, Vanos J, Jay O (2021) Hot weather and heat extremes: health risks. Lancet 398(10301):698–708. https://doi.org/10.1016/S0140-6736(21)01208-3

Fatahi Nafchi R, Yaghoobi P, Reaisi Vanani H, Ostad-Ali-Askari K, Nouri J, Maghsoudlou B (2021) Eco-hydrologic stability zonation of dams and power plants using the combined models of SMCE and CEQUALW2. Appl Water Sci 11(7):109. https://doi.org/10.1007/s13201-021-01427-z

Gebremichael HB, Raba GA, Beketie KT, Feyisa GL, Siyoum T (2022) Changes in daily rainfall and temperature extremes of upper Awash Basin. Ethiop Sci Afr 16:e01173. https://doi.org/10.1016/j.sciaf.2022.e01173

González-Hidalgo JC, Beguería S, Peña-Angulo D, Sandonis L (2021) Variability of maximum and minimum monthly mean air temperatures over mainland Spain and their relationship with low-variability atmospheric patterns for period 1916–2015. Int J Climatol 42(3):1723–1741. https://doi.org/10.1002/joc.7331

Gonzalez-Hidalgo JC, Beguería S, Peña-Angulo D, Trullenque-Blanco V (2023) MOPREDAS_century database and precipitation trends in mainland Spain, 1916–2020. Int J Climatol 43(8):3828–3840. https://doi.org/10.1002/joc.8060

Gonzalez-Hidalgo JC, Trullenque-Blanco V, Beguería S, Peña-Angulo D (2024) Seasonal precipitation changes in the western Mediterranean Basin: The case of the Spanish mainland, 1916–2015. Int J Climatol 44(5):1800–1815. https://doi.org/10.1002/joc.8412

Hamed KH (2009) Enhancing the effectiveness of prewhitening in trend analysis of hydrologic data. J Hydrol 368(1–4):143–155. https://doi.org/10.1016/j.jhydrol.2009.01.040

Hubert L, Schultz J (1976) Quadratic assignment as a general data analysis strategy. Br J Math Stat Psychol 29(2):190–241. https://doi.org/10.1111/j.2044-8317.1976.tb00714.x

IPCC: Masson-Delmotte V, Zhai P, Chen Y, Goldfarb L, Gomis MI, Matthews JBR, Berger S, Huang M, Yelekçi O, Yu R, Zhou B, Lonnoy E, Maycock TK, Waterfield T, Leitzell K, Caud N (2021) Working Group I contribution to the sixth assessment report of the intergovernmental panel on climate change edited by. Climate Change 2021: The Physical Science Basis. www.ipcc.ch

Jaagus J (2006) Climatic changes in Estonia during the second half of the 20th century in relationship with changes in large-scale atmospheric circulation. Theoret Appl Climatol 83(1–4):77–88. https://doi.org/10.1007/s00704-005-0161-0

Javadinejad S, Hannah D, Ostad-Ali-Askari K, Krause S, Zalewski M, Boogaard F (2019a) The impact of future climate change and human activities on hydro-climatological drought, analysis and projections: using CMIP5 climate model simulations. Water Conserv Sci Eng 4(2–3):71–88. https://doi.org/10.1007/s41101-019-00069-2

Javadinejad S, Ostad-Ali-Askari K, Eslamian S (2019b) Application of multi-index decision analysis to management scenarios considering climate change prediction in the Zayandeh Rud River Basin. Water Conserv Sci Eng 4(1):53–70. https://doi.org/10.1007/s41101-019-00068-3

Kaufman L, Rousseeuw PJ (1990). Finding groups in data: an introduction to cluster analysis. In: Kaufman L, Rousseeuw PJ (eds), John Wiley & Sons, Inc.

Kendall MG (1975) Rank correlation methods, 4th edn

Kohonen T, Oja E (1996) Engineering applications of the self-organizing map. Proc IEEE. https://doi.org/10.1109/5.537105

Kuriqi A, Ali R, Pham QB, Montenegro Gambini J, Gupta V, Malik A, Linh NTT, Joshi Y, Anh DT, Nam VT, Dong X (2020) Seasonality shift and streamflow flow variability trends in central India. Acta Geophys 68(5):1461–1475. https://doi.org/10.1007/s11600-020-00475-4

Li J, He X, Tao L (2022) Assessing multiscale variability and teleconnections of monthly precipitation in Yangtze River Basin based on multiscale information theory method. Theoret Appl Climatol 147(1–2):717–735. https://doi.org/10.1007/s00704-021-03845-0

Lorenzo MN, Alvarez I (2022) Future changes of hot extremes in Spain: towards warmer conditions. Nat Hazards 113(1):383–402. https://doi.org/10.1007/s11069-022-05306-x

Mann HB (1945) Nonparametric tests against trend. Econometrica 13(3):245. https://doi.org/10.2307/1907187

Mcclain JO, Rao VR (1975) CLUSTISZ: a program to test for the quality of clustering of a set of objects. In: Source: journal of marketing research, vol 12, issue 4

Mckee TB, Doesken NJ, Kleist J (1993) The relationship of drought frequency and duration to time scales. In: Eighth conference on applied climatology
Meseguer-Ruiz O, Lopez-Bustins JA, Arbiol-Roca L, Martin-Vide J, Miró J, Estrela MJ (2021) Temporal changes in extreme precipitation and exposure of tourism in Eastern and South-Eastern Spain. Theoret Appl Climatol 144(1–2):379–390. https://doi.org/10.1007/s00704-021-03548-6
Młyński D, Wałęga A, Kuriqi A (2021) Influence of meteorological drought on environmental flows in mountain catchments. Ecol Ind 133:108460. https://doi.org/10.1016/j.ecolind.2021.108460
Murtagh F, Legendre P (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? J Classif 31(3):274–295. https://doi.org/10.1007/s00357-014-9161-z
Newell RG, Prest BC, Sexton SE (2021) The GDP-temperature relationship: Implications for climate change damages. J Environ Econ Manag 108:102445. https://doi.org/10.1016/j.jeem.2021.102445
Noguera I, Domínguez-Castro F, Vicente-Serrano SM (2021) Flash DROUGHT RESPONSE TO PRECIPITATION AND ATMOSPHERIC EVAPORATIVE DEMAND in Spain. Atmosphere 12(2):165. https://doi.org/10.3390/atmos12020165
Ostad-Ali-Askari K (2022) Review of the effects of the anthropogenic on the wetland environment. Appl Water Sci 12(12):260. https://doi.org/10.1007/s13201-022-01767-4
Parracho AC, Melo-Gonçalves P, Rocha A (2016) Regionalisation of precipitation for the Iberian Peninsula and climate change. Phys Chem Earth Parts a/b/c 94:146–154. https://doi.org/10.1016/J.PCE.2015.07.004
Peña-Angulo D, Gonzalez-Hidalgo JC, Sandonís L, Beguería S, Tomas-Burguera M, López-Bustins JA, Lemus-Canovas M, Martin-Vide J (2021) Seasonal temperature trends on the Spanish mainland: a secular study (1916–2015). Int J Climatol 41(5):3071–3084. https://doi.org/10.1002/joc.7006
Pereira SC, Carvalho D, Rocha A (2021) Temperature and precipitation extremes over the iberian peninsula under climate change scenarios: a review. Climate 9(9):139. https://doi.org/10.3390/cli9090139
Ray S, Turi RH (1999) Determination of number of clusters in K-means clustering and application in colour image segmentation. In The 4th international conference on advances in pattern recognition and digital techniques, pp 137–143
Rodrigo FS (2023) Spatiotemporal variability of the relationship between seasonal temperatures and precipitation in Spain, 1951–2019. Theor Appl Climatol 153(3–4):1371–1391. https://doi.org/10.1007/s00704-023-04550-w
Roushangar K, Alizadeh F (2018) A multiscale spatio-temporal framework to regionalize annual precipitation using k-means and self-organizing map technique. J Mt Sci 15(7):1481–1497. https://doi.org/10.1007/s11629-017-4684-5
Sen PK (1968) Estimates of the Regression Coefficient Based on Kendall's Tau. J Am Stat Assoc 63(324):1379–1389. https://doi.org/10.1080/01621459.1968.10480934
Senent-Aparicio J, López-Ballesteros A, Jimeno-Sáez P, Pérez-Sánchez J (2023) Recent precipitation trends in Peninsular Spain and implications for water infrastructure design. J Hydrol Reg Stud 45:101308. https://doi.org/10.1016/j.ejrh.2022.101308
Shi Y, Li B, Du G, Dai W (2021) Clustering framework based on multi-scale analysis of intraday financial time series. Phys A Stat Mech Appl 567:71932008. https://doi.org/10.1016/j.physa.2020.125728
Stagge JH, Kingston DG, Tallaksen LM, Hannah DM (2017) Observed drought indices show increasing divergence across Europe. Sci Rep 7(1):14045. https://doi.org/10.1038/s41598-017-14283-2
Tessier Y, Lovejoy S, Hubert P, Schertzer D, Pecknold S (1996) Multifractal analysis and modeling of rainfall and river flows and scaling, causal transfer functions. J Geophys Res Atmos 101(D21):26427–26440. https://doi.org/10.1029/96JD01799
United Nations Convention to Combat Desertification (2022) Drought in Numbers 2022 - restoration for readiness and resilience. https://www.unccd.int/sites/default/files/2022-06/Drought%20in%20Numbers%20%28English%29.pdf
Wang L, Chen S, Zhu W, Ren H, Zhang L, Zhu L (2021) Spatiotemporal variations of extreme precipitation and its potential driving factors in China's North-South Transition Zone during 1960–2017. Atmos Res 252:105429. https://doi.org/10.1016/j.atmosres.2020.105429
Wang L, Yuan X (2018) Two Types of flash drought and their connections with seasonal drought. Adv Atmos Sci 35(12):1478–1490. https://doi.org/10.1007/s00376-018-8047-0
Wang W, Chen Y, Becker S, Liu B (2015) Variance correction prewhitening method for trend detection in autocorrelated data. J Hydrol Eng. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001234
Zhang X, Alexander L, Hegerl GC, Jones P, Tank AK, Peterson TC, Trewin B, Zwiers FW (2011) Indices for monitoring changes in extremes based on daily temperature and precipitation data. In Wiley

interdisciplinary reviews: climate change, vol 2, Issue 6, pp 851–870. Wiley-Blackwell. https://doi.org/10.1002/wcc.147

Capítulo 5

Capítulo 5

Uso recursivo de variables no climáticas en el proceso de regionalización climática para optimizar las agrupaciones finales

Resumen: *Por lo general, la regionalización climática se realiza agrupando únicamente series temporales de información climática. En este estudio se presenta un procedimiento de regionalización que no solo considera variables climáticas para obtener las agrupaciones finales.*

En el presente capítulo, el procedimiento de regionalización climática propuesto, además de incorporar indicadores cuantificables del cambio climático como input en el proceso de agrupación, emplea PLS-DA para introducir variables geográficas de forma recursivas en el proceso de agrupamiento. Esto no solo permite cuantificar su impacto en el cambio climático sino también optimizar los resultados de la agrupación mediante la creación de un índice que equilibra las principales métricas del modelo.

5.1. Objetivos

Emplear indicadores cuantificables del cambio climático y el impacto de variables geográficas sobre el clima y sus variaciones para identificar regiones climáticamente diferenciadas en España entre 1951 y 2021.

Evaluar la influencia de factores geográficos sobre patrones del cambio climático en España entre 1951 y 2021.

Emplear de manera recursiva la introducción de variables no climáticas en los procesos de agrupación de series temporales climáticas para optimizar los resultados finales del agrupamiento y por consiguiente de la regionalización climática.

5.2. Metodología

En primera instancia, este estudio genera series temporales multivariantes a partir de las series temporales de temperaturas extremas y de precipitaciones, luego emplea modelos autorregresivos para estimar los coeficientes asociados a las medias, tendencias y autocorrelaciones de las diferentes variables climáticas generadas.

Tras obtener las diferentes estimaciones, se utiliza una nueva distancia definida que tiene en cuenta pesos asignados a las estimaciones de acuerdo con el tipo de variable, temperaturas o precipitaciones, y en base a tres combinaciones de pesos genera tres matrices de similitudes. A continuación, se tienen en cuenta tres de los algoritmos de agrupamiento definidos en el capítulo 2 para generar agrupaciones, y teniendo en cuenta dos de los índices de validación definidos en la sección 2.5 del capítulo 2, selecciona 9 configuraciones de clustering en una primera optimización, 3 por algoritmo de clustering y 3 por matriz de similitudes.

A continuación, se emplea PLS-DA para generar 9 modelos de discriminación, empleando las variables geográficas consideradas para cada zona como variables explicativas X y las agrupaciones finales de cada configuración como variables predichas Y, permitiendo cuantificar el impacto de las variables X sobre los patrones de cambio climático representados por Y. Finalmente, se crea un índice que tienen en cuenta las principales métricas del modelo para seleccionar la configuración final de clústeres, se definen los prototipos de cada clúster y se analizan teniendo en cuenta herramientas de análisis climáticos.

5.3. Resumen de resultados

Tras considerar 15,992 series temporales distribuidas en la España Peninsular entre 1951 y 2021, los resultados del agrupamiento permitieron regionalizar a España en siete zonas climáticas distintas, permitiendo evidenciar diferencias regionales significativas en los patrones de cambio climático en toda la región de estudio.

En general, se encontró que las temperaturas máximas aumentaron más en las regiones montañosas y en las zonas centrales de España, mientras que las regiones meridionales experimentaron incrementos menores. De igual forma, se notó un

descenso de las precipitaciones, sobre todo en las regiones del sur y del interior, así como un marcado aumento de los días secos consecutivos, especialmente en las zonas del sur de España, y una reducción de las temperaturas extremas frías en la mayoría de las regiones.

5.4. Conclusiones

La integración de indicadores cuantificables del cambio climático y de variables geográficas en el proceso de agrupación permite un análisis exhaustivo de la variabilidad climática regional. El procedimiento de agrupación de series temporales propuesto permitió vincular eficazmente las tendencias climáticas espaciales y temporales a las variables geográficas, lo que conduce a clasificaciones regionales más refinadas.

5.5. Publicación JCR

De acuerdo con la estructura de esta tesis, como cuerpo de este capítulo se incluye la publicación del presente estudio, identificable en las bases de datos científicas de acuerdo con la siguiente referencia.

Estado: Publicado

Año de Publicación: 2025

Nombre de la Revista: Earth Systems and Environment

Editorial: Springer

Indicadores de Calidad: Factor de Impacto año 2023: 5.3; Rango: Q1, 26/254 de la categoría *GEOSCIENCES, MULTIDISCIPLINARY*; CiteScore: 15.5 (Scopus), SNIP 2.018

DOI: https://doi.org/10.1007/s11069-024-07082-2

Earth Systems and Environment
https://doi.org/10.1007/s41748-025-00568-4

ORIGINAL ARTICLE

Quantifying Impact of Geographical Variables on Climate Change Patterns over Spain by Time Series Clustering

Arnobio Palacios-Gutiérrez[1,2] · **Jose Luis Valencia-Delfa**[1] · **María Villeta**[1]

Received: 6 June 2024 / Revised: 2 December 2024 / Accepted: 1 January 2025

Abstract
One of the greatest environmental threats worldwide arises from temperature and precipitation variations driven by climate change. Recent studies have increasingly focused on climatic regionalization, generating clusters to analysing climate change patterns. However, most of these studies have analysed the effect of climate change on the groups once they have been formed. In this context, the present study proposes a novel regionalization approach by including climatic change estimates into each meteorological time series as input for clusters generation. This innovative methodology allows us quantify the impact of geographical variables, such as distance to the sea, height, latitude and longitude, on climate change patterns within a territory. Partial Least Squares Discriminant Analysis (PLS-DA) was employed to qualitatively describe the clusters and assess the influence of geographical variables. Additionally, from the PLS-DA analysis, a new index that optimizes clustering by balancing the main metrics from the model was generated. This approach was applied to variations in temperature and precipitation of Peninsular Spain from 1951 to 2021, analysing 15,992 multivariate time series. The results reveal significant regional differences in the observed climate change patterns. Maximum temperatures have increased the most in mountainous regions and central areas of Spain, while the smallest increases occurred in Southern Spain. A decrease in precipitation was observed, with most pronounced reductions in southern and inland regions. Furthermore, there was a marked increase in consecutive dry days, particularly in the South. Trends in cold temperature extremes have diminished across most regions. These findings provide valuable information for future climate adaptation and mitigation strategies. The proposed methodology is flexible and scalable, making it suitable for application to large regions with high climatic variability.

Keywords Climate Change · Clustering · Geographical Effects · Multivariate Time Series · PLS-DA · Spain

1 Introduction

Climate variations, manifested through changes in temperatures and precipitation, represent one of the principal global environmental threats (Gupta et al. 2019). These changes have a profound effect in the society and ecosystems by altering extreme meteorological and hydrological events (Kuriqi et al. 2020; Młyński et al., 2021; Malede et al. 2022). Such impacts are evident in critical issues such as drought and climate change, which affect food security, human health, and other key aspects (Ciais et al. 2005; Patz et al. 2005; de Lucena et al. 2009; O'neill and Ebi 2009; Choularton et al. 2012; Schaeffer et al. 2012; Sathaye et al. 2013; Brown et al. 2015; Khan et al. 2021). These climate issues and their impacts can vary considerably at a regional scale, as atmospheric conditions in different geographies, and their variations, are influenced by physical factors such

✉ Arnobio Palacios-Gutiérrez
arnobiop@ucm.es

Jose Luis Valencia-Delfa
joseval@ucm.es

María Villeta
mvilleta@ucm.es

[1] Faculty of Statistical Studies, Complutense University of Madrid, Avenida Puerta de Hierro, nº 1, 28040 Madrid, Spain

[2] Group Valoración y Aprovechamiento de la Biodiversidad, Technological University of Chocó, Quibdó, Chocó, Colombia

Published online: 07 February 2025

as topography, vegetation distribution, urbanization, water bodies, and other characteristics that affect surface climate (Chen et al. 2006; Teodoro et al. 2021). Consequently, as suggested by Yue & Hashino (2003) and Palacios Gutiérrez et al. (2023), climatic studies may be more effective if differences or similarities between sub-regions within a territory are identified. To detect significant climatic changes across different regional zones, it is crucial to identify spatio-temporal patterns in climate variations (Laepple and Huybers 2014).

In response to these regional climate variations, regionalization emerges as a crucial tool for identifying homogeneous zones within a large area by grouping localities based on shared climatic variables. This approach enables more precise analysis and a deeper understanding of spatial and temporal patterns, facilitating a more detailed study of climatic changes at the regional level.

In recent years, various studies have addressed climatic regionalization and its relevance for analysing climate change at different scales. Parracho et al. (2016) proposed a regionalization approach based on daily precipitation in the Iberian Peninsula using the k-means method, successfully identifying areas with homogeneous precipitation patterns. Similarly, Samantaray et al. (2021) employed a Markov random field model to investigate climate change impacts on hydroclimatic patterns in India, identifying homogeneous regions from monthly and monsoon precipitation data. Another notable approach was by Saunders et al. (2021), who used extreme precipitation dependence to group climatic stations in Australia, applying a hierarchical technique based on the F-madogram distance. Although these studies demonstrate the utility of regionalization for understanding climatic variations, they focus primarily on specific climatic data without deeply exploring the influence of geographic factors or climate change trends.

In contrast, Imam et al. (2022) present a different approach by basing their regionalization on broader and traditional geographic divisions, such as natural regions or cardinal divisions. While this method has the advantage of being easy to apply and providing a general overview of geographic differences, it lacks the precision needed to identify specific climatic patterns within regions. This general approach does not capture local particularities of climatic variations, limiting its applicability in regional climate analysis.

On the other hand, Palacios Gutiérrez et al. (2023) introduced an innovative method for clustering maximum temperature time series using trend, seasonality, and noise components extracted through a sequential singular spectrum analysis decomposition. This approach divided the Iberian Peninsula into three climatic zones, identifying temperature change patterns associated with the region's climatic gradient. Although this study focuses specifically on temperature behaviour, its methodology exemplifies how advanced time series analysis techniques can be applied to enhance understanding of regional climate patterns. Moreover, global approaches, such as those by Shi et al. (2016), provide a broad view of climate trends and fluctuations.

In summary, the existing literature on climatic regionalization explores at various scales to generate clusters based on climatic variables and subsequently examines the impact of climate change on these clusters. However, most studies focus on analysing the groups after their formation, without integrating an estimate of climate change that could influence the clustering process.

