Facultad de Ciencias.

Aplicación de técnicas de agrupamiento para la detección de tipos de tiempo atmosféricos condicionados a observaciones en superficie.

Application of clustering techniques for detecting types of weather conditioned to surface observations.

Trabajo de fin de grado para acceder al

GRADO EN MATEMÁTICAS.

Manuel Margallo Balbás

Dirigido por

Juan Antonio Cuesta Albertos
Departamento de Matemáticas, Estadística y Computación
y Daniel San Martín
Predictia S.L.

17 de octubre de 2016

Índice

Abstract

This research attemps to classify the climate in the Iberian Peninsula with Multivariant Analysis techniques. Statistical analyses on climate have been developed since the early twentieth century and have laid the foundations of today's state of the art methods. This work merges two fields, statistics and environmental science.

We use variables both experimental and products of a model and we transform them into indicators grouped in fewer factors that summarize and explain the data. Finally, we try to group the days according to their similarity to develop a list of types of days in Spain.

Keywords: climatology, climate, Iberian Peninsula, Principal Component Analysis, k-means.

Resumen

Este trabajo trata de clasificar el clima en la Península ibérica con técnicas de Análisis Multivariante. Análisis estadísticos sobre el clima se han desarrollado desde principios del siglo XX y constituyen las bases de los métodos de vanguardia. Este trabajo une dos disciplinas, la estadística y las ciencias medioambientales.

Usamos variables tanto experimentales como producidas por un modelo y las transformamos en indicadores y agrupamos en una pequeña cantidad de factores que resumen y explican los datos. Finalmente, tratamos de agrupar los días según su semejanza para desarrollar una lista de tipos de días que hay en España.

Palabras Clave: climatología, clima, Península Ibérica, Análisis de Componentes Principales, k-medias.

1. Objetivos.

En este trabajo aplicamos técnicas de clasificación no supervisada a datos de reanálisis (que reproducen el estado de la atmósfera día a día) para identificar distintos tipos de tiempo a escala diaria y que separe también un fenómeno en superficie, buscando entender este último.

Clasificamos los días de la región de la Península ibérica, el sur de Francia, el norte de Marruecos y el Mediterráneo durante 30 años usando dos fuentes de información: datos observados sobre el terreno y proporcionados por un modelo repetiendo el análisis varias veces dando distinta importancia a ambas fuentes para entender el efecto de los datos observados sobre el terreno.

Después, tratamos de interpretar los resultados y veremos si los grupos corresponden con un mes o estación concretos y si son un tipo de día que se corresponde con los de la península Ibérica.

2. La teoría detrás de la clasificación del clima.

El ser humano trata de entender su entorno para su beneficio y un ejemplo de ello es su relación con el clima por su influencia en las cosechas y la caza entre otros intereses de la sociedad.

Desde los faraones se predice sobre tiempo atmosférico hasta a un año o dos vista y estas predicciones profundamente mitológicas pero con procedimientos rigurosos son muestras de la curiosidad humana y su celo por la metodología que siglos después lleva a entender el clima.

El limitado desarrollo en geografía o la medición de la temperatura limitó hasta el siglo XVII en Europa la climatología con lo que las observaciones acerca del clima mezcladas con leyendas y tradición fueron la única ayuda para entenderlo durante la Edad Media.

La sociedad moderna necesitaba desarrollar una clasificación global y científica del clima que desarrolla en el siglo XX uno de los pioneros, el ruso Vladimir Peter Köppen publicando un método de discriminación climática aún vigente.

La Figura 1 procedente de [1] y muestra un mapa climático de Europa clasificado según Köppen que se basa en las pautas de precipitación y temperaturas anuales durante largos períodos de tiempo.

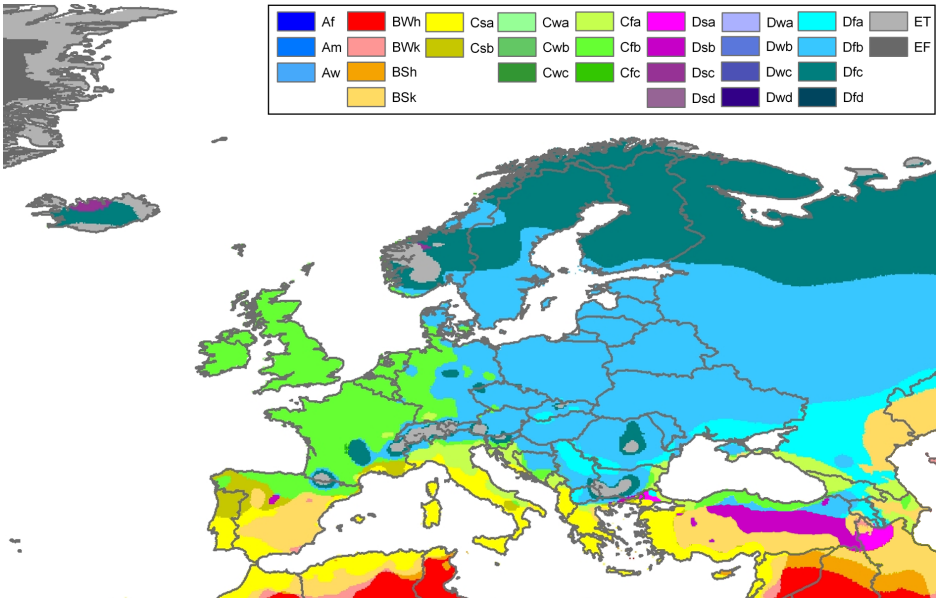

Figura 1: mapa de Europa coloreado por climas según la clasificación de Köppen. [1]

Es un ejemplo de inversión en ciencia básica que produce avances potentes ya que las condiciones climáticas son impredecibles para el ojo inexperto pero la investigación permite desarrollar aplicaciones revolucionarias.

La pesca es más segura y los efectos de sequías o inundaciones son menores con la monitorización del clima; las obras públicas combaten la erosión costera con décadas de antelación por las predicciones sobre el nivel del mar; el estudio del clima ayuda a planear la agricultura mundial con alcance cada vez mayor.

El estudio del clima no sólo alcanza el presente sino que nos muestra el futuro, previendo el efecto de la sociedad en el planeta y necesitamos que la paleoclimatología busque evidencias de climas pasados para dar contexto al poco tiempo que llevamos recopilando información.

2.1. Los datos experimentales en España.

La Agencia Estatal de Meteorología (AEMET) es el organismo español que presta servicios meteorológicos y la única organización meteorológica estatal siendo su historia paralela a la del aparato del estado.

El propósito de la institución en el siglo XIX es climatológico, recopilando y confeccionando datos que describan el clima, carente de medios y estando la previsión del tiempo fuera de su alcance.

Gracias al ejemplo de otras agencias europeas y la rapidez de las redes telegráficas el Instituto Central Meteorológico publica el 1 de marzo de 1893 un boletín con la predicción del tiempo 'probable' para ese día, el primero de la historia de España.

La medición es deficiente al ser los colaboradores voluntarios y sin criterios unificados pero en 1910 la red añade 400 estaciones fundamentalmente pluviométricas con algunas termopluviométricas asistidas técnicamente por la agencia estatal.

En los años 20 para asistir a la industria aeronaútica, la agencia aumenta la precisión administrando estaciones y creando otras en puntos clave; la plantilla aumenta entre 1905 y 1920 de dos a cien empleados entre ellos físicos españoles de calidad como Arturo Duperier.

Tras la Guerra Civil, un mando militar supervisa el servicio que cuenta con colaboración alemana, la adición de personal civil y la creación de un cuerpo de medición militar.

En los años 40 la universidad coopera y en los 50 la agencia coopera internacionalmente para coordinar y procesar datos, estos cambios mejoran los resultados y una muestra es la progresiva divulgación de sus productos como el parte meteorológico televisado.