In this context, our study introduces a novel approach that extends beyond traditional regionalization. This clustering process identifies how specific geographical factors —such as Distance to the sea, Height, Latitude and Longitude— affect climate levels and their changes within a territory. For regionalization and discrimination according to climate levels and its changes, the clustering process includes an estimate of climate change in each time series, which is used as an input for generating clusters. This method aligns with the approaches used in previous studies (Shi et al. 2016; Roushangar and Alizadeh 2018; Palacios Gutiérrez et al. 2023), which also deviate from traditional regionalisation methods. By including climate trends and autocorrelations in the clustering features —rather than just considering mean values—, it has been found improvements in model performance, enhancing the regionalisation process. This approach is particularly valuable for regions seeking to develop adaptive strategies to climate change.

The proposed methodology is applicable to any geographic area, and it is especially relevant for regions where climatic changes exhibit significant variations between neighbouring zones, as observed in Spain (Furió and Meneu 2011; del Río et al. 2012; Ramos et al. 2012; Peña-Angulo et al. 2015; Fonseca et al. 2016; García 2022; Palacios Gutiérrez et al. 2023). By integrating both climate change estimation and the influence of geographic variables into the regionalization process, this approach provides a more detailed understanding of climate change patterns, overcoming the limitations of previous approaches.

This study analyses three variables to discriminate climate changes in Spain using geographical information: maximum (T_{max}) and minimum (T_{min}) temperatures, along with precipitation ($Prcp$). For each location, nine annual time series were generated, comprising three series for each climatic variable. Autoregressive models were used to estimate three parameters per series, resulting in a total of 27 parameters per location (9 for T_{max}, 9 for T_{min} and 9 for $Prcp$). These parameters were used to define the distance between two zones. The calculated distance incorporates a

weighting scheme to adjust the relative importance of precipitation and temperature parameters. Three weight combinations were tested, resulting in three distinct distance matrices.

Three unsupervised learning algorithms —Hierarchical Agglomerative (HA) (Murtagh and Legendre 2014), K-medoids (PAM) (Kaufman and Rousseeuw 1990), and Kohonen's Self-Organizing Maps (SOM) (Kohonen & Oja 1996)— were applied to each distance matrix, yielding nine clustering configurations. The optimal number of clusters for each configuration was determined using the Dunn and Xie-Beni internal validation indices. A Partial Least Squares Discriminant Analysis (PLS-DA) model was then adjusted for each clustering configuration to evaluate the discriminative power of geographical variables —Distance to the sea, Height, Latitude and Longitude— in explaining the resulting clusters. The configuration that maximizes the combination of metrics derived from PLS-DA through a newly developed index was selected.

The remainder of this document is structured as follows: Sect. 2 describes the study area, data sources, models applied, the distance metrics and weighting schemes, and the clustering methods used. Section 3 presents the main results and discussion, including the optimization of cluster numbers, the selection of the optimal configuration, and the analysis of spatio-temporal characteristics. Finally, Sect. 4 outlines the study's key conclusions.

2 Data and Method

2.1 Area of Study

The study focuses on the Spanish peninsular territory. It is situated in southwestern Europe, bordered by the Atlantic Ocean to the west, the Mediterranean Sea to the east, and neighboring Portugal and France. The warm temperatures of the Mediterranean Sea typically influence southern and eastern Spain, while the northern and western regions tend to be influenced by the low temperatures of the Atlantic Ocean. Due to its geographical positioning, Spain exhibits a pronounced climatic gradient running from north to south, alongside notable diurnal and seasonal thermal gradients from coastal regions and the inland areas (Dasari et al. 2014).

2.2 Observed Data

This study utilizes a dataset comprising 15,992 multivariate time series derived from daily records of maximum and minimum temperatures, as well as precipitation. These data were obtained from the "Servicio de Desarrollos Climatológicos" of the Spanish Meteorological Agency (AEMET). The data are accessible via the website:

https://www.aemet.es/es/serviciosclimaticos/cambio_climat/datos_diarios?w=2&w2=0. The initial data was processed using spatial interpolation via kriging, applied to a 5×5 km^2 grid covering the entire territory of Peninsular Spain. Specifically, for each grid point nine annual climatic variables were computed, including:

- Total annual precipitation (PRCP).
- Maximum daily precipitation ($PRCP_{max}$).
- Number of dry days (precipitation$\leqslant$ 0.25 mm.) (DRY)
- Mean daily maximum temperature (TMAX).
- Maximum of daily maximum temperatures ($TMAX_{max}$).
- Minimum of daily maximum temperatures ($TMAX_{min}$).
- Mean daily minimum temperature (TMIN).
- Maximum of daily minimum temperatures ($TMIN_{max}$).
- Minimum of daily minimum temperatures ($TMIN_{min}$).

The resulting datasets include 15,992 multivariate time series, each with nine variables recorded over 71 years (1951–2021), with one observation per year for each variable. The analysis is restricted to grid cells located within Peninsular Spain (see Fig. 1). Additionally, geographic variables —Longitude, Latitude, Height, and Distance to the Sea— were included for each grid point. These variables were incorporated into the clustering and discrimination process of the time series.

2.3 Methodology

The primary objective of this study is to examine the influence of geographic location on climate change patterns in Spain. To achieve this, a clustering and discrimination methodology for climate time series was developed, integrating geographic information. The detailed steps of the proposed methodology are outlined below.

2.3.1 Extraction of Characteristics from Time Series of Climate Data

In this study, each grid point, represented by a multivariate time series, is treated as an object within a feature-based clustering framework. The attributes extracted from these time series capture their essential characteristics, enabling clustering algorithm to distinguish similarities and differences among them (Wang and Hyndman 2006).

To extract the relevant features, an autoregressive model, expressed in Eq. (1), was applied to each univariate time series (Gallant and Goebel 1976). To incorporate the influence of climate change on the time series, the model includes a coefficient $p_{ij,2}$ that evaluates the trend. Additionally,

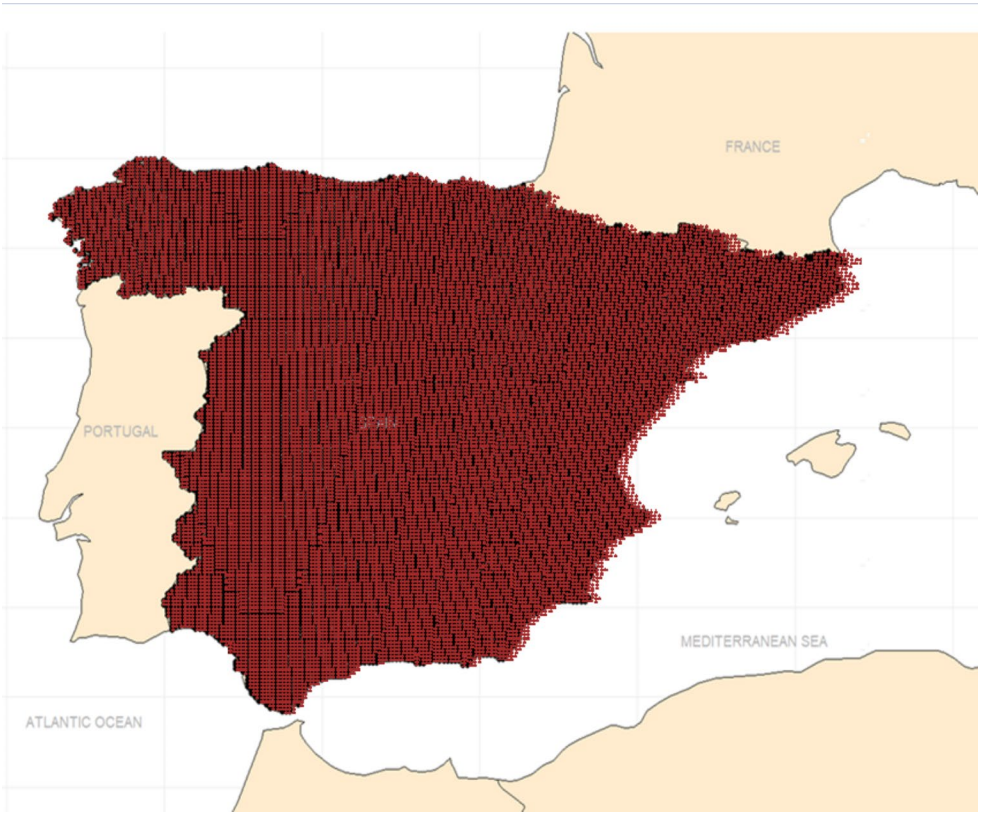


Fig. 1 Spatial distribution of the grid cells in Spain

the coefficient $p_{ij,3}$ captures first-order autocorrelation, accounting for temporal dependencies within the series.

$$y_{ij}(t) = p_{ij,1} + p_{ij,2}t + p_{ij,3}y_{ij}(t-1) \tag{1}$$

where $y_{ij}(t)$ represents the climate time series for zone $i = 1,2,\ldots,15,992$ and variable $j = 1\ldots 9$.

After estimating the parameters of the autoregressive model for each of the nine univariate climate time series, a feature vector consisting of 27 elements is constructed for each grid point.

$$\begin{aligned} C_i = &[p_{i1,1}, p_{i1,2}, p_{i1,3}, p_{i2,1}, p_{i2,2}, p_{i2,3}, \\ &\ldots, p_{i9,1}, p_{i9,2}, p_{i9,3}] \end{aligned} \tag{2}$$

This vector C_i is referred to as the feature vector or representation of grid point i.

2.3.2 Measurement of the Similarity between Different Grid Cells

The feature vector C_i was subdivided into two sub-vectors based on the type of climatic information represented by the parameters. Specifically, the first 18 parameters correspond to temperature variables, forming the sub-vector $C_{i,TE}$ while the remaining nine parameters are associated with precipitation, forming the sub-vector $C_{i,Prcp}$. This subdivision ensures that the clustering process accounts for the distinct contributions of temperature and precipitation in defining climatic patterns across the grid cells.

2.3.2.1 Weightings of Characteristics for Assessing Similarity between Grid Cells.

To define a distance that measures the similarity between grid cells based on their climate parameters, we have assigned different weights to these parameters to optimize the discrimination process.

Given the study's goal of aims to determine the influence of geographical factors on climate changes within a territory, it is important to minimize uncertainty in the proposed approach or improve its optimization in terms of explanatory capacity. Previous studies (Weigel et al. 2008; Knutti 2010; Kolusu et al. 2021; Wootten et al. 2023) have demonstrated that weighting strategies in climate models or projections often yield greater accuracy than unweighted multi-model means in several cases.

Weighting approaches are typically adjusted based on the variables, domains, data types (raw or downscaled), and specific weighting schemes under consideration (Wootten et al. 2023). For this study the weighting of parameters was strategically designed to reflect the type of climate variable being analysed. Shin et al. (2020) emphasized the need for multiple weighting schemes to address uncertainties in climate change projections. Therefore, we decided to

implement three weighting combinations based on the type of climate variable.

Furthermore, regional characteristics can influence the weights assigned (Brunner et al. 2020), making specific weighting schemes valuable for addressing biases and interdependencies (Wootten et al. 2023). In Peninsular Spain, the diverse climatic variations across regions make it reasonable to adjust the weights assigned to temperature and precipitation parameters based on the region's sensitivity to these variables. The three types of weightings considered were as follows.

Weighting Scheme M1 This scheme accounts for the high correlation between maximum and minimum temperatures T_{max} and T_{min}, which may reduce the independent contribution of temperature parameters. Here a weight of 60% is assigned to temperature parameters and 40% to precipitation parameters. As a result, the weighting for parameters in $C_{i,TE}$ are set at 1/30, and for $C_{i,PRCP}$ at 2/45. This scheme is suitable for regions where temperature heterogeneity dominates over precipitation variability.

Weighting Scheme M2 In this scheme, greater importance (60%) is given to precipitation parameters, reflecting regions with significant rainfall variability, while 40% is assigned to temperature parameters. Therefore, the parameters $C_{i,TE}$ in are weighted at 1/15, and those $C_{i,PRCP}$ for at 1/45. This scheme is ideal for areas where precipitation variations are more critical for defining climatic differences.

Weighting Scheme M3 In this case, an equal weight is assigned to all types of parameters, assuming no dominant relationship between T_{max} and T_{min}, in certain regions. Both $C_{i,TE}$ and $C_{i,PRCP}$parameters are weighted equally at 1/27. This scheme assumes all parameters with equal importance. It is suited for regions with balanced contributions of temperature and precipitation variability.

The flexibility of these weighting schemes allows them to be adapted to the specific characteristics of the regions being analysed. In Peninsular Spain, certain regions are more influenced by temperature, while others are affected by precipitation. These weighting strategies enhance the discriminatory power of the clustering process and improve its explanatory capacity by capturing regional climatic differences.

2.3.2.2 Distance between Grid Cells. In this study, the observations are zones represented by feature vectors derived from time series data. These observations pertain to characteristic values of states and changes, such as means, trends, and autocorrelations. To quantify the overall similarity between two grid cells, a distance metric has been defined to identify groups with general profiles based on their magnitudes. Initially, the Manhattan distance (Eq. (3)) is employed which, like the Euclidean distance, is magnitude-based and particularly suitable for directly assessing differences across various attributes.

$$d\left(zone_i, zone_h\right) = \sqrt{\sum_{j=1}^{6}\sum_{r=1}^{3} w_1\left(|p_{i,j,r} - p_{h,j,r}|\right) + \sum_{j=7}^{9}\sum_{r=1}^{3} w_2\left(|p_{i,j,r} - p_{h,j,r}|\right)} \tag{3}$$

where $d\left(zone_i\ ,\ zone_h\right)$ is the final distance from zone i to zone h and w_1 is the weight of parameters related to extreme temperatures, w_2 is the weight of parameters related to precipitation and r is the type of parameter (intercept, trend and autoregressive coefficient).

Assigning on Eq. (3) the combinations of weights defined in 2.3.2.1 section, the following distances were used to measure the similarity between two zones:

$$d_{M1}\left(zone_i, zone_h\right) = \sqrt{\sum_{j=1}^{6}\sum_{r=1}^{3} \frac{1}{30}\left(|p_{i,j,r} - p_{h,j,r}|\right) + \sum_{j=7}^{9}\sum_{r=1}^{3} \frac{2}{45}\left(|p_{i,j,r} - p_{h,j,r}|\right)} \tag{3.1}$$

$$d_{M2}\left(zone_i, zone_h\right) = \sqrt{\sum_{j=1}^{6}\sum_{r=1}^{3} \frac{1}{45}\left(|p_{i,j,r} - p_{h,j,r}|\right) + \sum_{j=7}^{9}\sum_{r=1}^{3} \frac{1}{15}\left(|p_{i,j,r} - p_{h,j,r}|\right)} \tag{3.2}$$

$$d_{M3}\left(zone_i\ , zone_h\right) = \sqrt{\sum_{j=1}^{6}\sum_{r=1}^{3} \frac{1}{27}\left(|p_{i,j,r} - p_{h,j,r}|\right) + \sum_{j=7}^{9}\sum_{r=1}^{3} \frac{1}{27}\left(|p_{i,j,r} - p_{h,j,r}|\right)} \tag{3.3}$$

Considering that the parameters used in these distances collect characteristics with different magnitudes, it is recommended to standardize them to ensure that all are within the same range, avoiding any particular parameter from disproportionately influencing the final dissimilarity measure. To address this potential issue, the parameters were standardized to have mean 0 and standard deviation of 1 (Bro and Smilde 2003).

2.3.3 Time Series Clustering Configurations

Clustering algorithms can generally be classified into six primary categories: partitioning, hierarchical, grid-based, model-based, density-based, and multi-step clustering (Aghabozorgi et al. 2015). In this study, three distinct algorithms representing these categories were employed: Hierarchical Agglomerative Clustering (HA), K-medoids (PAM), and Self-Organizing Maps (SOM).

HA is a hierarchical method that operates in a two-step process (Murtagh and Legendre 2014). K-medoids, a partitioning algorithm closely related to K-means, optimizes the selection of the medoid to minimize within-cluster dissimilarity (Hartigan and Wong 1979). SOM, a model-based algorithm, employs a neural network structure to cluster time series based on topological similarities (Kohonen and Oja 1996). The variant employed in this study was adapted to handle dissimilarity matrices.

Three different dissimilarity matrices (M1, M2 and M3) were derived from the data, resulting in nine clustering configurations: three for each algorithm (HA-M1, HA-M2, HA-M3, PAM-M1, PAM-M2, PAM-M3, SOM-M1, SOM-M2 and SOM-M3). For instance, HA-M1 refers to the HA algorithm applied with the M1 distance matrix, and similar interpretations applying for the other configurations.

The HA algorithm was implemented using the hclust function in R with Ward's method (Murtagh and Legendre 2014). For the PAM algorithm, the pam function from the R $cluster$ package was used (Kaufman and Rousseeuw 1990). For the SOM algorithm, the $\mathrm{trainSOM}$ function from the $SOMbrero$ package was employed (Olteanu and Villa-Vialaneix 2015), using an 8×8 grid for the topological configuration, which ensured effective performance across all matrices.

To determine the optimal number of clusters and assess the clustering performance, two internal validation metrics focused on compactness and separation were used (Zhao and Karypis 2002; Kraus et al. 2011). These metrics evaluate the clustering quality solely based on the dataset, without external information (Brun et al. 2007; Kremer et al. 2011; Song and Zhang 2008). The two indices used were Dunn's Index, which should be maximized to promote tighter clusters with larger inter-cluster distances (Dunn 1974), and the Xie-Beni Index, which should be minimized (Xie and Beni 1991). Originally designed for fuzzy clustering, the Xie-Beni Index is also applicable to crisp clustering, as described by Desgraupes (2017).

Since the clustering algorithms require specifying the number of clusters (k) a priori, we systematically varied k from 3 to 16 for each configuration. The optimal number of clusters for each configuration was selected based on the validation indices, resulting in nine distinct clustering results, each with a specific k value.

2.3.4 Discrimination of the Clusters Obtained Based on the Geographic Location

After optimizing the nine cluster configurations and determining the appropriate number of clusters k for each, the next step involves labelling the clusters. The primary objective is to use these labeled clusters or groups of time series to construct a model with suitable parameters. This model aims to optimize the direction of maximum discrimination for the time series within each group, based on the geographical information variables, for each configuration.

For this purpose, the Partial Least Squares Discriminant Analysis (PLS-DA) model was employed a robust supervised multivariate analysis method (Varmuza and Filzmoser 2009). The PLS-DA classification technique merges Partial Least Squares Regression (PLS) with Linear Discriminant Analysis (LDA) (Sjöström et al. 1986). This technique establishes a PLS model using a matrix of explanatory variables X (geographical information variables, with a dimension of $15{,}992\ \text{zones} \times 4\ \text{geographical variables}$) and a response matrix Y with a dimension of ($15{,}992\ \text{zones} \times \text{number of clusters}$) which represents the cluster or groups of time series. Since Y is a categorical, it is recoded into dummy variables (0,1) one for each cluster creates. Consequently, the PLS regression is executed as if Y were a continuous matrix. Thus, for the discrimination scenario, the PLS model can be represented as:

$$Y_{15992\times\, n^{\circ}\, clusters} = X_{15992\times\, 4zones}\beta + E_{15992\times\, n^{\circ}\, clusters} \tag{4}$$

where β is a regression coefficient matrix ($4 \times n^{\circ}\, clusters$) and E is a residual matrix.

This PLS model establishes a relationship between the variations of the observations (i.e., the time series) and the group to which they belong to by first transforming the original explanatory variables into a set of latent variables or factors. These latent variables are then used for regression with the response variable (Varmuza and Filzmoser 2009). More specifically, $\beta = W^{*}V^{T}$, where V is the matrix containing the right singular vectors of the SVD decomposition (loading vectors) in column form, and $W^{*} = W\left(U^{T}W\right)^{-1}$, where W denotes the matrix containing the regression coefficients of X on the latent variables and U contains the left singular vectors from the SVD decomposition (loading vectors) in column form. Further details of the PLS algorithm and the PLS-DA model can be found in (Sjöström et al. 1986; Varmuza and Filzmoser 2009).

By analysing the directions of maximum variability with this discrimination model, the variables that contribute most to the separation can be identified. This can be understood by closely examining the behaviour of the variables and observations in the loadings and scores plots of the model, respectively.