En la década de los 70 los meteorólogos dirigen la organización y encargan la medición al cuerpo de observadores recién creado y mejor formado, que reemplaza al personal anterior.

En los 80 renuevan las instalaciones y comienzan a predecir el tiempo atmosférico con ordenadores además de usar satélites y estaciones automáticas de medición.

En los años noventa crean la red de medición de descargas eléctricas y la

de radares meteorológicos y como algunos datos de este trabajo son de esas fechas. Esta sección resume dos discursos sobre la agencia; [2, 3]

2.2. Datos provenientes de modelo

Pero usamos también datos sintéticos, de un modelo meteorológico actual a lo que dedicaremos unas líneas en esta introducción.

Desde los años 20 el paradigma de los modelos meteorológicos consiste en que la atmósfera es un fluido y catamos en puntos interesantes para que la dinámica de fluidos y la termodinámica expliquen su comportamiento inmediato ya que la naturaleza caótica del problema impide pronósticos a largo plazo.

En los años 50, las organizaciones implicadas desarrollan ordenadores que consiguieron que los modelos numéricos pronosticasen el tiempo y el servicio meteorológico de Estados Unidos incorpora en esa época sus predicciones.

Las técnicas estándar aprovechan su fuerza de computación y los métodos que simplifican los cálculos, como el de los elementos finitos, distando los dos puntos más cercanos que calculan entre 5 y 200km, lo que complica tratar fenómenos de reducido tamaño como cúmulos.

Las redes de medición y los satélites artificiales meteorológicos miden los océanos con un factor de error que debemos contar: los promedios de datos conocidos reemplazan las zonas no medidas.

Un beneficio claro de los modelos son sus aplicaciones como la predicción de los incendios combinando variables climáticas y orografía siendo estos fenómenos influenciados por el clima modelados para beneficiarnos como se ve en la figura 2, tomada de [4]

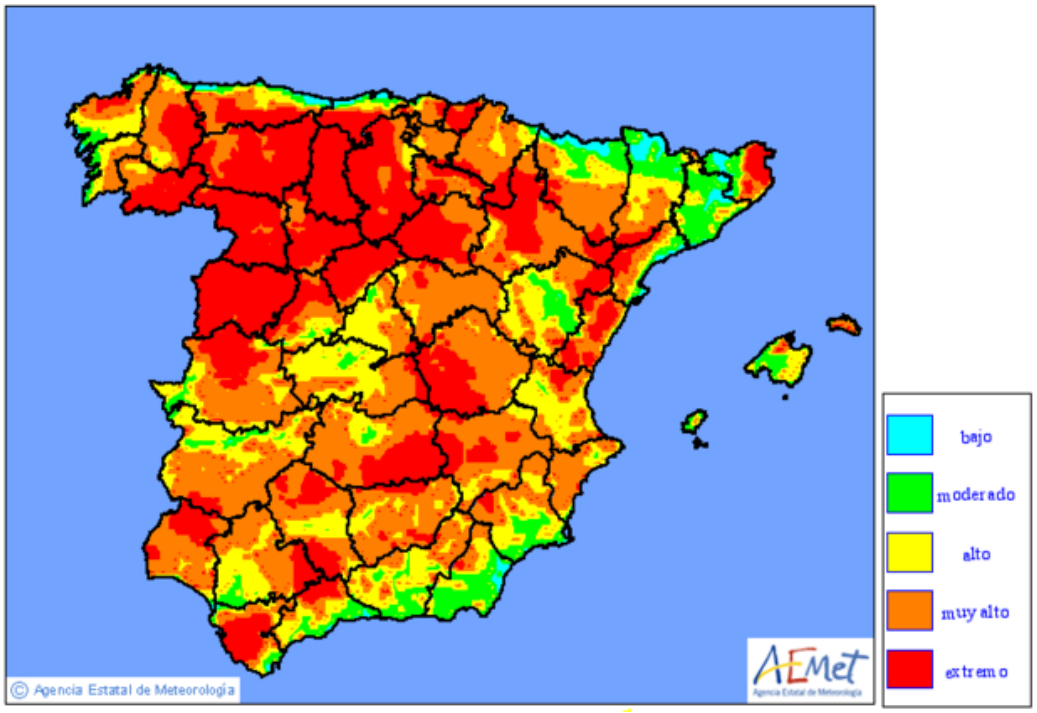

Figura 2: modelización AEMET de las condiciones propicias para incendios en España.

por otro lado, necesitamos evitar la incertidumbre del modelo y desde los años 90 se usa el sistema de predicción por conjuntos, en el que varios datos meteorológicos parecidos entre sí se utilizan para predecir el tiempo y luego se promedian.

3. Las herramientas matemáticas.

La Estadística Multivariante estudia los vectores aleatorios, conjuntos de variables aleatorias y de sus relaciones. Tiene aplicaciones en muchas ciencias y de las herramientas que proporciona nosotros usaremos el Análisis de Componentes Principales, las k-medias y la evaluación de clústeres. [5, 6]

En el caso de este estudio poseemos una base de datos en la que hay datos meteorológicos de varios años en nuestra zona de análisis: la Península ibérica, el sur de Francia y la costa noroeste de África. A su vez hay unos puntos

donde hay medidas físicas relevantes: altura, longitud, latitud, presión, viento o temperatura siendo este experimento con otras regiones un ejemplo natural para estas técnicas. Entre los autores que utilizan este ejemplo, destaca[7].

La secciones: en la 3.1 presentamos el Análisis de Componentes Principales; en la sección 3.2 proponemos el Análisis Clúster; en la 3.3 exponemos algunas herramientas para evaluar los análisis clúster y en la 3.4 analizamos modos de medir la correlación entre variables.

3.1. Análisis de Componentes Principales.

El Análisis de Componentes Principales reduce el número de variables buscando en cada dimensión el subespacio que mejor aproxima los datos. Un subespacio es un conjunto de vectores aleatorios que también contiene a sus sumas y su multiplicación por escalares.

Por su interés explico el concepto de varianza explicada por un vector que es la varianza de los datos proyectados en un vector del subespacio y la varianza explicada acumulada es la varianza del subespacio que hemos creado, siempre menor a la dispersión total de los datos, esto hace que podamos asignarle un porcentaje calculando qué proporción de la varianza de los datos representa.

La varianza explicada tiene relación con nuestros criterios de parada ya que escogemos la menor dimensión posible del subespacio que contenga suficiente información y hay varios métodos populares:

Buscando la eficiencia del método incluyendo variables según la varianza explicada y parar cuando la razón de varianza explicada por variable baje demasiado, visualizamos mejor esto con una gráfica de varianza contra dimensión donde se busca el 'codo' o punto en que la pendiente de la gráfica cambia bruscamente. Escogida esa dimensión obtenemos un conjunto de trabajo que preserva la información más compacta, que es favorable porque consigue el objetivo sin perder información sustancial.

Otra manera es buscar tantas variables como podamos explicar, para explorar los datos y entender un grupo de datos podemos buscar en las componentes principales nexos naturales entre variables para inferir su estructura interna pero no las usamos con este el motivo, así que no usaremos este criterio.

Nosotros utilizamos el Análisis de Componentes Principales para trabajar con datos manejables de menor dimensión así que pediremos un umbral de varianza explicada acumulada a un experto que garantice que la información relevante queda en el subespacio que creamos, que será el menor posible que cumpla esa condición.

Esta técnica tiene una ventaja sobre el ruido de los datos ya que la distribución del ruido gaussiano es invariante por la transformación lineal que utilizamos con el Análisis de Componentes Principales y al concentrar la varianza en los primeros componentes principales aumenta la proporción señal a ruido y como consecuencia, el ruido dominará las últimas componentes pero nosotros no las usaremos. Esta técnica tambien tiene desventajas, como dar prioridad a las relaciones lineales sobre tipo de relaciones.