Given that there are nine clustering configurations, each with its corresponding discrimination model, we selected the configuration that maximized a new index incorporating metrics from the PLS-DA. Such index is designed to balance all metrics while compensating for discrepancies

arising from the fact that each configuration has a different number of clusters. The metrics to be used are:

- **R²Y**: Evaluates the predictability of the model, which may decrease due to granularity effects as the number of clusters increases.
- **RMSEE (**Root Mean Squared Error of Estimation**)**: Measures model fit. It is generally lower with a greater number of clusters.
- **Sensitivity and Specificity**: They reflect the supervised classification and evaluation capabilities of the clusters.

To balance the effects of the number of clusters (*ncl*), we assumed a maximum entropy scenario, where all categories have equal priori probabilities. Probability of correct random classification is defined as 1/*ncl* (the random failure probability is (*ncl*−1)/*ncl*). This value is used to adjust both R²Y and (1 - RMSEE), normalizing them following the methodology of the Practical Significance Index (PSI) introduced by Hubert and Levin (1976). This approach allows us to create an index that comprehensively balances the predictive power and accuracy of the clusters, as reflected in Eq. (5):

$$Index = \frac{1}{4}\left(\frac{0.01R^2y - \left({}^{1}/_{ncl}\right)}{\left(\frac{ncl-1}{ncl}\right)} + \frac{(1 - RMSSE) - \left({}^{1}/_{ncl}\right)}{\left(\frac{ncl-1}{ncl}\right)} + Specificity + Sensitivity\right) \quad (5)$$

The index is consistent, as all metrics used (R²Y, RMSEE, Sensitivity and Specificity) are normalized within the range of 0 to 1. Additionally, the factor 0.01 is applied only if the R²Y is expressed as a percentage. If it was expressed as a proportion, then the 0.01 factor would not be used.

The clustering that is found by maximizing the index is the one that best captures the relationship between the dummy variables (representing the clusters) and the geographical variables. After identifying the optimal cluster configuration and its corresponding discrimination model, the characteristics of each cluster were analysed to identify the variables that most significantly contributed to discrimination.

3 Results and Analysis

In this section, we present the outcomes of our analysis, focusing on the discrimination of climatic time series across Peninsular Spain based on geographical location. The results include the metrics used to determine optimal cluster configurations, as well as the final discrimination model, allowing for a detailed assessment of the changing climatic patterns across different regions. All computations were conducted using R and SAS programming environments.

3.1 Clustering Configurations

Using the three similarity measures outlined in 2.3.2.2 section and the clustering algorithms described in 2.3.3 section, we derived nine distinct clustering configurations (HA-M1, HA-M2, …, SOM-M3). Figure 2a and b illustrates the performance of Dunn's and Xie Beni's indices, which served as internal validation measures to identify the optimal number of clusters for each configuration.

Figure 2a and b illustrates that the configurations generated using the HA algorithm produced more distinct and compact clusters compared to those generated by the PAM and SOM algorithms. This is evidenced by the fact that HA consistently achieved the highest Dunn Index values and the lowest Xie-Beni Index values, whereas the PAM and SOM configurations showed less optimal performance. As shown, the number of clusters varied between 3 and 16 across the configurations.

Table 1 contains the metrics for the optimal number of clusters in each configuration. It provides detailed information about the performance of the PLS-DA models associated with each clustering configuration. This includes the percentage of variance explained by the predictor variables (X, representing geographic variables) and the response variables (Y, representing clusters), as well as key performance metrics: sensitivity (the ability of the model to correctly identify observations belonging to a specific class), specificity (the ability to correctly identify observations that do not belong to a specific class), accuracy (which evaluates the overall performance of the model), and RMSEE, which measures model precision.

The proportion of variance explained by the predictor variables (R²X) remains consistently high—exceeding 84% across all models— indicating that the geographic variables are well-represented. However, the R²Y values, reflecting the variance explained by the response variables (the clusters), are more moderate, ranging from 13.90 to 28.40%. Configurations with a higher number of clusters, particularly those based on the M3 distance or configurations

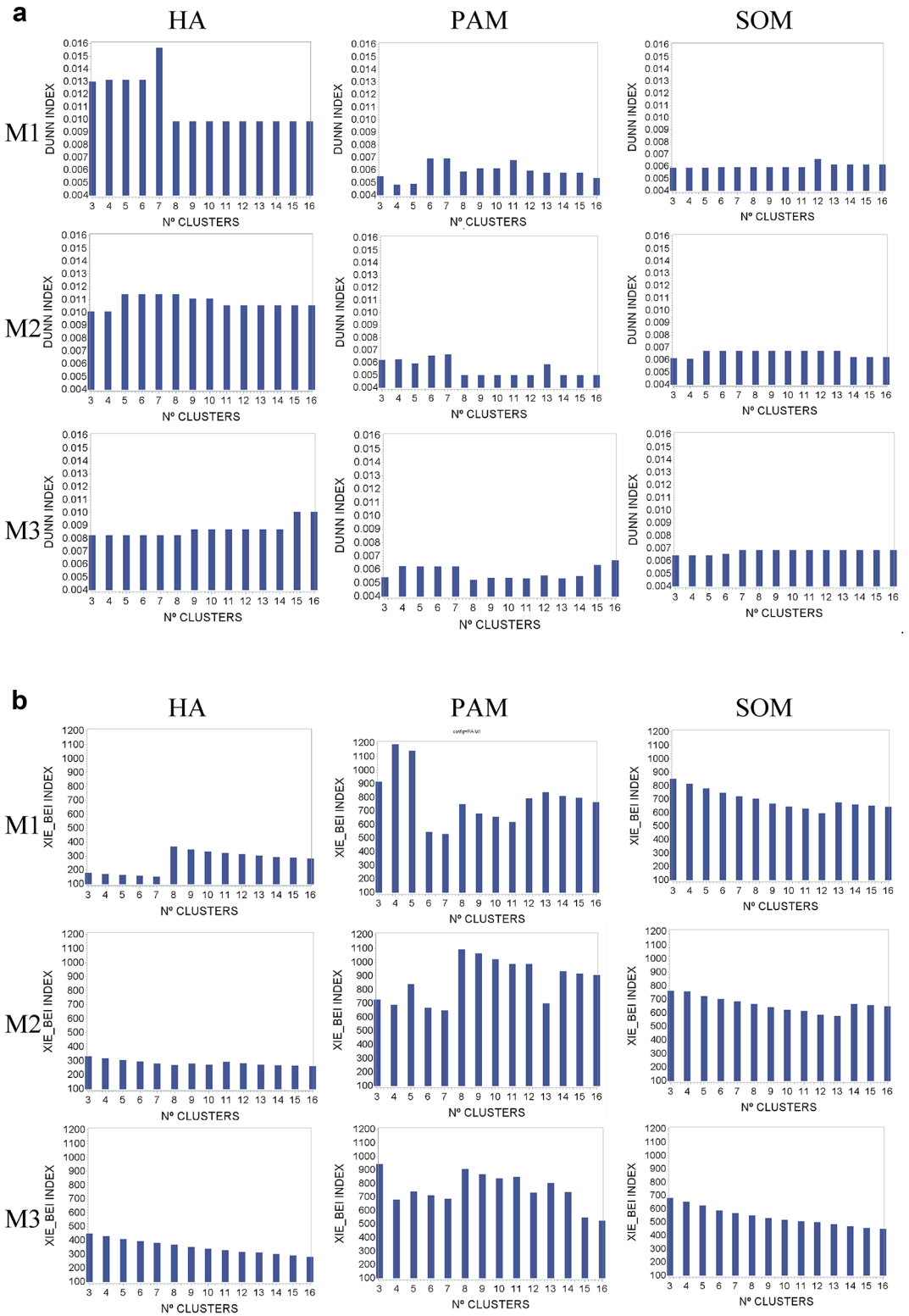


Fig. 2 **a** Dunn Index for K values from 3 to 16 with the different configurations, **b** Xie_Beni Index for K values from 3 to 16 whit the different configurations

Table 1 Explained Variance and Performance Metrics for PLS-DA models by Clustering configuration

Cluster Configuration	Number of Clusters (K)	Explained Variability R^2X	Explained Variability R^2Y	Specificity	Sensitivity	Accuracy	RMSEE	Index
HA - M1	7	85.2	24.9	0.936	0.682	0.606	0.294	0.600
PAM - M1	7	85.2	**28.4**	0.951	**0.728**	**0.697**	0.295	**0.625**
SOM - M1	12	85.6	17.0	0.956	0.521	0.500	0.247	0.576
HA - M2	8	84.2	21.5	0.952	0.692	0.667	0.268	0.610
PAM - M2	7	84.4	24.2	0.939	0.598	0.622	0.301	0.575
SOM - M2	13	84.5	14.1	0.951	0.442	0.413	0.244	0.550
HA - M3	16	**85.7**	15.3	**0.965**	0.479	0.463	**0.219**	0.578
PAM - M3	16	**85.7**	14.7	0.964	0.444	0.438	0.221	0.566
SOM - M3	16	85.3	13.9	0.960	0.406	0.380	0.222	0.553

Fig. 3 Geographical Distribution of clusters in Spain by PAM-M1

derived using the SOM algorithm, tend to exhibit the lowest R^2Y values.

A robust PLS-DA model should achieve high sensitivity, specificity, and accuracy, as these metrics are critical for evaluating classification performance. Most models derived from the various clustering configurations show moderate performance across these metrics. Notably, the PLS-DA model associated with the PAM-M1 configuration stands out, achieving high specificity, moderate sensitivity, and reasonable accuracy, making it the best performer overall. Regarding RMSEE, which measures the average estimation error, its interpretation in this context —where all dependent variables are dummy-coded— becomes proportional, like R^2Y. The estimated values for RMSEE range between 0 and 1, serving as a comparative reference for model precision.

Using these metrics, Eq. 5 computes an index that balances all these characteristics. The index aggregates sensitivity, specificity, accuracy, and RMSEE, providing a comprehensive evaluation of each model's overall performance. Table 1 summarizes the explained variance and performance metrics for the PLS-DA models across all configurations. In conclusion, the PAM-M1 configuration

Table 2 Centroid values for climate variables by PAM-M1 clusters

Cluster	$TMAX_{mean}$ (°C)	$TMAX_{max}$ (ºC)	$TMIN_{mean}$ (ºC)	$TMIN_{min}$ (ºC)	$PRCP$ (mm)	$PRCP_{max}$ (mm)	Consecutive Dry Days
1	23.0	39.2	10.8	−0.6	643.7	45.4	60.0
2	15.4	32.7	5.3	−6.3	1381.3	49.8	19
3	19.8	38.1	7.3	−4.6	551.3	37.0	37
4	21.6	39.1	9.5	−2.8	656.9	38.6	46
5	19.5	36.8	8.0	−6.4	510.5	34.7	31
6	15.4	35.0	3.7	−7.4	741.0	37.6	23
7	17.3	35.4	4.9	−7.1	579.6	35.1	31

Table 3 Rate annual of change of climate variables by cluster with PAM-M1 configuration

Cluster	$TMAX_{mean}$ (°C)	$TMAX_{max}$(ºC)	$TMIN_{mean}$(ºC)	$TMIN_{min}$ (ºC)	$PRCP$ (mm)	$PRCP_{max}$ (mm)	Consecutive Dry days
1	0.015	0.034	0.015	0.005	−1.42	0.199	0.183
2	0.018	0.030	0.021	0.028	−1.53	0.165	0.025
3	0.022	0.036	0.020	0.014	−1.71	0.054	0.150
4	0.018	0.031	0.010	0.004	−1.79	0.091	0.135
5	0.017	0.033	0.015	0.012	−0.507	0.188	0.064
6	0.025	0.035	0.021	0.032	−1.28	0.115	0.063
7	0.021	0.028	0.012	0.016	−1.29	0.037	0.012

Table 4 Geographical characteristics of PAM-M1 clusters

Cluster	Sea_distance (km)	Height (m)	Latitude (º)	Longitude (º)
1	50.63	264.8	37.39	−4.696
2	63.15	727.5	42.64	−4.895
3	204.10	754.8	39.25	−3.218
4	185.70	485.0	38.82	−5.382
5	88.44	579.9	40.26	−0.866
6	133.86	1095.0	41.52	−1.809
7	185.20	839.3	41.47	−4.758

achieves the highest index value, indicating a well-balanced performance across all metrics and suggesting that it is the most suitable.

3.2 Regional Characteristics

Based on the information presented, particularly as detailed in Table 1, the PAM-M1 configuration emerges as the most effective for discriminating climate changes across Peninsular Spain. Figure 3 illustrates the spatial distribution of clusters. Notably, some clusters exhibit significant dispersion throughout the Iberian Peninsula, with orientations extending from north to south and others from east to west. The proposal of various clustering configurations, helps reducing uncertainty in the approach and improves optimization in terms of both explanatory power and performance.

Spain is one of the most affected European countries by climate change. Tables 2 and 3 summarise the centroid values for each cluster across several climate parameters, while Table 4 provides the average geographical characteristics of each cluster. These tables reveal critical trends: droughts are anticipated to become more frequent, daily maximum precipitation is expected to increase, overall rainfall is projected to decline, and extreme maximum temperatures are likely to intensify.

The annual increase in average maximum temperatures is less pronounced than the rise in peak maximum temperatures. However, minimum temperatures are expected to mitigate, resulting in a decrease in the occurrence of severe frosts. The observed annual rise in the highest daily temperature is consistent across all clusters, with an average increase of 0.3 °C per decade. The following differences are observed among the seven identified clusters.

Cluster 1 This region shows very high summer temperatures and mild winters. It is characterized by a significant increase in torrential rainfall (maximum precipitation in a single day) and a more pronounced advance in droughts (a greater increase in consecutive dry days). It corresponds to most of western Andalucía and the Mediterranean coast in the south, corresponding to Murcia and the southern Valencia.

Cluster 2 Located in an area with the smooth climate, this cluster receives the highest annual rainfall and records the highest daily precipitation values. It shows high annual increase in minimum temperatures and a slower progression of droughts. This cluster is located along the Atlantic coast in northern of Spain.

Cluster 3 This cluster stands out for its significant reduction in annual precipitation and the slowest growth in torrential rains, while experiencing a very high increase of maximum

temperatures during warm summers. It is formed by the centre and south of Madrid, a large part of Castilla La Mancha and part of the interior of Eastern Andalucía.

Cluster 4 Characterized by extremely high maximum temperatures, this cluster exhibits the smallest increase in minimum temperatures and is the cluster whit highest rate of reduction of total annual precipitation. It is in Extremadura and northwest Andalucía.

Cluster 5 With low precipitation, this cluster shows the smaller reduction in total annual rainfall. It includes western Castilla La Mancha, the Valencian Community, southern Cataluña and the non-mountainous zone of Aragón.

Cluster 6 This cluster has the lowest average maximum and minimum temperatures, but these temperatures are projected to increase more than in other clusters. It is the most geographically dispersed, representing mountainous areas with the highest altitudes.

Cluster 7 It is characterized by having the lowest rates of extreme precipitations, both in terms of drought (smallest increase in consecutive dry days) and torrential rainfall. Although the rates are very similar to those of the other clusters, it also exhibits the lowest increase in extreme maximum temperatures. It includes the interior regions of Castilla León.

Figure 4 shows longitudinal analyses, clearly illustrating significant long-term increases in extreme temperatures and the frequency of dry days across all studied regions. Simultaneously, a significant decrease in precipitation over time underscores the growing prevalence of drought in Peninsular Spain in recent years.

The results presented in this figure align with other studies and the IPCC reports, which indicate that temperatures across Europe, including Spain, are increasing at a rate exceeding 1.5 °C per century target (Masson-Delmotte et al., 2021). Similarly, Del Río et al. (2012) observed that maximum temperatures in Spain have risen faster than minimum temperatures, a trend also is identified in this study.

The trends in extreme temperatures identified in this work are consistent with previous research, though they indicate slightly higher and more concerning growth rates. For instance, Sigro et al. (2024) analysed temperature and precipitation trends in the Pyrenees and Sierra Nevada, two regions located in the north and south of Spain, respectively. Their study reported temperature increases of approximately 0.17 °C and 0.13 °C per decade, respectively, over the period 1930–2020. These regions correspond to Cluster 6 in the present study, characterized by mountainous zones, where higher increases in extreme temperatures were observed, around 0.25 °C per decade for T_{max} and approximately 0.21 °C for T_{min}.

Recently, Moreno et al. (2024) reported rising maximum temperatures and intensified drought over the past two

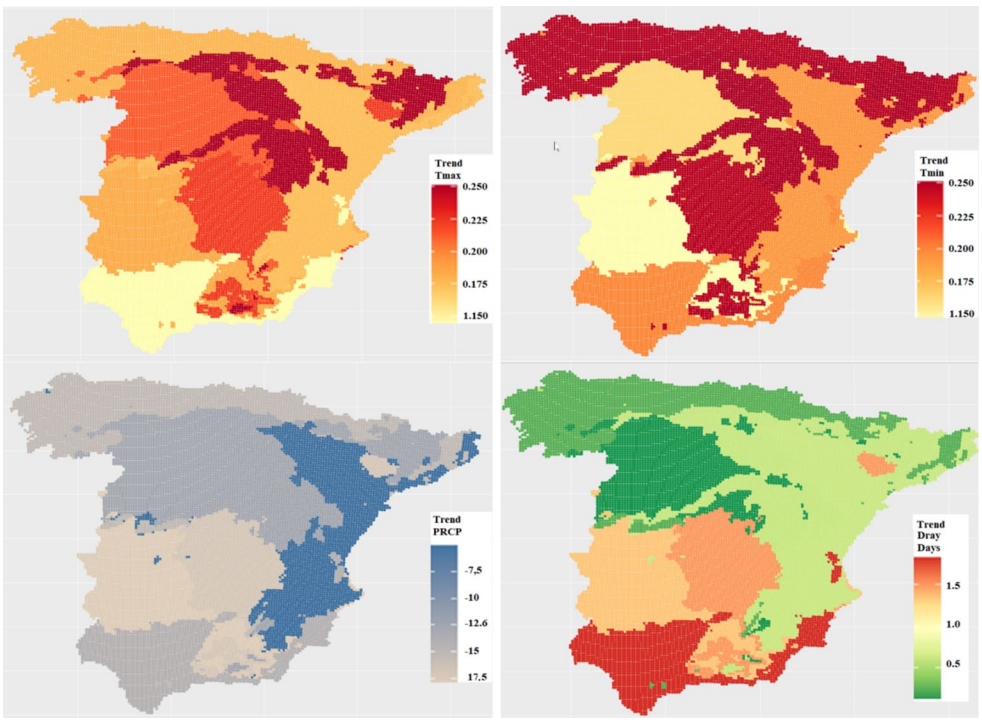


Fig. 4 Rates of change for Tmax, Tmin, PRCP and consecutive Dry Days by decade in each cluster (PAM-M1 configuration)

decades in Seville and Almería, two southern Spanish cities located within Cluster 1. This cluster not only exhibits the highest average T_{max} and the largest number of dry days but also shows the fastest growth rate in the number of dry days.

On average, the results of Table 3 and Fig. 4 show decreasing annual precipitation trends across Peninsular Spain. These findings are consistent with those of Senent-Aparicio et al. (2023), who reported decreasing precipitation trends on an annual scale, throughout most of Spain, except for some isolated areas in the north. Cluster 5, which covers large parts of Aragón, Navarra, and La Rioja in northern Spain, shows the smallest reduction in precipitation. This suggests that certain regions within this cluster may exhibit stable or non-declining trends.

The observed increases in consecutive dry days follow a pattern similar to the precipitation trends, further supporting Senent-Aparicio et al. (2023) findings of a general decline in the number of rainy days across the country.

Additionally, Meseguer-Ruiz et al. (2021) found an upward trend in heavy rainfall events in eastern and southeastern Spain between 1950 and 2016, while González Hidalgo et al. (2003) noted a rise in extreme precipitation events in the Valencian Community. These regions are mostly contained within Cluster 5, where certain areas show positive trends in maximum precipitation levels.

3.3 Quantifying the Impacts of Geographic Variables on Regional Climate Change

To further explore the role of geographical variables in differentiating clusters within the PAM-M1 configuration, we analysed the components derived from the PLS-DA model. Figures 5 and 6 illustrate the loadings and scores for the first two components.

The loadings plot in Fig. 5 reveals distinct patterns of climate variability across geographic locations. The first component differentiates between areas characterised by high latitude, elevated altitude, and proximity to the sea, and those with lower latitude, below-average altitude, and greater distance from the coast. The second component captures variability based on longitudinal differences, underscoring climatic distinctions between regions with broader and narrower longitudinal spans. This spatial distribution of scores reveals marked contrasts between northern and southern regions and between eastern and western areas, illustrating the climatic diversity of Spain. These findings highlight the well-established north-south climate gradient, with additional influences along the east-west axis. Future research could delve further into the roles of altitude and coastal proximity in shaping Spain's climate and their potential variations over time.