Esta explicación resume el funcionamiento de esta técnica, pero se omiten propiedades y detalles de cómo se lleva a cabo, para hallar más información sobre este tipo de análisis, consultar:[8]

3.2. Análisis clúster.

Nuestro objetivo es agrupar los días por tipos y para ello necesitamos métodos independientes de un modelo previo, así que utilizamos las técnicas del análisis clúster. El problema general parte de un conjunto de datos multidimensional y queremos que los datos que se parecen pertenezcan al mismo grupo y los datos en distintos grupos sean diferentes entre sí para lo que usaremos las técnicas del análisis clúster que buscan los mejores grupos en un conjunto de datos y difieren en el sentido en el que son mejores. Como referencias, dejamos:[10]

Estas técnicas se dividen según sus características. Pueden buscar soluciones globales, hallar la mejor división, o conformarse con óptimos locales, eligiendo una agrupación aceptable. Asimismo, pueden ser jerárquicas o no jerárquicas.

El análisis jerárquico crea una jerarquía entre los grupos: de uno grande desgrana otros más pequeños (clúster divisivo) o agrupa los pequeños (clúster aglomerativo), en ambos el resultado es un árbol que va de un grupo con todos los datos a grupos de un elemento. El clúster viene cuando cortamos el árbol por el número de grupos deseado.

El análisis no jerárquico trabaja con un número de grupos determinado de antemano y construye una partición homogénea. Sus métodos son variados: puede buscar la homogeneidad de los grupos o puede suponer una distribución de los grupos y aplicarla a los datos. Nosotros usamos un algoritmo creado en los años 60 que se llama k-medias.

Para hallar los grupos definitivos, k-medias usa la estabilidad de los centroides[1] de unas agrupaciones iniciales en las que en cada paso halla los centroides de los grupos, clasifica de nuevo todos puntos según el centroide más cercano y vuelve a empezar. Da el proceso por terminado cuando los centroides quedan casi estáticos con lo que no analiza todas las configuraciones posibles lo que le hace un método local.

Al ser un método local, no sabemos si hallaremos la agrupación óptima y para solucionarlo ejecutamos el algoritmo varias veces con valores iniciales distintos escogiendo la mejor configuración obtenida. Tampoco hemos precisado qué es lo que minimiza el algoritmo, cuál es su medida de homogeneidad. Este factor es su varianza intra grupos: si la varianza de cada grupo sumada es la varianza explicada, que es una fracción de la total y que representa la cantidad de variación explicada por su pertenencia al grupo, la varianza intragrupo es la otra parte de la varianza, lo que los grupos no explican. Dejamos como referencia [9]

3.2.1. Calidad del análisis clúster.

El análisis clúster produce grupos que hacen que los datos de un grupo se parezcan más entre sí que a los de otros grupos. Según los parámetros que demos un algoritmo, como el número de grupos que tiene que buscar o la distancia que debe usar, un mismo conjunto de datos produce resultados muy distintos.

Entonces analizando su validez elegimos la mejor agrupación lo que es más importante en conjuntos como nuestros datos con muchas dimensiones donde visualizarlos en mapas de 2D o 3D no es una opción.

Hay dos conceptos fundamentales para validar agrupamientos: la compacidad y la separación. La primera implica que cada miembro de un grupo debería estar cerca de los demás del grupo, siendo una buena medida la

[1]El punto formado por las medias de las coordenadas los puntos de un grupo.

varianza; la separación de los grupos la observamos mediante diferentes medidas: la distancia de sus centroides, la distancia mínima y la máxima entre los puntos de esos grupos.

Hay tres tipos de evaluaciones de los agrupamientos: interna, que se fija en las características internas de los grupos, sus índices de calidad valoran los grupos sólo por sus propios datos; externa, con información extra, como puntos bien clasificados para validar la solución propuesta; por último, la intermedia o relativa que compara entre distintos clústeres de un mismo grupo de datos.