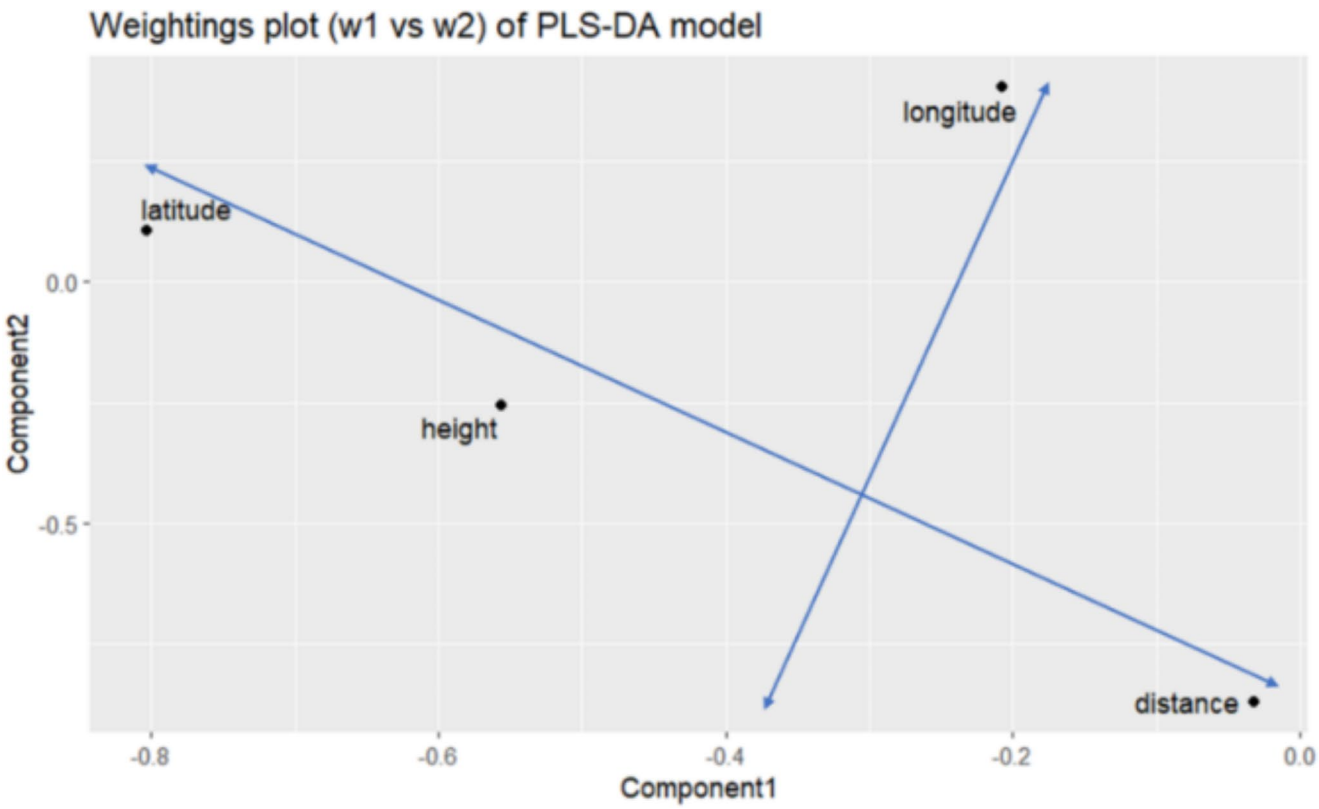


Fig. 5 Loadings Plot for the first two components of the PLS-DA model (PAM-M1 configuration)

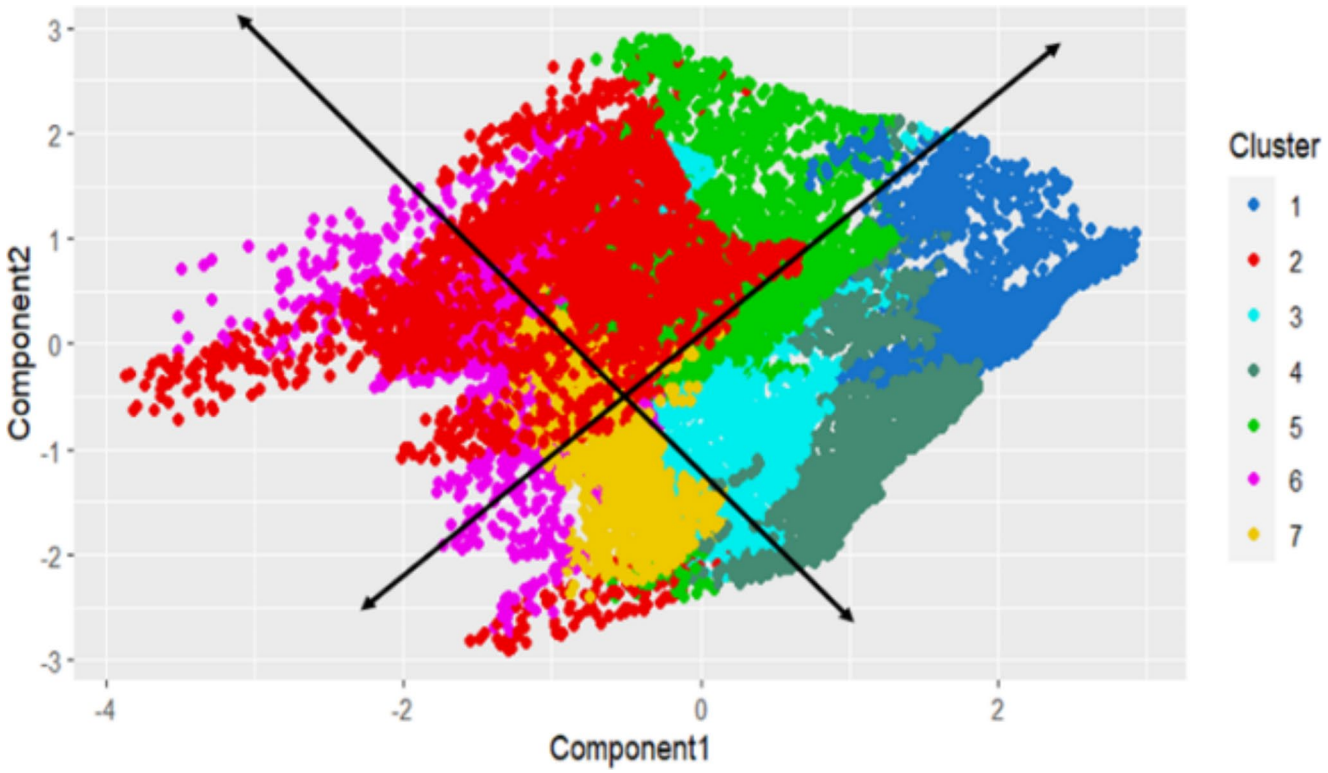


Fig. 6 Score plot for the first two components of the PLS-DA model (PAM-M1 configuration)

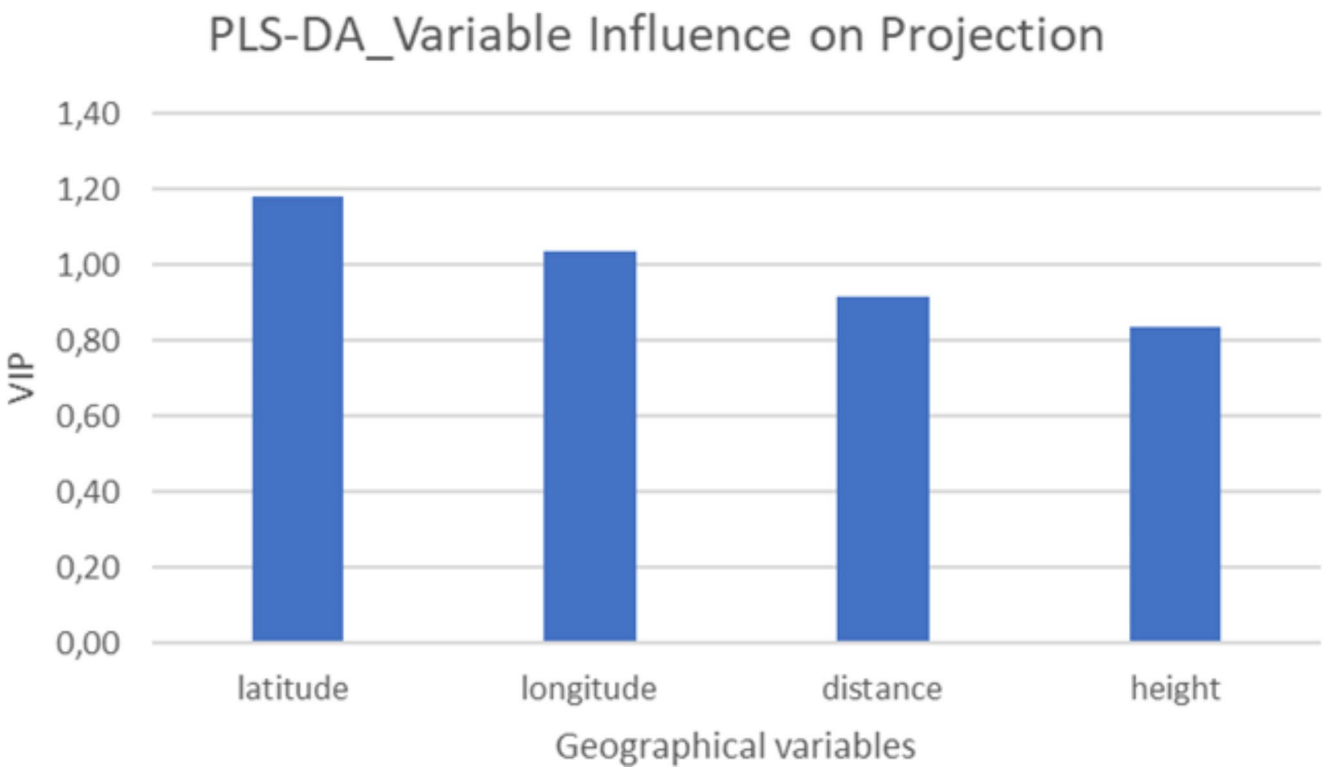


Fig. 7 Importance of geographical variables in the PLS-DA model for climate change discrimination

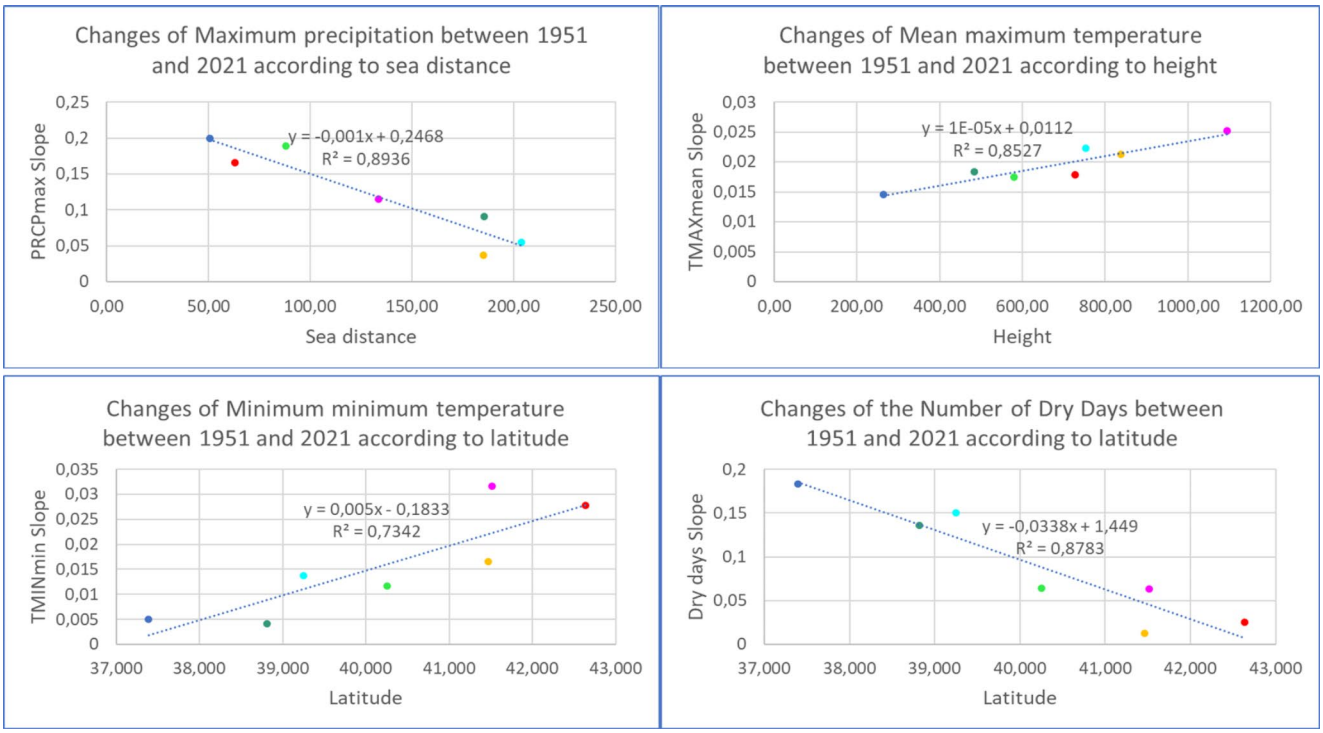


Fig. 8 Rates of change of climatic variables in cluster from geographic variables

An analysis of the Variable Importance in the Projection (VIP) scores (Fig. 7) identifies longitude and latitude are the most influential geographical variables, reinforcing the significance of the north-south and east-west climate gradients in Peninsular Spain. These results highlight the country's climatic diversity, shaped by the north-south gradient, and the distinct contrast between cooler Atlantic currents and the warmer Mediterranean waters, which typically influence weather patterns from east to west (Dasari et al. 2014).

To further investigate the influence of geographical factors on climate change in Peninsular Spain, we analysed key relationships between geographical variables and climatic trends across the clusters. Figure 8 provides a comprehensive overview of these relationships.

The regression models displayed in Fig. 8 are statistically significant, with geographical variables explaining a substantial portion of the variability in climate changes indicators. Regions closer to the sea, including parts of Andalucía, Murcia, Valencia, Aragón, Cataluña, and northern Spain, exhibit the largest increases in maximum precipitation. Unlike average precipitation, maximum precipitation shows positive trends across all clusters.

A strong correlation exists between increases in mean maximum temperature and altitude, indicating that higher-altitude areas, particularly in mountainous regions, have the greatest increases in maximum temperature. Conversely, reductions of extremes of minimum temperature are more pronounced in mountainous and northern regions, which are located at higher latitudes.

The relationship between the number of dry days and altitude reveals that northern and mountainous region, characterized by higher latitudes, experience fewer drought conditions compared to the southern zones, particularly Andalucía, which shows higher drought levels.

Additionally, the increase in the lowest annual minimum temperature is more significant in mountainous and northern regions, where higher latitudes correspond to greater temperature rises. Similarly, the analysis of consecutive dry days and latitude indicates that northern regions and mountainous areas, with higher latitudes, exhibit lower drought levels than southern regions, such as Andalucía, which experiences more frequent and prolonged droughts.

By contrasting the geographical data from the identified areas, certain climate characteristics and variations across Spain have emerged. These findings offer valuable insights into the overall climate change observed in Peninsular Spain.

4 Conclusions

This study presents a comprehensive analysis of climate variability across Peninsular Spain, employing an innovative methodology that integrates climate change estimates

and geographic variables into the clustering process. The analysis incorporates the influence of geographical factors and uses various clustering techniques (PAM, HA and SOM) alongside PLS-DA. This approach provides a clearer understanding of climate change patterns across different regions. The methodology effectively links spatial and temporal climate trends to geographical variables, leading to more refined regional classifications. The key findings from the application of this methodology in Spain are:

- Temperature Trends. Maximum temperatures have increased at a faster rate than minimum temperatures, aligning with prior studies. The most significant rises in extreme temperatures are observed in mountainous regions and areas at a moderate distance from the sea, particularly in northern regions. Southern regions, such as Andalucía and Murcia, exhibit smaller increases in both maximum and minimum temperatures.
- Cold Temperature Extremes. The study confirms an increase of the lowest minimum temperatures, therefore a notable reduction in cold temperature extremes is observed, particularly in the coldest areas. Contrary to the idea that climate change increases the number of extreme events in the Iberian Peninsula, this research shows that only the warm temperature extremes are increasing.
- Precipitation. A substantial decrease in precipitation is observed, especially in inland areas like Castilla-La Mancha and Extremadura. In contrast, coastal regions near the Mediterranean exhibit smaller reduction in precipitation.
- Extreme Precipitation Events. Annual maximum precipitation in a single day is increasing across all areas, particularly in regions near the Mediterranean Sea. This trend indicates a growing risk of extreme weather events, including flooding, consistent with findings from previous works.
- Drought Conditions. The frequency of dry days has markedly increased, particularly in southern Spain, heightening the vulnerability of regions such as Andalucía and Murcia to intensified drought conditions, echoing previous research.
- Geographical Influence. Latitude, longitude, and altitude play a pivotal role in determining regional climate trends. Northern and mountainous regions experience fewer droughts but larger increases in minimum temperatures, while southern and coastal areas are more susceptible to drought and exhibit greater variability in extreme precipitation.

These findings demonstrate that the proposed methodology enhances the explanatory power of PLS-DA models for climate change discrimination based on geographic variables. Future research could refine the method by assigning different weights to parameters: initial points, slope, and autocorrelations, instead of focusing solely on weight on climate variables, or exploring a combination of both approaches. This innovative method, which incorporates position, trend, and autocorrelation parameters for each variable in cluster formation, admits modifications. The key concept underlying this approach is the necessity of including these three parameters. In certain cases, autocorrelation, although necessarily included in the model as described in Eq. 1, might be omitted from the vector, or its influence could be reduced by assigning it a lower weight relative to the other two parameters. This adjustment is applied to cases where the autocorrelations do not have significantly influence on cluster formation. In this study although many parameters are zero due to their lack of significance, the presence of autocorrelation in certain areas acts as a discrimination factor that requires consideration. Moreover, it directly influences of the slope and position parameter values.

Additionally, while the generated clusters using HA were more compact and better defined according to internal validation indices, they did not significantly improve the explained variance of the PLS-DA models. Therefore, the applicability of other indices could be explored in future studies to further optimize the method's performance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41748-025-00568-4.

Acknowledgements For generating and distributing the data we have used in this study, we express our gratitude to "Servicio de Desarrollos Climatológicos" of the Meteorological Spanish State Agency. Likewise, we are grateful to the Colombian Ministry of Science and the Technological University of Chocó for supporting the doctoral formation of Arnobio Palacios. The research is also supported by a Grant from Agencia Estatal de Investigación of Spain, with reference (PID2019-106433GB-I00/AEI/https://doi.org/10.13039/501100011033).

Author contributions Conceptualization: Arnobio Palacios, José Luis Valencia, María Villeta; Methodology: Arnobio Palacios, José Luis Valencia; Data acquisition and resources: José Luis Valencia; Data processing: Arnobio Palacios, José Luis Valencia; Validation: Arnobio Palacios, José Luis Valencia, María Villeta; Formal analysis and investigation: Arnobio Palacios, José Luis Valencia, María Villeta; Writing - original draft preparation: Arnobio Palacios; Writing - review and editing: José Luis Valencia, María Villeta; Funding acquisition: María Villeta; Supervision: José Luis Valencia, María Villeta.

Declarations

Conflict of interest We declare that no financial interests or personal relationships influenced the work reported in this article.

References

Aghabozorgi S, Seyed Shirkhorshidi A, Ying Wah T (2015) Time-series clustering - A decade review. Inform Syst 53:16–38. https://doi.org/10.1016/j.is.2015.04.007

Bro R, Smilde AK (2003) Centering and scaling in component analysis. J Chemom 17(1):16–33. https://doi.org/10.1002/cem.773

Brown ME, Antle JM, Backlund P, Carr ER, Easterling WE, Walsh MK, Ammann C, Attavanich W, Barrett CB, Bellemare MF, Dancheck V, Funk C, Grace K, Ingram JSI, Jiang H, Maletta H, Mata T, Murray A, Ngugi M, Tebaldi C (2015) *Climate Change, Global Food Security, and the U.S. Food System*. https://doi.org/10.7930/J0862DC7

Brun M, Sima C, Hua J, Lowey J, Carroll B, Suh E, Dougherty ER (2007) Model-based evaluation of clustering validation measures. Pattern Recogn 40(3):807–824. https://doi.org/10.1016/j.patcog.2006.06.026

Brunner L, McSweeney C, Ballinger AP, Befort DJ, Benassi M, Booth B, Coppola E, de Vries H, Harris G, Hegerl GC, Knutti R, Lenderink G, Lowe J, Nogherotto R, O'Reilly C, Qasmi S, Ribes A, Stocchi P, Undorf S (2020) Comparing methods to Constrain Future European Climate projections using a consistent Framework. J Clim 33(20):8671–8692. https://doi.org/10.1175/JCLI-D-19-0953.1

Chen XL, Zhao HM, Li PX, Yin ZY (2006) Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. Remote Sens Environ 104(2):133–146. https://doi.org/10.1016/j.rse.2005.11.016

Choularton R, Krishnamurthy PK, Lewis K (2012) *Climate impacts on food security and nutrition - A Review of Existing Knowledge*

Ciais P, Reichstein M, Viovy N, Granier A, Ogée J, Allard V, Aubinet M, Buchmann N, Bernhofer C, Carrara A, Chevallier F, De Noblet N, Friend AD, Friedlingstein P, Grünwald T, Heinesch B, Keronen P, Knohl A, Krinner G, Valentini R (2005) Europe-wide reduction in primary productivity caused by the heat and drought in 2003. Nature 437(7058):529–533. https://doi.org/10.1038/nature03972

Dasari HP, Pozo I, Ferri-Yáñez F, Araújo MB (2014) A Regional Climate study of heat waves over the Iberian Peninsula. Atmospheric Clim Sci 04(05):841–853. https://doi.org/10.4236/acs.2014.45074

de Lucena AFP, Szklo AS, Schaeffer R, de Souza RR, Borba BSMC, da Costa IVL, Júnior AOP, da Cunha SH F (2009) The vulnerability of renewable energy to climate change in Brazil. Energy Policy 37(3):879–889. https://doi.org/10.1016/j.enpol.2008.10.029

del Río S, Cano-Ortiz A, Herrero L, Penas A (2012) Recent trends in mean maximum and minimum air temperatures over Spain (1961–2006). Theoret Appl Climatol 109(3–4):605–626. https://doi.org/10.1007/s00704-012-0593-2

Desgraupes B (2017) *clusterCrit: Clustering Indices*. https://CRAN.R-project.org/package=clusterCrit

Dunn JC (1974) Well-separated clusters and optimal fuzzy partitions. J Cybernetics 4(1):95–104. https://doi.org/10.1080/01969727408546059

Fonseca D, Carvalho MJ, Marta-Almeida M, Melo-Gonçalves P, Rocha A (2016) Recent trends of extreme temperature indices for the Iberian Peninsula. Phys Chem Earth 94:66–76. https://doi.org/10.1016/j.pce.2015.12.005

Furió D, Meneu V (2011) Analysis of extreme temperatures for four sites across Peninsular Spain. Theoret Appl Climatol 104(1–2):83–99. https://doi.org/10.1007/s00704-010-0324-5

Gallant AR, Goebel JJ (1976) Nonlinear regression with autocorrelated errors. J Am Stat Assoc 71(356):961–967. https://doi.org/10.2307/2286869

García DH (2022) Analysis of Urban Heat Island and heat waves using Sentinel-3 images: a study of andalusian cities in Spain. Earth Syst Environ 6(1):199–219. https://doi.org/10.1007/s41748-021-00268-9

González Hidalgo JC, De Luís M, Raventós J, Sánchez JR (2003) Daily rainfall trend in the Valencia Region of Spain. Theoret Appl Climatol 75(1):117–130. https://doi.org/10.1007/s00704-002-0718-0

Gupta AK, Negi M, Nandy S, Alatalo JM, Singh V, Pandey R (2019) Assessing the vulnerability of socio-environmental systems to climate change along an altitude gradient in the Indian Himalayas. Ecol Ind 106:105512. https://doi.org/10.1016/j.ecolind.2019.105512

Hartigan JA, Wong MA (1979) Algorithm AS 136: a K-Means Clustering Algorithm. Appl Stat 28(1):100. https://doi.org/10.2307/2346830

Hubert LJ, Levin JR (1976) A general statistical framework for assessing categorical clustering in free recall. Psychol Bull 83(6):1072–1080

Imam MH, Rahman MM, Roy S, Hoque F, Ahsan U, Abdullah SMA, Hossain MS, Rahim MA (2022) Analysis of diurnal air temperature range variation over Bangladesh. Earth Syst Environ 6(2):361–373. https://doi.org/10.1007/s41748-021-00282-x

IPCC: Masson-Delmotte V, Zhai P, Chen Y, Goldfarb L, Gomis MI, Matthews JBR, Berger S, Huang M, Yelekçi O, Yu R, Zhou B, Lonnoy E, Maycock TK, Waterfield T, Leitzell K, Caud N (2021) *Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change Edited by*. Climate Change 2021: The Physical Science Basis. www.ipcc.ch

Kaufman L, Rousseeuw PJ (1990) In: Kaufman L, Rousseeuw PJ (eds) Finding groups in data: an introduction to Cluster Analysis. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316801

Khan AA, Zhao Y, khan J, Rahman G, Rafiq M, Moazzam MFU (2021) Spatial and temporal analysis of Rainfall and Drought Condition in Southwest Xinjiang in Northwest China, using various climate indices. Earth Syst Environ 5(2):201–216. https://doi.org/10.1007/s41748-021-00226-5

Knutti R (2010) The end of model democracy? Clim Change 102(3–4):395–404. https://doi.org/10.1007/s10584-010-9800-2

KOHONEN T, OJA E (1996) Engineering applications of the self-organizing map. https://doi.org/10.1109/5.537105

Kolusu SR, Siderius C, Todd MC, Bhave A, Conway D, James R, Washington R, Geressu R, Harou JJ, Kashaigili JJ (2021) Sensitivity of projected climate impacts to climate model weighting: multi-sector analysis in eastern Africa. Clim Change 164(3–4):36. https://doi.org/10.1007/s10584-021-02991-8

Kraus JM, Müssel C, Palm G, Kestler HA (2011) Multi-objective selection for collecting cluster alternatives. Comput Stat 26(2):341–353. https://doi.org/10.1007/s00180-011-0244-6

Kremer H, Kranen P, Jansen T, Seidl T, Bifet A, Holmes G, Pfahringer B (2011) An effective evaluation measure for clustering on evolving data streams. Proc ACM SIGKDD Int Conf Knowl Discovery Data Min 868:876. https://doi.org/10.1145/2020408.2020555

Kuriqi A, Ali R, Pham QB, Gambini M, Gupta J, Malik V, Linh A, Joshi NTT, Anh Y, Nam DT, V. T., Dong X (2020) Seasonality shift and streamflow flow variability trends in central India. Acta Geophys 68(5):1461–1475. https://doi.org/10.1007/s11600-020-00475-4

Laepple T, Huybers P (2014) Ocean surface temperature variability: large model–data differences at decadal and longer periods. Proc Natl Acad Sci 111(47):16682–16687. https://doi.org/10.1073/pnas.1412077111

Malede DA, Agumassie TA, Kosgei JR, Andualem TG, Diallo I (2022) Recent approaches to Climate Change impacts on Hydrological

extremes in the Upper Blue Nile Basin, Ethiopia. Earth Syst Environ 6(3):669–679. https://doi.org/10.1007/s41748-021-00287-6

Meseguer-Ruiz O, Lopez-Bustins JA, Arbiol-Roca L, Martin-Vide J, Miró J, Estrela MJ (2021) Temporal changes in extreme precipitation and exposure of tourism in Eastern and South-Eastern Spain. Theoret Appl Climatol 144(1–2):379–390. https://doi.org/10.1007/s00704-021-03548-6

Młyński D, Wałęga A, Kuriqi A (2021) Influence of meteorological drought on environmental flows in mountain catchments. Ecol Ind 133:108460. https://doi.org/10.1016/j.ecolind.2021.108460

Moreno M, Barea R, Castro L, Cagigas D, Ortiz R, Ortiz P (2024) Climate Change monitoring with art-risk 5: New approach for environmental hazard assessment in Seville and Almería Historic centres (Spain). Procedia Struct Integr 55:9–17. https://doi.org/10.1016/j.prostr.2024.02.002

Murtagh F, Legendre P (2014) Ward's hierarchical agglomerative clustering method: which Algorithms Implement Ward's Criterion? J Classif 31(3):274–295. https://doi.org/10.1007/s00357-014-9161-z

O'neill MS, Ebi KL (2009) Temperature extremes and health: impacts of climate variability and change in the. Source: J Occup Environ Med 51(1):13–25. https://doi.org/10.1097/JOM.0b013e318173el22

Olteanu M, Villa-Vialaneix N (2015) Using SOMbrero for clustering and visualizing graphs Titre: Utiliser SOMbrero pour la classification et la visualisation de graphes. J Soc Fr Stat 156(3):95–119. http://www.sfds.asso.fr/journal

Palacios Gutiérrez A, Valencia Delfa JL, Villeta López M (2023) Time series clustering using trend, seasonal and autoregressive components to identify maximum temperature patterns in the Iberian Peninsula. Environ Ecol Stat 30(3):421–442. https://doi.org/10.1007/s10651-023-00572-9

Parracho AC, Melo-Gonçalves P, Rocha A (2016) Regionalisation of precipitation for the Iberian Peninsula and climate change. Phys Chem Earth Parts A/B/C 94:146–154. https://doi.org/10.1016/J.PCE.2015.07.004

Patz JA, Campbell-Lendrum D, Holloway T, Foley JA (2005) Impact of regional climate change on human health. Nat 438(7066):310–317.https://doi.org/10.1038/nature04188

Peña-Angulo D, Cortesi N, Brunetti M, González-Hidalgo JC (2015) Spatial variability of maximum and minimum monthly temperature in Spain during 1981–2010 evaluated by correlation decay distance (CDD). Theoret Appl Climatol 122(1–2):35–45. https://doi.org/10.1007/s00704-014-1277-x

Ramos MC, Balasch JC, Martínez-Casasnovas JA (2012) Seasonal temperature and precipitation variability during the last 60 years in a Mediterranean climate area of Northeastern Spain: a multivariate analysis. Theoret Appl Climatol 110(1–2):35–53. https://doi.org/10.1007/s00704-012-0608-z

Roushangar K, Alizadeh F (2018) A multiscale spatio-temporal framework to regionalize annual precipitation using k-means and self-organizing map technique. J Mt Sci 15(7):1481–1497. https://doi.org/10.1007/s11629-017-4684-5

Samantaray AK, Mitra A, Ramadas M, Panda RK (2021) Regionalization of hydroclimatic variables using Markov random field model for climate change impact assessment. J Hydrol 596:126071. https://doi.org/10.1016/J.JHYDROL.2021.126071

Sathaye JA, Dale LL, Larsen PH, Fitts GA, Koy K, Lewis SM, de Lucena AFP (2013) Estimating impacts of warming temperatures on California's electricity system. Glob Environ Change 23(2):499–511. https://doi.org/10.1016/j.gloenvcha.2012.12.005

Saunders KR, Stephenson AG, Karoly DJ (2021) A regionalisation approach for rainfall based on extremal dependence. Extremes 24(2):215–240. https://doi.org/10.1007/s10687-020-00395-y

Schaeffer R, Szklo AS, Pereira de Lucena AF, Cesar Borba M, Nogueira BSP, Fleming LP, Troccoli FP, Harrison A, M., Boulahya MS (2012) Energy sector vulnerability to climate change: A review. In *Energy* (Vol. 38, Issue 1, pp. 1–12). Elsevier Ltd. https://doi.org/10.1016/j.energy.2011.11.056

Senent-Aparicio J, López-Ballesteros A, Jimeno-Sáez P, Pérez-Sánchez J (2023) Recent precipitation trends in Peninsular Spain and implications for water infrastructure design. J Hydrology: Reg Stud 45:101308. https://doi.org/10.1016/j.ejrh.2022.101308

Shi P, Sun S, Gong D, Zhou T (2016) World regionalization of Climate Change (1961–2010). Int J Disaster Risk Sci 7(3):216–226. https://doi.org/10.1007/s13753-016-0094-5

Shin Y, Lee Y, Park J-S (2020) A weighting Scheme in a Multi-model Ensemble for Bias-Corrected Climate Simulation. Atmosphere 11(8):775. https://doi.org/10.3390/atmos11080775

Sigro J, Cisneros M, Perez-Luque AJ, Perez-Martinez C, Vegas-Vilarrubia T (2024) Trends in temperature and precipitation at high and low elevations in the main mountain ranges of the Iberian Peninsula (1894–2020): the Sierra Nevada and the Pyrenees. Int J Climatol 44(9):2897–2920. https://doi.org/10.1002/joc.8487

Sjöström M, Wold S, Söderström B (1986) PLS DISCRIMINANT PLOTS. In *Pattern Recognition in Practice* (pp. 461–470). Elsevier. https://doi.org/10.1016/B978-0-444-87877-9.50042-X

Song M, Zhang L (2008) Comparison of cluster representations from partial second- to full fourth-order cross moments for data stream clustering. Proc - IEEE Int Conf Data Min ICDM 560–569. https://doi.org/10.1109/ICDM.2008.143

Teodoro TA, Reboita MS, Llopart M, da Rocha RP, Ashfaq M (2021) Climate Change impacts on the South American Monsoon System and its surface–atmosphere processes through RegCM4 CORDEX-CORE projections. Earth Syst Environ 5(4):825–847. https://doi.org/10.1007/s41748-021-00265-y

Varmuza K, Filzmoser P (2009) Introduction to Multivariate Statistical Analysis in Chemometrics. CRC. https://doi.org/10.1201/9781420059496

Wang X, Hyndman R (2006) Characteristic-based clustering for Time Series Data. Data Min Knowl Disc 13:335–364. https://doi.org/10.1007/s10618-005-0039-x

Weigel AP, Liniger MA, Appenzeller C (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q J R Meteorol Soc 134(630):241–260. https://doi.org/10.1002/qj.210

Wootten AM, Massoud EC, Waliser DE, Lee H (2023) Assessing sensitivities of climate model weighting to multiple methods, variables, and domains in the south-central United States. Earth Sys Dyn 14(1):121–145. https://doi.org/10.5194/esd-14-121-2023

Xie XL, Beni G (1991) A validity measure for fuzzy clustering. IEEE Trans Pattern Anal Mach Intell 13(8):841–847. https://doi.org/10.1109/34.85677

Yue S, Hashino M, LONG TERM TRENDS OF ANNUAL AND MONTHLY PRECIPITATION IN JAPAN 1 (2003) JAWRA J Am Water Resour Association 39(3):587–596. https://doi.org/10.1111/j.1752-1688.2003.tb03677.x

Zhao Y, Karypis G (2002) Evaluation of hierarchical clustering algorithms for document datasets. Int Conf Inform Knowl Manage Proc 515–524. https://doi.org/10.1145/584792.584877

Capítulo 6

Capítulo 6

Discusión de Resultados

Resumen: *En este capítulo se presenta un análisis integrador de las principales contribuciones y hallazgos de la tesis, enmarcados en los procedimientos generados para la identificación de patrones de cambio climático mediante el clustering de series temporales con reducción de la dimensionalidad.*

Discusión

Al igual que en otros ámbitos del conocimiento, los sistemas climáticos generan datos de series temporales que se caracterizan por su complejidad y alta dimensionalidad. Los métodos de reducción de la dimensionalidad pretenden simplificar este tipo de datos conservando sus características esenciales (Aghabozorgi et al., 2015), lo que permite un análisis y una agrupación más eficaz. En este contexto, estos métodos actúan como etapas de preprocesamiento para comprimir la información conservando sus estructuras, por lo que además de ser útil para mejorar la eficiencia computacional, pueden ayudar al tratamiento de datos multivariantes. Esto ocurre cuando se analizan simultáneamente múltiples variables climáticas (por ejemplo, temperaturas, precipitaciones, etc.). En tales casos las técnicas de reducción de la dimensionalidad permiten abordar de manera adecuada la complejidad de las series temporales multivariantes, logrando una representación más completa y precisa de la dinámica climática regional.

En esta tesis doctoral se han desarrollado nuevos procedimientos para agrupar series temporales haciendo uso de los métodos de representación o de reducción de la dimensionalidad para identificar patrones de cambio climático. Las contribuciones de esta tesis subrayan colectivamente la importancia de las técnicas avanzadas de agrupación y descomposición para comprender la variabilidad climática. A continuación, se presenta una interpretación de los hallazgos generales de esta investigación en términos de tendencias y cambios detectados, avances en la regionalización y la investigación climática, y de implicaciones para las investigaciones climáticas en general.

- **Tendencias y cambios detectados**

Revelando perspectivas únicas, los estudios de los capítulos 3, 4 y 5 ponen de manifiesto una serie de dinámicas climáticas.

Tendencias en las temperaturas:

En el capítulo 3 se evidenció un aumento de las temperaturas máximas en las zonas norte y centro de la Península Ibérica, mientras que la zona sur muestra un ligero descenso, posiblemente vinculado a diferencias en las condiciones meteorológicas. Este leve descenso podría estar asociado a variaciones dentro del mismo clúster, dado que este agrupa puntos ubicados tanto en zonas oceánicas como en áreas no oceánicas del sur de España. La variabilidad climática regional identificada entre zonas destaca la influencia de factores meteorológicos específicos en el comportamiento térmico.

Los resultados del capítulo 5 coinciden con estos hallazgos, al identificar incrementos pronunciados de las temperaturas máximas, especialmente en las regiones montañosas y en el centro de España. Sin embargo, las regiones meridionales muestran aumentos más moderados, lo que subraya la disparidad espacial en las tendencias de temperatura.

El capítulo 4 complementa estos estudios al destacar el crecimiento de las temperaturas máximas extremas en Andalucía oriental y el sur de Murcia, regiones previamente señaladas por su alta variabilidad.

Cambios en las precipitaciones:

Tanto el capítulo 4 como el capítulo 5 subrayan las tendencias a la disminución de las precipitaciones en las regiones del sur y del interior de España, lo que concuerda con los hallazgos globales más amplios sobre el aumento de la aridez en los climas mediterráneos. Los aumentos detectados de la duración de la sequía y de los días secos consecutivos, especialmente en el sur de España, generan preocupación por la gestión de los recursos hídricos, la sostenibilidad agrícola y los impactos ecológicos.

Componentes estacionales y cíclicos:

En esta investigación se destaca de manera sistemática la importancia de las variaciones estacionales y cíclicas. En el capítulo 3 se incorporan componentes estacionales para agrupar regiones, mientras que en el capítulo 4 se demuestra cómo el análisis multiescala descubre patrones cíclicos de sequía. Por lo que estos resultados enriquecen la comprensión del funcionamiento de la variabilidad climática a corto y largo plazo.

Influencias geográficas:

En el capítulo 5 se incorpora de forma explícita el papel de las variables geográficas, mostrando cómo factores como la latitud, longitud, altura y la distancia al mar configuran los impactos locales del cambio climático. Todo esto ayuda a contextualizar la heterogeneidad espacial observada en los estudios de los otros dos capítulos.

- **Avances en la regionalización y la investigación climática**

Los procedimientos propuestos en esta tesis suponen un avance significativo respecto a investigaciones previas sobre regionalización climática, tanto por las metodologías aplicadas como por la calidad de los resultados obtenidos.

Incorporación de metodologías avanzadas:

La utilización de técnicas como la Descomposición en Componentes Singulares (SSA), el análisis multiescala y el Análisis Discriminante de Mínimos Cuadrados Parciales (PLS-DA) representa una evolución en el rigor metodológico, ofreciendo resultados de agrupación más sólidos y matizados. Lo que permite descomponer y modelizar series temporales de forma que capten patrones subyacentes que a menudo se pasan por alto en métodos más simples.

Más allá de las métricas tradicionales:

A diferencia de los estudios climáticos tradicionales que a menudo se basan en variables estáticas como las tendencias medias, nuestros enfoques hacen hincapié en componentes dinámicos (por ejemplo, patrones estacionales, tendencias cíclicas, factores geográficos), permitiendo una visión más profunda de los factores que

influyen en la variabilidad. Estos cambios proporcionan una mejora para la representación de los climas regionales, haciendo posibles estrategias de adaptación climática más precisas y procesables.