Siendo nuestro algoritmo no supervisado y sin referencias externas, no esperamos resultados concretos y lo que necesitamos es un criterio para justificar nuestras conclusiones. Así, aspiramos a entender el clima con los datos proporcionados, recuperando los días típicos sin información externa y usaremos varios índices internos que nos servirán para medir la calidad de estas agrupaciones. Para hallar más información sobre los índices de validez interna y la calidad del análisis clúster, consultar [11]

3.2.2. Índices de validez

En esta sección describimos cuatro índices de evaluación interna con los que pretendemos evaluar los datos. Son los índices Dunn, el Davies-Bouldin, R-cuadrado (en adelante RS por sus siglas en inglés) y el SD. Es una discusión más larga que continúa en: [11, 12, 13]

Para hablar de estos índices escribiremos sus fórmulas, que tendrán símbolos con significados que aquí aclaramos. Dado un grupo ω dividido en k grupos, x_k un punto cualquiera del grupo G_k,$\|G_k\|$ el número de elementos y C_k su centroide, $\sigma(X)$ es la desviación típica del vector aleatorio X de dimensión d asociado a una distribución ω.

Definimos la norma de un vector, la varianza de un vector y la varianza de un grupo k.

$$\|X\| = \sqrt{XX^T} \tag{1}$$

$$\sigma(X)^p = \frac{1}{k}\sum_{i=1}^{k}({X_i}^p - \overline{X_i}^p)^2; \sigma(X) = \begin{bmatrix}\sigma(X)^1 \\ ... \\ \sigma(X)^d\end{bmatrix} \tag{2}$$

$$\sigma(C_k)^p = \frac{1}{\|G_k\|}\sum_{i=1}^{k}({X_i}^p - \overline{C_i}^p)^2\sigma(C_k) = \begin{bmatrix}\sigma(C_k)^p \\ ... \\ \sigma(C_k)^p\end{bmatrix} \tag{3}$$

El índice Dunn viene a ser la razón entre la mínima separación entre grupos y la mínima compacidad de los mismos. Expone el peor escenario posible, requiriendo un sólo grupo disperso para obtener bajos valores de este indicador. Esto le convierte en un indicador sensible al ruido, requiriendo unos pocos datos atípicos para variar mucho el diámetro de un grupo, lo cual es bastante para producir cambios drásticos en el índice, como menciona [13, p.16]. definimos el índice Dunn como:

$$D = \frac{\min\limits_{1\leq i<j\leq k}(d(C_i, C_j))}{\max\limits_{1\leq k\leq m}[\max_{x\in\omega} d(x, C_k)]}$$

El siguiente indicador es el Davies-Bouldin, la media aritmética de los máximos de las medidas de similitud entre clústeres. Esto quiere decir que elegimos una manera de medir la separación entre grupos, tomando en cada grupo el valor máximo de su separación del resto de grupos, y luego promediamos. Esta definición podría no tener en cuenta lo compacto que es ningún grupo, pero en nuestro caso elegimos una medida de similitud que dé peso a la compacidad. En la fórmula de la medida de similitud se da igual peso a la compacidad de cada grupo y a la distancia entre sus centroides. Definimos el índice Davies-Bouldin como:

$$DB = \frac{1}{k}\sum_{i=1}^{k} D_i$$

siendo los D_i indicadores de la bondad de ajuste entre clústeres.

$$D_i = \max_{i\neq j} R_{i,j}, R_{i,j} = \frac{S_i + S_j}{d(C_i, C_j)}, S_i = \frac{\sum_{x_k\in G_i} d(x_i, C_i)}{||G_i||}$$

El tercero es RS, usado para medir la disimilaridad de los clústeres. El algoritmo de las k-medias utiliza este índice como función objetivo.