Agrupaciones más detalladas:

Los análisis sobre zonas climáticas amplias pueden ocultar variaciones locales. Las agrupaciones detalladas que se identificaron en nuestras investigaciones, las 12 zonas climáticas de España encontradas en el capítulo 4 captan diferencias microclimáticas. Las zonas adyacentes pueden diferir en gran medida debido a la topografía o a la proximidad del mar, como se mostró en el capítulo 5. Los resultados de agrupaciones refinadas son más útiles para las políticas climáticas localizadas.

Integración de múltiples variables:

La integración de diferentes variables como las temperaturas, precipitaciones y factores geográficos hace que la regionalización sea eficaz, toda vez que permite que el enfoque tenga una interconexión de los procesos climáticos, al proporcionar una comprensión multidimensional del sistema.

- **Implicaciones para las investigaciones climáticas en general**

Los resultados de esta tesis trascienden el ámbito geográfico de la Península Ibérica y pueden aportar valor a la investigación climática global

Valor para los modelos predictivos:

Los procedimientos presentados en estos estudios, en particular la inclusión de variables geográficas y la integración de múltiples escalas para múltiples variables en los procesos de agrupamientos y regionalización, pueden resultar fácilmente adaptables a otras regiones del mundo. Su aplicación puede contribuir a mejorar las predicciones del cambio climático y sus impactos.

Perspectivas de climas similares:

Al ser parte de una zona mediterránea, las tendencias de España permiten indicadores valiosos para regiones comparables que están enfrentadas a situaciones de calentamiento, de aridez y a los patrones climáticos extremos.

Aplicación en políticas de adaptación climática:

Con el refinamiento de las regionalizaciones, junto con la atención a fenómenos extremos, permite generar información clave para el diseño de políticas de adaptación eficaces. Ámbitos como la agricultura, gestión hídrica y planificación urbana, pueden beneficiarse de estos hallazgos, tanto dentro de las zonas de estudio como en regiones expuestas a riesgos climáticos comparables.

En síntesis, esta tesis no sólo mejora en la comprensión de los patrones climáticos en España, sino que también establece nuevos referentes metodológicos en el campo de la regionalización climática. Al ofrecer una perspectiva multidimensional y localizada, contribuye a llenar lagunas existentes en la literatura y sienta las bases

para estrategias de adaptación más eficaces frente a los desafíos del cambio climático global.

Capítulo 7

Capítulo 7

Síntesis, Conclusiones y Futuras Líneas de Investigación

Resumen: *En esta sección se presenta un resumen de la tesis junto con las principales conclusiones extraídas de nuestras investigaciones y, además, se establecen futuras líneas de investigación que surgieron durante el desarrollo de la tesis.*

7.1. Síntesis

Con las modificaciones de los fenómenos meteorológicos e hidrológicos extremos (como el calentamiento global y las sequias) y los consiguientes daños sobre la sociedad y los ecosistemas, es crucial desarrollar estrategias y procedimientos para investigar los patrones del clima y sus cambios. A partir de las evidencias de las variaciones espaciales y temporales que se presentan naturalmente en el clima de los territorios, emerge desarrollar estrategias que tengan en cuenta un marco fiable para identificar y analizar variaciones regionales del clima y sus cambios. Es posible que los comportamientos climáticos únicos de regiones distintas queden empañados en los análisis agregados que utilizan los enfoques tradicionales al analizar los patrones climáticos. Por tanto, en este punto, resulta interesante y pertinente la cuestión sobre *cómo desarrollar estrategias que permitan identificar los patrones del clima y sus cambios en un territorio determinado, permitiendo establecer diferencias climáticas subregionales al considerar procesos univariantes o multivariantes, en una o múltiples escalas e identificando el posible impacto de factores no climáticos*.

A continuación, en este capítulo, se recoge una síntesis de la tesis, acompañada de un breve análisis de las contribuciones más destacadas realizadas mediante el desarrollo de esta investigación.

Descubrir patrones espaciotemporales interesantes sobre el clima y sus cambios en un territorio determinado es una tarea importante para las investigaciones que involucran el clima en general, y particularmente para los meteorólogos, ya que contribuye de manera significativa a establecer tendencias y anomalías climáticas específicas y a mejorar la predicción y la previsión. No obstante, los procesos naturales y la influencia de factores físicos como la topografía, la distribución de la vegetación, las urbanizaciones y las masas de agua, que afectan el clima de la superficie (Chen et al., 2006; Teodoro et al., 2021), conducen a que estos patrones varíen considerablemente a escala regional dentro de un territorio. Y si se tienen en cuenta los retos que supone el cambio climático, tales patrones, diferenciables espacial y temporalmente, se hacen aún más inusuales e inesperados, haciendo que la tarea de identificarlos sea compleja, suponiendo un reto. Bajo la concepción de estos retos, las investigaciones sobre el clima ponen de manifiesto que, dentro de un territorio determinado, el clima y sus cambios pueden variar de una región a otra, junto con alguna influencia de variación temporal. En gran medida, esta tesis fue motivada por esta constatación y por la pretensión inherente de generar procedimientos eficientes capaces de regionalizar el cambio climático de un territorio determinado.

A continuación, se exponen algunas de las observaciones que se derivan de esta investigación:

1. En el capítulo 3 se introdujo un nuevo método para identificar patrones climáticos considerando procesos definidos por series temporales univariantes, en la sección titulada “Agrupación de series temporales utilizando componentes de tendencia, estacionalidad y autorregresivos

para identificar patrones de temperaturas máximas en la Península Ibérica", donde se empleó el SSA como técnica de preprocesamiento principal para descomponer las series de temperatura máxima y reducir su dimensionalidad.

Las principales observaciones que se desprenden del estudio son las siguientes:

- La investigación sobre los cambios del clima empleando SSA confirma que las temperaturas máximas en la Península Ibérica muestran variabilidad espacial y temporal, y ha permitido establecer diferencias regionales climáticas características entre zonas.
- Los parámetros de tendencia, estacionalidad y ruido incluidos en los vectores de características han permitido caracterizar zonas climáticas de manera eficiente, pues la presencia de estos componentes ha permitido que mediante este enfoque fuera posible identificar patrones de cambio de temperaturas asociados con el gradiente climático del área de estudio.
- Mediante el análisis de los datos se identificaron distintos patrones de cambio en la temperatura máxima de la Península Ibérica desde 1931 hasta 1956. El estudio encontró tres zonas diferenciadas: las zonas norte y central que mostraron un aumento de la temperatura en el tiempo, mientras que la zona sur exhibió una ligera disminución.
- El procedimiento de agrupamiento propuesto aporta precisión a los resultados de la agrupación al ser capaz de lidiar con el ruido de las series, pues al emplear el SSA para extraer las componentes de cada serie temporal original, aprovecha una ventaja en la capacidad de SSA para eliminar el ruido de las series, una de sus principales aplicaciones, junto con el estudio del perfil espectral.

Las observaciones anteriores han resultado interesantes y han forjado la idea de que la regionalización del cambio climático mediante la agrupación de series temporales promete una identificación eficiente de patrones interesantes de cambio climático que pueden quedar ocultos al analizar el sistema mediante análisis agregados. Sin embargo, los procesos físicos que influyen en las variables climáticas a menudo operan en un amplio rango de escalas de tiempo, por lo que estudiar la posibilidad de analizar el sistema considerando este aspecto, ayudaría a mejorar los resultados. Además, al considerar los procesos multivariantes del clima, la regionalización puede mejorar su precisión. De ahí, que el siguiente reto fue determinar *cómo identificar patrones climáticos diferenciados regionalmente en múltiples escalas y en múltiples procesos*.

2. Para responder a la cuestión anterior, se ha propuesto un procedimiento de agrupamiento basado en análisis multiescala de series temporales en el capítulo denominado "Identificación de patrones extremos de temperatura y precipitación en España a partir del análisis multiescalar de series temporales (Capitulo 4)" que tiene en cuenta características de series temporales multivariantes definidas en múltiples escalas de tiempo

para agrupar las series temporales e identificar los patrones climáticos regionales.

Para llevar a cabo este estudio, se adaptó al ámbito climático la metodología propuesta a partir del estudio de análisis multiescala de series financieras propuesto por Shi et al. (2021). Esta adaptación incluye: (1) la consideración de la asimetría de las series de precipitación, (2) la extensión al análisis multidimensional para obtener una mejor representación de la realidad climática, y (3) la modificación de la distancia entre series univariadas a series multivariadas lo que mejora la capacidad de capturar la complejidad del cambio climático.

Los resultados importantes del estudio son los siguientes:

- El método propuesto ha demostrado que la inclusión de análisis teniendo en cuenta diferentes escalas de tiempo y considerando series temporales multivariantes, permite mejorar de manera significativa la identificación de patrones climáticos regionales, mucho más específicos e informativos.
- Una fortaleza clave de este enfoque radica en su capacidad para incorporar múltiples escalas temporales, lo que permite la identificación de tendencias graduales a largo plazo y cambios climáticos abruptos a corto plazo.
- La investigación puso de manifiesto la presencia de patrones cíclicos y tendencias al considerar características de las series temporales climáticas definidas en diferentes escalas temporales, lo que hace que esta metodología sea muy útil para detectar diferencias entre la evolución de microclimas en pequeñas áreas territoriales vecinas.
- Los resultados de la agrupación dividieron a España en doce zonas distintas, cada una con un patrón único de temperaturas (máximas y mínimas) y de precipitaciones. Entre otros aspectos, se observó un crecimiento más marcado de las temperaturas máximas en áreas que abarcan Andalucía Oriental y el Sur de Murcia, así como una reducción de los eventos de frío extremo, principalmente en zonas montañosas y humedales atlánticos.
- En la reducción de la dimensionalidad, además de incorporar información de las series temporales determinadas en múltiples escalas de tiempo haciendo especial esfuerzo en considerar el orden temporal, se tiene en cuenta la distribución de las variables a la hora de definir las características. Por ejemplo, dada la asimetría de la distribución de las temperaturas y de las precipitaciones, se consideró más factible tener en cuenta medianas en lugar de medias para la construcción de vectores de características, ya que el estimador de la mediana refleja de manera más precisa el comportamiento de estas variables climáticas.

Este estudio pone de manifiesto la importancia de incorporar el análisis multiescala en la agrupación para regionalizar el cambio climático de manera efectiva. Hasta este punto, se han considerado múltiples variables estrictamente climáticas. Pero si consideramos que hay factores físicos definidos por variables no climáticas, como

por ejemplo algunas delimitaciones geográficas, que determinan el clima y sus variaciones, valdría la pena investigar si la inclusión de estas variables aportaría a mejorar los resultados de la regionalización climática.

3. Para responder a la cuestión anterior, en el capítulo 5, se propone un procedimiento de regionalización climática que además de incorporar indicadores cuantificables del cambio climático por cada serie temporal meteorológica como insumo para la generación de clústeres, tiene en cuenta de manera estratégica algunas variables geográficas para optimizar las agrupaciones finales. Este enfoque titulado "Cuantificación del impacto de las variables geográficas en los patrones de cambio climático en España mediante agrupamiento de series temporales" además permite cuantificar el impacto de las variables geográficas en el cambio climático de un territorio mediante un proceso de discriminación.

Los principales hallazgos revelan que:

- En este enfoque se definió una métrica de distancia que permite considerar el nivel de importancia de las variables o series temporales consideradas en el proceso como un primer paso para optimizar las agrupaciones finales. Este enfoque facilitó la inclusión coherente de indicadores cuantificables del cambio climático dentro del modelo de clustering.

- Al asignar diferentes combinaciones de pesos a los indicadores cuantificables del cambio climático, fue posible generar varias configuraciones de clustering con base en diferentes matrices de distancias y algoritmos de agrupamiento. Esto permitió incorporar de manera eficiente las variables geográficas mediante modelos PLS-DA y construir un índice final para optimizar la agrupación equilibrando las principales métricas del modelo.

- La metodología propuesta vincula de manera efectiva las tendencias espaciales y temporales del cambio climático con las variables geográficas, lo que conduce a clasificaciones regionales más refinadas. Los resultados de la aplicación empírica revelaron diferencias regionales significativas en los patrones de cambio climático, y permitió identificar siete zonas diferenciadas en la España Peninsular empleando series temporales del periodo 1951-2021.

- El modelo permitió identificar de forma clara la influencia de las variables geográficas sobre el clima y sus cambios al dejar ver que la latitud, la longitud y la altura desempeñan un papel fundamental en la determinación de las tendencias climáticas regionales.

- Este estudio permitió constatar que las temperaturas máximas han aumentado más en las regiones montañosas y en las zonas centrales de España, mientras que los menores incrementos se produjeron en el sur de España. Se observó un descenso de las precipitaciones, con las reducciones más pronunciadas en las regiones meridionales y del interior. Además, se

produjo un notable aumento de los días secos consecutivos, sobre todo en el sur.

En conjunto este enfoque metodológico permitió optimizar las agrupaciones finales mediante la vinculación de indicadores cuantificables del cambio climático en la agrupación y de variables no climáticas (geográficas) que muestran una influencia sobre el clima y sus cambios, con lo cual, la regionalización climática y los patrones identificados son más fieles a la realidad.

7.2. Conclusiones

La presente tesis, titulada "Nueva Metodología para Identificar Patrones de Cambio Climático mediante Análisis Clúster de Series Temporales con Reducción de la Dimensionalidad" ofrece contribuciones significativas tanto en el campo de la investigación climática como en el desarrollo metodológico para el análisis de series temporales. Las principales aportaciones pueden agruparse en dos bloques: conclusiones metodológicas y conclusiones climáticas.

Conclusiones metodológicas

1. El uso del SSA para la descomposición y agrupación de las series temporales resultó eficaz para capturar los componentes de tendencia, estacionalidad y de ruido, lo que ha permitido una comprensión más detallada de la dinámica regional de la temperatura.

2. La investigación define nuevas métricas de distancias para medir la similitud entre series temporales, permitiendo ponderar la importancia relativa de las variables, características o parámetros en los procesos de agrupamiento, mejorando así la sensibilidad del análisis a las particularidades de cada serie.

3. La integración de múltiples escalas de tiempo para diferentes variables en los procesos de agrupamiento permitió detectar patrones cíclicos y tendencias, poniendo de relieve la variabilidad espacial de los fenómenos climáticos extremos.

4. La incorporación del PLS-DA en la optimización de los resultados del agrupamiento ha facilitado la interpretación de las agrupaciones, al vincularlas con factores geográficos y climáticos, y ha servido como herramienta para optimizar la validez y robustez de los resultados.

5. Se ha desarrollado un nuevo índice que combina técnicas de aprendizaje supervisado y no supervisado, orientado a optimizar los resultados del clustering.

6. A diferencia de la gran mayoría de estudios previos, esta tesis aplica los métodos propuestos a un conjunto muy amplio de series temporales (más de 15.000) con alta resolución espacial (cuadriculas de hasta $5\times5 km^2$), lo cual refuerza la robustez de los resultados y permite detectar variaciones climáticas locales con mayor detalle.

Conclusiones climáticas

1. Este estudio resalta la heterogeneidad espacial de las manifestaciones del cambio climático, con patrones regionales diferenciados tanto en las tendencias térmicas como en las precipitaciones.
2. El enfoque adoptado proporciona claridad en la identificación y comprensión de los patrones regionales y sus variaciones espaciotemporales en los sistemas climáticos.
3. Estudia los procesos climáticos tanto a nivel univariante como multivariante, considerando en algunos casos múltiples escalas temporales en las que pueden representarse las series temporales. Además, propone diversas técnicas y herramientas para su análisis.
4. La tesis motiva la incorporación de variables no climáticas, que pueden tener un impacto en los procesos del sistema, como estrategia para optimizar la regionalización.
5. El trabajo realizado subraya la necesidad crítica de estrategias de adaptación y mitigación específicas para cada región, puesto que las variaciones del clima difieren de manera considerable entre zonas.

Por un lado, las conclusiones generadas en esta tesis ponen de manifiesto la importancia de considerar las diferencias climáticas subregionales a la hora de estudiar los procesos del clima y sus cambios dentro de un territorio determinado, y por otro lado, contribuyen a una comprensión más profunda de la dinámica climática en España y la Península Ibérica, proporcionando valiosos conocimientos para los responsables políticos y los investigadores que trabajan en la resiliencia climática y la sostenibilidad. Los procedimientos de agrupación de series temporales con reducción de la dimensionalidad que aquí se han propuesto, resultarán de interés para la sociedad de científicos climatológicos y de modelización de procesos, dados los problemas a los que se enfrentan para comprender los patrones del clima y sus variaciones. Las distancias definidas y las técnicas empleadas ofrecen cierta flexibilidad y su aplicación puede ser sencilla dependiendo del contexto.

7.3. Futuras líneas de investigación

Mediante el desarrollo de esta tesis y su alcance, se han realizado algunas contribuciones innovadoras relacionadas con la identificación de patrones climáticos regionales, los métodos de representación de series temporales, las medidas de similitud entre series temporales y en general con la agrupación de series temporales, lo que supone algunos avances importantes, por un lado, en la regionalización climática y, por otro lado, en la minería de datos. Lo cual, durante el transcurso y a la fecha, ha permitido pensar en algunas líneas o investigaciones futuras de interés.

Con respecto a métodos de representación o de reducción de la dimensionalidad, e incluso quizás con respecto a las medidas de similitud, los enfoques se han empleado sobre conjuntos de datos en los que todas las observaciones corresponden a series

temporales con igual longitud, pero si tuviéramos un dataset en el que no todas las series temporales tuvieran igual longitud, sería interesante (1) *estudiar como representar conjuntos de series temporales de diferentes longitudes* y/o (2) *indagar si es factible establecer similitudes entre las series de diferentes longitudes y cómo medirla*.

Como se detalla en los capítulos 1 y 2, existen principalmente tres enfoques para agrupar series temporales, de los cuales en esta tesis hemos aplicado solo dos, conforme a la metodología propuesta, centrada en la reducción de la dimensionalidad. El enfoque no utilizado es el agrupamiento basado en datos brutos, donde las series se agrupan según la forma de sus trayectorias. Las revisiones indican que este método suele ser más adecuado para series de corta duración, principalmente debido a las exigencias de memoria y su sensibilidad al ruido. Considerando este enfoque como una posible línea de investigación futura, una alternativa sería segmentar series de larga duración conservando el orden temporal y, para cada segmento, comparar la forma siguiendo la metodología habitual. Si además se aplica un tratamiento adecuado al ruido, se podría (3) *explorar la posibilidad de establecer una similitud global, basada en la comparación de segmentos cortos dentro de series extensa*s.

Por otra parte, en esta investigación solo estudiamos el cambio del clima considerando precipitaciones y temperaturas extremas. Por tanto, sería interesante (4) *proponer enfoques incluyendo otras variables climáticas* como la humedad relativa, la radiación solar, la velocidad del viento o indicadores de evapotranspiración, lo que permitiría enriquecer los modelos de agrupamiento y ofrecer una visión más integral de los sistemas climáticos

Finalmente, como se notó en el capítulo 5, al comparar las configuraciones de clustering, las configuraciones de clustering generadas empleando HA mostraron mejor rendimiento según los índices de validación empleados, pero hubo otras configuraciones que mejoraron la varianza explicada de los modelos PLS-DA, por lo que sería interesante (5) *explorar la aplicabilidad de otros índices de validación en el enfoque planteado con el fin de obtener una mayor optimización del método planteado*. De igual forma, si vamos a las distancias propuestas en los capítulos 4 y 5, podríamos pensar en (6) *establecer otros criterios para la asignación de pesos*, como, por ejemplo, no de acuerdo con las variables sino con base en los tipos de parámetros o combinaciones de ambos.

Anexo

Anexo I

Resúmenes de ponencias presentadas en congresos internacionales

Resumen: Aquí se presentan algunos de los resúmenes de contribuciones presentadas en simposios y congresos internacionales.

Publications of the Institute of Geophysics, Polish Academy of Sciences
International Symposium on Drought and Climate Change
November 24-25, 2022

Time series clustering using trend, seasonal and autoregressive components: patterns of change of maximum temperature in Iberian Peninsula

Arnobio PALACIOS GUTIERREZ[1,2] ✉,
and Jose Luis VALENCIA DELFA[1]

[1]Complutense University of Madrid, Faculty of Statistical Studies, Madrid, Community of Madrid

[2]Technological University of Chocó, Group Valoración y Aprovechamiento de la Biodiversidad, Quibdó, Chocó-Colombia

✉ arnobiop@ucm.es

ABSTRACT

Time series clustering is an important field of data mining and can be used to identify interesting patterns. This study introduces a new way to obtain clusters of time series by representing them with feature vectors that define the trend, seasonality and noise components of each series, in order to identify areas of the Iberian Peninsula that follow the same pattern of change in their maximum temperature during 1931-2009. Singular spectrum analysis decomposition in a sequential manner is used for dimensionality reduction, which allows the extraction of the trend, seasonality and residual components of each time series corresponding to an area of the Iberian region; then, the feature vectors of the time series are obtained by modelling the extracted components and estimating the parameters. Finally, the series are clustered using a clustering algorithm, and the clusters are defined according to the centroids. The results identified three differentiated zones, allowing to describe how the maximum temperature varied: in the north and central zones, an increase in temperature was noted over time, and in the south, a slight decrease, moreover different seasonal variations were noted according to zones.

KEYWORDS

Clustering; Maximum temperature time series; Singular spectrum analysis; Feature vectors of time series; Iberian Peninsula

EGU General Assembly 2023

Vienna, Austria & Online | 23–28 April 2023

Abstract EGU23-4466

EGU23-4466
https://doi.org/10.5194/egusphere-egu23-4466
EGU General Assembly 2023

Identification of precipitation and maximum temperature patterns in Spain during 1951 to 2021 using clustering based on multiscale analysis of time series

Arnobio Palacios Gutiérrez[1,2] and Jose Luis Valencia Delfa[1]

[1]Complutense University of Madrid, Faculty of Statistical Studies, Madrid, Community of Madrid, Spain
[2]Technological University of Chocó, Group Valoración y Aprovechamiento de la Biodiversidad, Quibdó, Chocó-Colombia

Spain is a territory with spatial variations highly influenced by its distance from the sea and its complex orography, where it is possible to note an uneven distribution of both temperature and precipitation. This study presents an analysis of trends in maximum temperature and precipitation by zone over the period 1951-2021 using monthly data. The database used includes 16156 multivariate time series (maximum temperatures and precipitation) corresponding to different areas of the Spanish territory, distributed over a grid of 5x5km^2. The methodology used starts by reducing the dimensionality of the time series and with this version are clustered using an approach based on multiscale analysis using a clustering algorithm. In the following, the prototypes of each group are defined, which allows to identify and analyse patterns of change in maximum temperatures and precipitation by zones. An increase in average maximum temperature has been identified in eight zones distributed in Spain from 1951 to 2021. The rate of change of maximum temperature was between 0.060ºC and 0.2155ºC per decade. Areas further south showed a higher rate of increase than areas found in the north. It has been observed that May was the month with the highest variation for all areas in maximum temperature, nevertheless, differences in seasonal variation are evident when passing from one zone to other, as in some there is greater variation in spring months and in others in winter months. An analysis of trends and seasonal variations of precipitation in the identified zones will be carried out and the correlation between patterns of maximum temperature and precipitation will be studied in each of the eight zones.

How to cite: Palacios Gutiérrez, A. and Valencia Delfa, J. L.: Identification of precipitation and maximum temperature patterns in Spain during 1951 to 2021 using clustering based on multiscale analysis of time series, EGU General Assembly 2023, Vienna, Austria, 24–28 Apr 2023, EGU23-4466, https://doi.org/10.5194/egusphere-egu23-4466, 2023.

Bibliografía

Abbasi, F., Bazgeer, S., Kalehbasti, P. R., Oskoue, E. A., Haghighat, M., & Kalehbasti, P. R. (2022). New climatic zones in Iran: a comparative study of different empirical methods and clustering technique. *Theoretical and Applied Climatology*, *147*(1–2), 47–61. https://doi.org/10.1007/s00704-021-03785-9

Abu Alfeilat, H. A., Hassanat, A. B. A., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. B. S. (2019). Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data*, *7*(4), 221–248. https://doi.org/10.1089/big.2018.0175

Agarwal, A., Maheswaran, R., Marwan, N., Caesar, L., & Kurths, J. (2018). Wavelet-based multiscale similarity measure for complex networks. The European Physical Journal B, 91(11), 296. https://doi.org/10.1140/epjb/e2018-90460-6

Agarwal, A., Maheswaran, R., Sehgal, V., Khosa, R., Sivakumar, B., & Bernhofer, C. (2016). Hydrologic regionalization using wavelet-based multiscale entropy method. *Journal of Hydrology*, *538*, 22–32. https://doi.org/10.1016/j.jhydrol.2016.03.023

Agarwal, A., Marwan, N., Rathinasamy, M., Merz, B., & Kurths, J. (2017). Multi-scale event synchronization analysis for unravelling climate processes: a wavelet-based approach. Nonlinear Processes in Geophysics, 24(4), 599–611. https://doi.org/10.5194/npg-24-599-2017

Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering - A decade review. *Information Systems*, *53*, 16–38. https://doi.org/10.1016/j.is.2015.04.007

Ahmad, A., & Khan, S. S. (2019). Survey of State-of-the-Art Mixed Data Clustering Algorithms. *IEEE Access*, *7*, 31883–31902. https://doi.org/10.1109/ACCESS.2019.2903568

Alqahtani, A., Ali, M., Xie, X., & Jones, M. W. (2021). Deep Time-Series Clustering: A Review. *Electronics*, *10*(23), 3001. https://doi.org/10.3390/electronics10233001

Alsalal, S., Tan, M. L., Samat, N., Al-Bakri, J. T., & Zhang, F. (2024). Temperature and precipitation changes under CMIP6 projections in the Mujib Basin, Jordan. *Theoretical and Applied Climatology*. https://doi.org/10.1007/s00704-024-05087-2

Bagnall, A., & Janacek, G. (2005). Clustering time series with clipped data. *Machine Learning*, *58*(2–3), 151–178. https://doi.org/10.1007/s10994-005-5825-6

Baker, F. B., & Hubert, L. J. (1975). Measuring the Power of Hierarchical Cluster Analysis. In *Source: Journal of the American Statistical Association* (Vol. 70, Issue 349).

Banerjee, A., & Davé, R. N. (2004). Validating clusters using the Hopkins statistic. *IEEE International Conference on Fuzzy Systems*, *1*, 149–153. https://doi.org/10.1109/FUZZY.2004.1375706

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms* (Springer Science & Business Media, Ed.). Springer US. https://doi.org/10.1007/978-1-4757-0450-1

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control* (John Wiley & Sons, Ed.; 5th ed.). https://books.google.es/books?id=rNt5CgAAQBAJ&dq=%22time+series+analysis+forecasting+and+control%22&lr=&hl=es&source=gbs_navlinks_s

Bro, R., & Smilde, A. K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics*, *17*(1), 16–33. https://doi.org/10.1002/cem.773

Brown, M. E., Antle, J. M., Backlund, P., Carr, E. R., Easterling, W. E., Walsh, M. K., Ammann, C., Attavanich, W., Barrett, C. B., Bellemare, M. F., Dancheck, V., Funk, C., Grace, K., Ingram, J. S. I., Jiang, H., Maletta, H., Mata, T., Murray, A., Ngugi, M., … Tebaldi, C. (2015). *Climate Change, Global Food Security, and the U.S. Food System*. https://doi.org/10.7930/J0862DC7

Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., & Dougherty, E. R. (2007). Model-based evaluation of clustering validation measures. *Pattern Recognition*, *40*(3), 807–824. https://doi.org/10.1016/j.patcog.2006.06.026

Brunner, L., McSweeney, C., Ballinger, A. P., Befort, D. J., Benassi, M., Booth, B., Coppola, E., de Vries, H., Harris, G., Hegerl, G. C., Knutti, R., Lenderink, G., Lowe, J., Nogherotto, R., O'Reilly, C., Qasmi, S., Ribes, A., Stocchi, P., & Undorf, S. (2020). Comparing Methods to Constrain Future European Climate Projections Using a Consistent Framework. *Journal of Climate*, *33*(20), 8671–8692. https://doi.org/10.1175/JCLI-D-19-0953.1

Calheiros, T., Pereira, M. G., & Nunes, J. P. (2021). Assessing impacts of future climate change on extreme fire weather and pyro-regions in Iberian Peninsula. *Science of The Total Environment*, *754*, 142233. https://doi.org/10.1016/j.scitotenv.2020.142233

Caruso, G., Gattone, S. A., Fortuna, F., & Di Battista, T. (2018). Cluster analysis as a decision-making tool: A methodological review. *Advances in Intelligent Systems and Computing*, *618*, 48–55. https://doi.org/10.1007/978-3-319-60882-2_6

Cha, S. (2007). Comprehensive Survey on Distance / Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, *1*(4).

Chen, X. L., Zhao, H. M., Li, P. X., & Yin, Z. Y. (2006). Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. *Remote Sensing of Environment*, *104*(2), 133–146. https://doi.org/10.1016/j.rse.2005.11.016

Choularton, R., Krishnamurthy, P. K., & Lewis, K. (2012). *Climate impacts on food security and nutrition - A Review of Existing Knowledge*.

Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogée, J., Allard, V., Aubinet, M., Buchmann, N., Bernhofer, C., Carrara, A., Chevallier, F., De Noblet, N., Friend, A. D., Friedlingstein, P., Grünwald, T., Heinesch, B., Keronen, P., Knohl, A., Krinner, G., … Valentini, R. (2005). Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature*, *437*(7058), 529–533. https://doi.org/10.1038/nature03972

Cross, G. R., & Jain, A. K. (1982). Measurement of Clustering Tendency. *IFAC Proceedings Volumes*, *15*(1), 315–320. https://doi.org/10.1016/s1474-6670(17)63365-2

CWang, C., & XWang, X. S. (2000). Supporting content-based searches on time series via approximation. *Proceedings of the International Conference on Scientific and Statistical Database Management, SSDBM*, 69–81. https://doi.org/10.1109/ssdm.2000.869779

Dasari, H. P., Pozo, I., Ferri-Yáñez, F., & Araújo, M. B. (2014). A Regional Climate Study of Heat Waves over the Iberian Peninsula. *Atmospheric and Climate Sciences*, *04*(05), 841–853. https://doi.org/10.4236/acs.2014.45074

de Lucena, A. F. P., Szklo, A. S., Schaeffer, R., de Souza, R. R., Borba, B. S. M. C., da Costa, I. V. L., Júnior, A. O. P., & da Cunha, S. H. F. (2009). The vulnerability of renewable energy to climate change in Brazil. *Energy Policy*, *37*(3), 879–889. https://doi.org/10.1016/j.enpol.2008.10.029

del Río, S., Cano-Ortiz, A., Herrero, L., & Penas, A. (2012). Recent trends in mean maximum and minimum air temperatures over Spain (1961-2006). *Theoretical and Applied Climatology*, *109*(3–4), 605–626. https://doi.org/10.1007/s00704-012-0593-2

Desgraupes, B. (2017). *clusterCrit: Clustering Indices*. https://CRAN.R-project.org/package=clusterCrit

Diday, E., & Simon, J. C. (1976). Clustering Analysis. In *In: Fu, K.S. (eds) Digital Pattern Recognition. Communication and Cybernetics: Vol. 10. Springer* (pp. 47–94). https://doi.org/10.1007/978-3-642-96303-2_3

Dosio, A., Mentaschi, L., Fischer, E. M., & Wyser, K. (2018). Extreme heat waves under 1.5 °C and 2 °C global warming. *Environmental Research Letters*, *13*(5), 054006. https://doi.org/10.1088/1748-9326/aab827

Dunn, J. C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, *4*(1), 95–104. https://doi.org/10.1080/01969727408546059

Ergüner Özkoç, E. (2021). Clustering of Time-Series Data. In D. Birant (Ed.), *Data Mining - Methods, Applications and Systems*. https://doi.org/10.5772/intechopen.84490

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743. https://doi.org/10.1016/j.engappai.2022.104743

Fonseca, D., Carvalho, M. J., Marta-Almeida, M., Melo-Gonçalves, P., & Rocha, A. (2016). Recent trends of extreme temperature indices for the Iberian Peninsula. *Physics and Chemistry of the Earth*, *94*, 66–76. https://doi.org/10.1016/j.pce.2015.12.005

Furió, D., & Meneu, V. (2011). Analysis of extreme temperatures for four sites across Peninsular Spain. *Theoretical and Applied Climatology*, *104*(1–2), 83–99. https://doi.org/10.1007/s00704-010-0324-5

Gallant, A. R., & Goebel, J. J. (1976). Nonlinear Regression with Autocorrelated Errors. *Journal of the American Statistical Association*, *71*(356), 961–967. https://doi.org/10.2307/2286869

Gao, J., & Shang, P. (2019). Analysis of complex time series based on EMD energy entropy plane. *Nonlinear Dynamics*, *96*(1), 465–482. https://doi.org/10.1007/s11071-019-04800-5

García, D. H. (2022). Analysis of Urban Heat Island and Heat Waves Using Sentinel-3 Images: a Study of Andalusian Cities in Spain. *Earth Systems and Environment*, *6*(1), 199–219. https://doi.org/10.1007/s41748-021-00268-9

García Díaz, J. C. (2016). *Predicción en el dominio del tiempo. Análisis de series temporales para ingenieros* (Editorial Universitat Politècnica de València, Ed.). http://hdl.handle.net/10251/72938

Gebremichael, H. B., Raba, G. A., Beketie, K. T., Feyisa, G. L., & Siyoum, T. (2022). Changes in daily rainfall and temperature extremes of upper Awash Basin, Ethiopia. *Scientific African*, *16*, e01173. https://doi.org/10.1016/j.sciaf.2022.e01173

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, *131*(1–2), 59–95. https://doi.org/10.1016/j.jeconom.2005.01.004

Golyandina, N. (2010). On the choice of parameters in Singular Spectrum Analysis and related subspace-based methods. In *Statistics and Its Interface* (Vol. 3).

Golyandina, N., & Korobeynikov, A. (2014). Basic Singular Spectrum Analysis and forecasting with R. *Computational Statistics and Data Analysis*, *71*, 934–954. https://doi.org/10.1016/j.csda.2013.04.009

Golyandina, N., Korobeynikov, A., & Zhigljavsky, A. (2018). Singular Spectrum Analysis with R. In *Springer*.

Golyandina, N., Nekrutkin, V., & Zhigljavsky, A. A. (2001). *Analysis of Time Series Structure: SSA and Related Techniques* (2001 CRC Press, Ed.; ilustrada). https://books.google.es/books?id=L0HjfJCIrNYC&lr=&hl=es&source=gbs_navlinks_s

Golyandina, N., Pepelyshev, A., & Steland, A. (2012). New approaches to nonparametric density estimation and selection of smoothing parameters. *Computational Statistics and Data Analysis*, *56*, 2206–2218. https://doi.org/10.1016/j.csda.2011.12.019

González Hidalgo, J. C., De Luís, M., Raventós, J., & Sánchez, J. R. (2003). Daily rainfall trend in the Valencia Region of Spain. *Theoretical and Applied Climatology*, *75*(1), 117–130. https://doi.org/10.1007/s00704-002-0718-0

González Velasco, M., & del Puerto García, I. M. (2009). *Series temporales* (Universidad de Extremadura, Ed.).

Granger, C. W. J. (1989). *Forecasting in Business and Economics* (Emerald Group Publishing Limited, Ed.; 2nd ed.).

Guo, C., Jia, H., & Zhang, N. (2008). Time series clustering based on ICA for stock data analysis. *2008 International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2008*, 1–4. https://doi.org/10.1109/WiCom.2008.2534

Gupta, A. K., Negi, M., Nandy, S., Alatalo, J. M., Singh, V., & Pandey, R. (2019). Assessing the vulnerability of socio-environmental systems to climate change along an altitude gradient in the Indian Himalayas. *Ecological Indicators*, *106*, 105512. https://doi.org/10.1016/j.ecolind.2019.105512

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. In Morgan Kaufmann (Ed.), *Journal of Chemical Information and Modeling* (Third Edit, Vol. 53, Issue 9). http://library.books24x7.com/toc.aspx?bkid=44712

Hansen, P., & Jaumard, B. (1997). Cluster analysis and mathematical programming. *Mathematical Programming 79 (1997) 191-215*, *79*, 191–215. https://doi.org/10.1080/01621459.1971.10482319

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

Hautamaki, V., Nykänen, P., & Fränti, P. (2008). Time-series clustering by approximate prototypes. *Proceedings - International Conference on Pattern Recognition*, 2–5. https://doi.org/10.1109/icpr.2008.4761105

Holder, C., Middlehurst, M., & Bagnall, A. (2024). A review and evaluation of elastic distance functions for time series clustering. *Knowledge and Information Systems*, *66*(2), 765–809. https://doi.org/10.1007/s10115-023-01952-0

Hsu, K.-C., & Li, S.-T. (2010). Clustering spatial–temporal precipitation data using wavelet transform and self-organizing map neural network. *Advances in Water Resources*, *33*(2), 190–200. https://doi.org/10.1016/j.advwatres.2009.11.005

Hubert, L., & Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology*, *29*(2), 190–241. https://doi.org/10.1111/j.2044-8317.1976.tb00714.x

Huhtala, Y., Karkkainen, J., & Toivonen, H. T. T. (1999). Mining for similarities in aligned time series using wavelets. *Data Mining and Knowledge Discovery: Theory, Tools, and Technology*, *3695*. https://doi.org/10.1117/12.339977

Huiting, L., Zhiwei, N., & Jianyang, L. (2006). Time series similar pattern matching based on empirical mode decomposition. *Proceedings - ISDA 2006: Sixth International Conference on Intelligent Systems Design and Applications*, *1*(050460402), 644–648. https://doi.org/10.1109/ISDA.2006.273

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Ilbay-Yupa, M., Lavado-Casimiro, W., Rau, P., Zubieta, R., & Castillón, F. (2021). Updating regionalization of precipitation in Ecuador. *Theoretical and Applied Climatology*, *143*(3–4), 1513–1528. https://doi.org/10.1007/s00704-020-03476-x

Imam, Md. H., Rahman, Md. M., Roy, S., Hoque, F., Ahsan, U., Abdullah, Sk. Md. A., Hossain, Md. S., & Rahim, M. A. (2022). Analysis of Diurnal Air Temperature Range Variation over Bangladesh. *Earth Systems and Environment*, *6*(2), 361–373. https://doi.org/10.1007/s41748-021-00282-x

IPCC. (2014). *Climate change 2014: Mitigation of climate change. Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, IPCC.* Cambridge University Press. https://scholar.google.com/scholar_lookup?title=Climate%20change%202014&publication_year=2014&author=IPCC&author=K.P.%20R&author=A.M.%20L

IPCC. (2022). Annex I: Glossary [van Diemen, R., J.B.R. Matthews, V. Möller, J.S. Fuglestvedt, V. Masson-Delmotte, C. Méndez, A. Reisinger, S. Semenov (eds)]. In IPCC, 2022: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, J. Malley, (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA. doi: 10.1017/9781009157926.020

IPCC: Masson-Delmotte, V., Zhai, P., Chen, Y., Goldfarb, L., Gomis, M. I., Matthews, J. B. R., Berger, S., Huang, M., Yelekçi, O., Yu, R., Zhou, B., Lonnoy, E., Maycock, T. K., Waterfield, T., Leitzell, K., & Caud, N. (2021). *Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change Edited by*. Climate Change 2021: The Physical Science Basis. www.ipcc.ch

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

Kalpakis, K., Gada, D., & Puttagunta, V. (2001). Distance Measures for Effective Clustering of ARIMA Time-Series. *Actas 2001 IEEE International Conference on Data Mining*, 273–280. https://doi.org/10.1109 / ICDM.2001.989529

Kassambara, A. (2017). Multivariate Analysis I: Practical Guide To Cluster Analysis in R. Unsupervised Machine Learning. *Taylor & Francis Group*, 188.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis* (L. Kaufman & P. J. Rousseeuw, Eds.). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316801

Keogh, E., & Smyth, P. (1997). A probabilistic approach to fast pattern matching in time series databases. *Proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining*, *M* (1994), 52–57. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+Probabilistic+Approach+to+Fast+Pattern+Matching+in+Time+Series+Databases#0

Khan, A. A., Zhao, Y., khan, J., Rahman, G., Rafiq, M., & Moazzam, M. F. U. (2021). Spatial and Temporal Analysis of Rainfall and Drought Condition in Southwest Xinjiang in Northwest China, Using Various Climate Indices. *Earth Systems and Environment*, *5*(2), 201–216. https://doi.org/10.1007/s41748-021-00226-5

King, A. D., & Karoly, D. J. (2017). Climate extremes in Europe at 1.5 and 2 degrees of global warming. *Environmental Research Letters*, *12*(11), 114031. https://doi.org/10.1088/1748-9326/aa8e2c

Knutti, R. (2010). The end of model democracy? *Climatic Change*, *102*(3–4), 395–404. https://doi.org/10.1007/s10584-010-9800-2

KOHONEN, T., & OJA, E. (1996). *Engineering applications of the self-organizing map*. https://doi.org/10.1109/5.537105

Kolusu, S. R., Siderius, C., Todd, M. C., Bhave, A., Conway, D., James, R., Washington, R., Geressu, R., Harou, J. J., & Kashaigili, J. J. (2021). Sensitivity of projected climate impacts to climate model weighting: multi-sector analysis in eastern Africa. *Climatic Change*, *164*(3–4), 36. https://doi.org/10.1007/s10584-021-02991-8

Kraus, J. M., Müssel, C., Palm, G., & Kestler, H. A. (2011). Multi-objective selection for collecting cluster alternatives. *Computational Statistics*, *26*(2), 341–353. https://doi.org/10.1007/s00180-011-0244-6

Kremer, H., Kranen, P., Jansen, T., Seidl, T., Bifet, A., Holmes, G., & Pfahringer, B. (2011). An effective evaluation measure for clustering on evolving data streams. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 868–876. https://doi.org/10.1145/2020408.2020555

Kuglitsch, F. G., Toreti, A., Xoplaki, E., Della-Marta, P. M., Zerefos, C. S., Türkeş, M., & Luterbacher, J. (2010). Heat wave changes in the eastern Mediterranean since 1960. *Geophysical Research Letters*, *37*(4). https://doi.org/10.1029/2009GL041841

Kumar, R. P., & Nagabhushan, P. (2006). Time Series as a Point - A Novel Approach for Time Series Cluster Visualization. *Conference on Data Mining | DMIN'06 |*, *0*(3), 24–29.

Kuriqi, A., Ali, R., Pham, Q. B., Montenegro Gambini, J., Gupta, V., Malik, A., Linh, N. T. T., Joshi, Y., Anh, D. T., Nam, V. T., & Dong, X. (2020). Seasonality shift and streamflow

flow variability trends in central India. *Acta Geophysica*, *68*(5), 1461–1475. https://doi.org/10.1007/s11600-020-00475-4

Laepple, T., & Huybers, P. (2014). Ocean surface temperature variability: Large model–data differences at decadal and longer periods. *Proceedings of the National Academy of Sciences*, *111*(47), 16682–16687. https://doi.org/10.1073/pnas.1412077111

Lee, A. J. T., Lin, M. C., Kao, R. T., & Chen, K. T. (2010). An effective clustering approach to stock market prediction. *PACIS 2010 - 14th Pacific Asia Conference on Information Systems*, 345–354.

Lee, Y., Na, J., & Lee, W. B. (2018). Robust design of ambient-air vaporizer based on time-series clustering. *Computers and Chemical Engineering*, *118*, 236–247. https://doi.org/10.1016/j.compchemeng.2018.08.026

Li, H., & Wei, M. (2020). Fuzzy clustering based on feature weights for multivariate time series. *Knowledge-Based Systems*, *197*, 105907. https://doi.org/10.1016/j.knosys.2020.105907

Li, J., He, X., & Tao, L. (2022). Assessing multiscale variability and teleconnections of monthly precipitation in Yangtze River Basin based on multiscale information theory method. *Theoretical and Applied Climatology*, *147*(1–2), 717–735. https://doi.org/10.1007/s00704-021-03845-0

Liu, H., Zhan, Q., Yang, C., & Wang, J. (2019). The multi-timescale temporal patterns and dynamics of land surface temperature using Ensemble Empirical Mode Decomposition. *Science of The Total Environment*, *652*, 243–255. https://doi.org/10.1016/j.scitotenv.2018.10.252

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., & Wu, S. (2013). Understanding and enhancement of internal clustering validation measures. *IEEE Transactions on Cybernetics*, *43*(3), 982–994. https://doi.org/10.1109/TSMCB.2012.2220543

Lorenzo, M. N., & Alvarez, I. (2022). Future changes of hot extremes in Spain: towards warmer conditions. *Natural Hazards*, *113*(1), 383–402. https://doi.org/10.1007/s11069-022-05306-x

Lukasová, A. (1979). Hierarchical agglomerative clustering procedure. *Pattern Recognition*, *11*(5–6), 365–381. https://doi.org/10.1016/0031-3203(79)90049-9

Luna, M. Y., Morata, A., Luisa Martin, M., Santos-Muñoz, D., De, J., & Cruz, L. A. (2008). Validación de la base de datos reticular de la AEMET: temperatura diaria máxima y mínima. https://doi.org/http://hdl.handle.net/20.500.11765/8563

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In U. of C. P. Volume 1: Statistics (Ed.), *5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297).

Malede, D. A., Agumassie, T. A., Kosgei, J. R., Andualem, T. G., & Diallo, I. (2022). Recent Approaches to Climate Change Impacts on Hydrological Extremes in the Upper

Blue Nile Basin, Ethiopia. *Earth Systems and Environment*, *6*(3), 669–679. https://doi.org/10.1007/s41748-021-00287-6

Mauricio, J. (2007). *Introducción al Análisis de series temporales* (Universidad Complutence de Madrid, Ed.). https://www.ucm.es/data/cont/docs/518-2013-11-11-JAM-IAST-Libro.pdf

Mcclain, J. O., & Rao, V. R. (1975). CLUSTISZ: A Program to Test for the Quality of Clustering of a Set of Objects. In *Source: Journal of Marketing Research* (Vol. 12, Issue 4).

Meseguer-Ruiz, O., Lopez-Bustins, J. A., Arbiol-Roca, L., Martin-Vide, J., Miró, J., & Estrela, M. J. (2021). Temporal changes in extreme precipitation and exposure of tourism in Eastern and South-Eastern Spain. *Theoretical and Applied Climatology*, *144*(1–2), 379–390. https://doi.org/10.1007/s00704-021-03548-6

Młyński, D., Wałęga, A., & Kuriqi, A. (2021). Influence of meteorological drought on environmental flows in mountain catchments. *Ecological Indicators*, *133*, 108460. https://doi.org/10.1016/j.ecolind.2021.108460

Molina, M. O., Sánchez, E., & Gutiérrez, C. (2020). Future heat waves over the Mediterranean from an Euro-CORDEX regional climate model ensemble. *Scientific Reports*, *10*(1), 8801. https://doi.org/10.1038/s41598-020-65663-0

Möller-Levet, C. S., Klawonn, F., Cho, K., & Wolkenhauer, O. (2003). Fuzzy Clustering of Short Time-Series and Unevenly Distributed Sampling Points. *Advances in Intelligent Data Analysis*, *2810*, 330–340.

Moreno, M., Barea, R., Castro, L., Cagigas, D., Ortiz, R., & Ortiz, P. (2024). Climate Change monitoring with Art-Risk 5: New approach for environmental hazard assessment in Seville and Almería Historic Centres (Spain). *Procedia Structural Integrity*, *55*, 9–17. https://doi.org/10.1016/j.prostr.2024.02.002

Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, *31*(3), 274–295. https://doi.org/10.1007/s00357-014-9161-z

Oliver, M. A., & Webster, R. (1990). Kriging: a method of interpolation for geographical information systems. International Journal of Geographical Information Systems, 4(3), 313–332. https://doi.org/10.1080/02693799008941549

Olteanu, M., & Villa-Vialaneix, N. (2015). Using SOMbrero for clustering and visualizing graphs Titre: Utiliser SOMbrero pour la classification et la visualisation de graphes. In *Journal de la Société Française de Statistique* (Vol. 156, Issue 3). http://www.sfds.asso.fr/journal

O'neill, M. S., & Ebi, K. L. (2009). Temperature Extremes and Health: Impacts of Climate Variability and Change in the. *Source: Journal of Occupational and Environmental Medicine*, *51*(1), 13–25. https://doi.org/10.1097/JOM.0b013e318173el22

Palacios Gutiérrez, A., & Valencia Delfa, J. L. (2023, April 25). Identification of precipitation and maximum temperature patterns in Spain during 1951 to 2021 using clustering based on multiscale analysis of time series. *EGU General Assembly 2023*. https://doi.org/https://doi.org/10.5194/egusphere-egu23-4466, 2023

Palacios Gutiérrez, A., Valencia Delfa, J. L., & Villeta López, M. (2023). Time series clustering using trend, seasonal and autoregressive components to identify maximum temperature patterns in the Iberian Peninsula. *Environmental and Ecological Statistics*, *30*(3), 421–442. https://doi.org/10.1007/s10651-023-00572-9

Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, *36*(2), 3336–3341. https://doi.org/10.1016/J.ESWA.2008.01.039

Parracho, A. C., Melo-Gonçalves, P., & Rocha, A. (2016). Regionalisation of precipitation for the Iberian Peninsula and climate change. *Physics and Chemistry of the Earth, Parts A/B/C*, *94*, 146–154. https://doi.org/10.1016/J.PCE.2015.07.004

Patz, J. A., Campbell-Lendrum, D., Holloway, T., & Foley, J. A. (2005). Impact of regional climate change on human health. In *Nature* (Vol. 438, Issue 7066, pp. 310–317). Nature Publishing Group. https://doi.org/10.1038/nature04188

Peña-Angulo, D., Cortesi, N., Brunetti, M., & González-Hidalgo, J. C. (2015). Spatial variability of maximum and minimum monthly temperature in Spain during 1981–2010 evaluated by correlation decay distance (CDD). *Theoretical and Applied Climatology*, *122*(1–2), 35–45. https://doi.org/10.1007/s00704-014-1277-x

Ramos, M. C., Balasch, J. C., & Martínez-Casasnovas, J. A. (2012). Seasonal temperature and precipitation variability during the last 60 years in a Mediterranean climate area of Northeastern Spain: A multivariate analysis. *Theoretical and Applied Climatology*, *110*(1–2), 35–53. https://doi.org/10.1007/s00704-012-0608-z

Rani, S., & Sikka, G. (2012). Recent Techniques of Clustering of Time Series Data: A Survey. *International Journal of Computer Applications*, *52*(15), 1–9. https://doi.org/10.5120/8282-1278

Ratanamahatana, C., Keogh, E., Bagnall, A. J., & Lonardi, S. (2005). A novel bit level time series representation with implication of similarity search and clustering. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3518 LNAI*, 771–777. https://doi.org/10.1007/11430919_90

Ray, S., & Turi, R. H. (1999). Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation. *The 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, 137–143.

Romesburg, C. (2004). *Cluster Analysis for Researchers* (Lulu.com, Ed.; ilustrada). https://books.google.es/books?id=ZuIPv7OKm10C&hl=es&source=gbs_navlinks_s

Roushangar, K., & Alizadeh, F. (2018). A multiscale spatio-temporal framework to regionalize annual precipitation using k-means and self-organizing map technique. *Journal of Mountain Science*, *15*(7), 1481–1497. https://doi.org/10.1007/s11629-017-4684-5

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Ruspini, E. H. (1969). A new approach to clustering. *Information and Control*, *15*(1), 22–32. https://doi.org/10.1016/S0019-9958(69)90591-9

Russo, S., Sillmann, J., & Fischer, E. M. (2015). Top ten European heatwaves since 1950 and their occurrence in the coming decades. *Environmental Research Letters*, *10*(12), 124003. https://doi.org/10.1088/1748-9326/10/12/124003

Samantaray, A. K., Mitra, A., Ramadas, M., & Panda, R. K. (2021). Regionalization of hydroclimatic variables using Markov random field model for climate change impact assessment. *Journal of Hydrology*, *596*, 126071. https://doi.org/10.1016/J.JHYDROL.2021.126071

Samset, B. H., Fuglestvedt, J. S., & Lund, M. T. (2020). Delayed emergence of a global temperature response after emission mitigation. *Nature Communications*, *11*(1), 3261. https://doi.org/10.1038/s41467-020-17001-1

Sathaye, J. A., Dale, L. L., Larsen, P. H., Fitts, G. A., Koy, K., Lewis, S. M., & de Lucena, A. F. P. (2013). Estimating impacts of warming temperatures on California's electricity system. *Global Environmental Change*, *23*(2), 499–511. https://doi.org/10.1016/j.gloenvcha.2012.12.005

Saunders, K. R., Stephenson, A. G., & Karoly, D. J. (2021). A regionalisation approach for rainfall based on extremal dependence. *Extremes*, *24*(2), 215–240. https://doi.org/10.1007/s10687-020-00395-y

Schaeffer, R., Szklo, A. S., Pereira de Lucena, A. F., Moreira Cesar Borba, B. S., Pupo Nogueira, L. P., Fleming, F. P., Troccoli, A., Harrison, M., & Boulahya, M. S. (2012). Energy sector vulnerability to climate change: A review. In *Energy* (Vol. 38, Issue 1, pp. 1–12). Elsevier Ltd. https://doi.org/10.1016/j.energy.2011.11.056

Sehgal, V., Lakhanpal, A., Maheswaran, R., Khosa, R., & Sridhar, V. (2018). Application of multi-scale wavelet entropy and multi-resolution Volterra models for climatic downscaling. *Journal of Hydrology*, *556*, 1078–1095. https://doi.org/10.1016/j.jhydrol.2016.10.048

Senent-Aparicio, J., López-Ballesteros, A., Jimeno-Sáez, P., & Pérez-Sánchez, J. (2023). Recent precipitation trends in Peninsular Spain and implications for water infrastructure design. *Journal of Hydrology: Regional Studies*, *45*, 101308. https://doi.org/10.1016/j.ejrh.2022.101308

Sharghi, E., Nourani, V., Soleimani, S., & Sadikoglu, F. (2018). Application of different clustering approaches to hydroclimatological catchment regionalization in

mountainous regions, a case study in Utah State. *Journal of Mountain Science*, *15*(3), 461–484. https://doi.org/10.1007/s11629-017-4454-4

Shi, P., Sun, S., Gong, D., & Zhou, T. (2016). World Regionalization of Climate Change (1961–2010). *International Journal of Disaster Risk Science*, *7*(3), 216–226. https://doi.org/10.1007/s13753-016-0094-5

Shi, Y., Li, B., Du, G., & Dai, W. (2021). Clustering framework based on multi-scale analysis of intraday financial time series. *Physica A: Statistical Mechanics and Its Applications*, *567*(71932008). https://doi.org/10.1016/j.physa.2020.125728

Shin, Y., Lee, Y., & Park, J.-S. (2020). A Weighting Scheme in A Multi-Model Ensemble for Bias-Corrected Climate Simulation. *Atmosphere*, *11*(8), 775. https://doi.org/10.3390/atmos11080775

Sigro, J., Cisneros, M., Perez-Luque, A. J., Perez-Martinez, C., & Vegas-Vilarrubia, T. (2024). Trends in temperature and precipitation at high and low elevations in the main mountain ranges of the Iberian Peninsula (1894–2020): The Sierra Nevada and the Pyrenees. *International Journal of Climatology*, *44*(9), 2897–2920. https://doi.org/10.1002/joc.8487

Sjöström, M., Wold, S., & Söderström, B. (1986). PLS DISCRIMINANT PLOTS. In *Pattern Recognition in Practice* (pp. 461–470). Elsevier. https://doi.org/10.1016/B978-0-444-87877-9.50042-X

Smyth, P. (1997). Clustering Sequences with Hidden Markov Models. *Adv. Neural Inf. Proceso. Syst*, *9*, 648–654.

Song, M., & Zhang, L. (2008). Comparison of cluster representations from partial second- to full fourth-order cross moments for data stream clustering. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 560–569. https://doi.org/10.1109/ICDM.2008.143

Tebaldi, C., Debeire, K., Eyring, V., Fischer, E., Fyfe, J., Friedlingstein, P., Knutti, R., Lowe, J., O'Neill, B., Sanderson, B., van Vuuren, D., Riahi, K., Meinshausen, M., Nicholls, Z., Tokarska, K. B., Hurtt, G., Kriegler, E., Lamarque, J.-F., Meehl, G., … Ziehn, T. (2021). Climate model projections from the Scenario Model Intercomparison Project (ScenarioMIP) of CMIP6. *Earth System Dynamics*, *12*(1), 253–293. https://doi.org/10.5194/esd-12-253-2021

Teodoro, T. A., Reboita, M. S., Llopart, M., da Rocha, R. P., & Ashfaq, M. (2021). Climate Change Impacts on the South American Monsoon System and Its Surface–Atmosphere Processes Through RegCM4 CORDEX-CORE Projections. *Earth Systems and Environment*, *5*(4), 825–847. https://doi.org/10.1007/s41748-021-00265-y

Tessier, Y., Lovejoy, S., Hubert, P., Schertzer, D., & Pecknold, S. (1996). Multifractal analysis and modeling of rainfall and river flows and scaling, causal transfer functions. *Journal of Geophysical Research: Atmospheres*, *101*(D21), 26427–26440. https://doi.org/10.1029/96JD01799

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

United Nations Convention to Combat Desertification. (2022). *Drought in Numbers 2022 - restoration for readiness and resilience*. https://www.unccd.int/sites/default/files/2022-06/Drought%20in%20Numbers%20%28English%29.pdf

Van Valkengoed, A. M., Steg, L., & Perlaviciute, G. (2021). Development and validation of a climate change perceptions scale. Journal of Environmental Psychology, 76, 101652. https://doi.org/10.1016/j.jenvp.2021.101652

Varmuza, K., & Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press. https://doi.org/10.1201/9781420059496

Vicedo-Cabrera, A. M., Guo, Y., Sera, F., Huber, V., Schleussner, C.-F., Mitchell, D., Tong, S., Coelho, M. de S. Z. S., Saldiva, P. H. N., Lavigne, E., Correa, P. M., Ortega, N. V., Kan, H., Osorio, S., Kyselý, J., Urban, A., Jaakkola, J. J. K., Ryti, N. R. I., Pascal, M., … Gasparrini, A. (2018). Temperature-related mortality impacts under and beyond Paris Agreement climate change scenarios. *Climatic Change*, *150*(3–4), 391–402. https://doi.org/10.1007/s10584-018-2274-3

Wang, X., & Hyndman, R. (2006). Characteristic-Based Clustering for Time Series Data. *Data Mining and Knowledge Discovery*, *13*, 335–364. https://doi.org/10.1007/s10618-005-0039-x

Warren Liao, T. (2005). Clustering of time series data - A survey. *Pattern Recognition*, *38*(11), 1857–1874. https://doi.org/10.1016/j.patcog.2005.01.025

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2008). Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, *134*(630), 241–260. https://doi.org/10.1002/qj.210

Wierzchoń, S. T., & Kłopotek, M. A. (2018). *Modern Algorithms of Cluster Analysis*. http://www.springer.com/series/11970

Wootten, A. M., Massoud, E. C., Waliser, D. E., & Lee, H. (2023). Assessing sensitivities of climate model weighting to multiple methods, variables, and domains in the south-central United States. *Earth System Dynamics*, *14*(1), 121–145. https://doi.org/10.5194/esd-14-121-2023

Xiao, Y., Liu, J. J., Hu, Y., Wang, Y., Lai, K. K., & Wang, S. (2014). A neuro-fuzzy combination model based on singular spectrum analysis for air transport demand forecasting. *Journal of Air Transport Management*, *39*, 1–11. https://doi.org/10.1016/j.jairtraman.2014.03.004

Xie, X. L., & Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(8), 841–847. https://doi.org/10.1109/34.85677

Xiong, Y., & Yeung, D. (2002). *Mixtures of ARMA Models for Model-Based Time Series Clustering **. 717–720.

Yue, S., & Hashino, M. (2003). Long term trends of annual and monthly precipitation in Japan 1. *JAWRA Journal of the American Water Resources Association*, *39*(3), 587–596. https://doi.org/10.1111/j.1752-1688.2003.tb03677.x

Zhang, X., Alexander, L., Hegerl, G. C., Jones, P., Tank, A. K., Peterson, T. C., Trewin, B., & Zwiers, F. W. (2011). Indices for monitoring changes in extremes based on daily temperature and precipitation data. In *Wiley Interdisciplinary Reviews: Climate Change* (Vol. 2, Issue 6, pp. 851–870). Wiley-Blackwell. https://doi.org/10.1002/wcc.147

Zhao, Y., & Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. *International Conference on Information and Knowledge Management, Proceedings*, 515–524. https://doi.org/10.1145/584792.584877

Zhigljavsky, A. (2011). Singular Spectrum Analysis for Time Series. In *International Encyclopedia of Statistical Science*. Springer, Berlin, Heidelberg. https://doi.org/https://doi.org/10.1007/978-3-642-04898-2_521

Las raíces de la educación son amargas, pero sus frutos son dulces.
Aristóteles