$$RS = \sum_{i=1}^{k} \sum_{x_i \in G_i} \|x_i - C_i\|^2$$

El cuarto es SD índice, que usa dos parámetros: uno de dispersión con la media de la norma de las varianzas de los clústeres entre la norma de los datos y otro de distancia que mide la compacidad basándose en la distancia entre los centroides. Las fórmulas son:

$$SD = Scatt \cdot Dis + Dis \tag{4}$$

$$Dis = \frac{\text{máx}_{i,j=1\ldots k}\|C_i - C_j\|}{\text{mín}_{i\neq j;i,j=1\ldots k}\|C_i - C_j\|} \sum_{m=1,m\neq z}^{k} \left(\sum_{z=1}^{k} \|C_z - C_m\| \right)^{-1} \tag{5}$$

$$Scatt = \frac{\sum_{i=1}^{m} \|\sigma(C_i)\|}{k\|\sigma(x)\|} \tag{6}$$

3.3. Correlación.

La correlación mide la magnitud de la dependencia entre dos o más variables siendo el indicador de la dependencia lineal es el 'coeficiente de correlación de Pearson' que toma valores en el intervalo [-1,1] y dejamos como referencia: [14]

Dicho coeficiente tiene dos valores de interés: la magnitud de la relación se mide con el módulo de su valor, siendo 1 total y 0 nula, y el signo define una relación decreciente si es negativa y creciente a la inversa. Dadas las variables x e y, se calcula por esta fórmula:

$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

donde $Cov(x, y)$ es la covarianza entre las variables X e Y, σ_X denota la desviación típica de la variable X y la σ_Y la de la variable Y.

Una correlación alta no implica causalidad y una correlación baja no agota la cuestión de la relación de las variables así que puede haber variables relacionadas entre sí por una causa subyacente, como la inteligencia en el

desempeño de distintas pruebas, y otras unidas por el azar. Es un indicador con sus limitaciones y conocerlas es tan importante como conocer sus usos.

3.4. Regresión Robusta.

Uno de nuestros objetivos es entender cómo afecta a la calidad del agrupamiento el peso que damos a los datos del modelo. Para eso usamos el análisis de regresión, que busca relaciones entre variables. En este caso, la variable explicativa será el peso que se le da a los datos del modelo, y la explicada el valor del índice de calidad de los datos.

En concreto usamos la regresión robusta que es insensible a valores atípicos y utilizamos una de sus variantes más simples, la que minimiza las desviaciones absolutas, LAD por sus siglas en inglés. Para más información sobre esta técnica: [15]

Empleamos el análisis de regresión cuando tenemos un conjunto de datos con valores atípicos y concluimos que el proceso no tiene errores y dichos valores no pertenecen a una población diferente. En este caso, sabemos las k-medias producen soluciones locales, lo que puede llevar a valores atípicos en soluciones que hacen necesaria esta técnica de regresión.

4. Resultados.

En este capítulo explicamos lo realizado y los resultados según esta estructura: primero aclaramos qué variables usaremos en la sección 4.1; procesamos dichas variables, explicando cada decisión en la Sección 4.2; aplicamos Análisis Componentes Principales en la Sección 4.3 y resumimos el análisis clúster en la Sección 4.4.

Los datos se analizan con el lenguaje de programación `R`, con las funciones básicas y las librerías `clv,plotrix`.

4.1. Variables.

La información de este estudio abarca la península Ibérica, el sur de Francia, el norte de África, una parte del Mediterráneo y el Atlántico por ser el estándar que se usa para medir el tiempo en España. Las variables de los datos experimentales provienen de los datos de AEMET, que proporciona datos históricos de entre 1920 y 2012 de más de 100 estaciones meteorológicas de España repartidas por su geografía, del cual usaremos el período entre 1970 y 2012. La información que ofrece son las siguientes variables: temperatura, presión, módulo y dirección del viento, horas de sol y precipitación. De ellas, usamos la fuerza del viento: pretendemos hacer una clasificación que separe especialmente bien una variable en superficie así que considerar más variables limitaría la discriminación del estudio.

La distribución de las 71 estaciones meteorológicas españolas activas en el período de 42 años que nosotros estudiamos aparece en la figura 3. Los datos proporcionados por las estaciones tienen distintos niveles de calidad y fiabilidad así que las omisiones e inconsistencias son frecuentes.

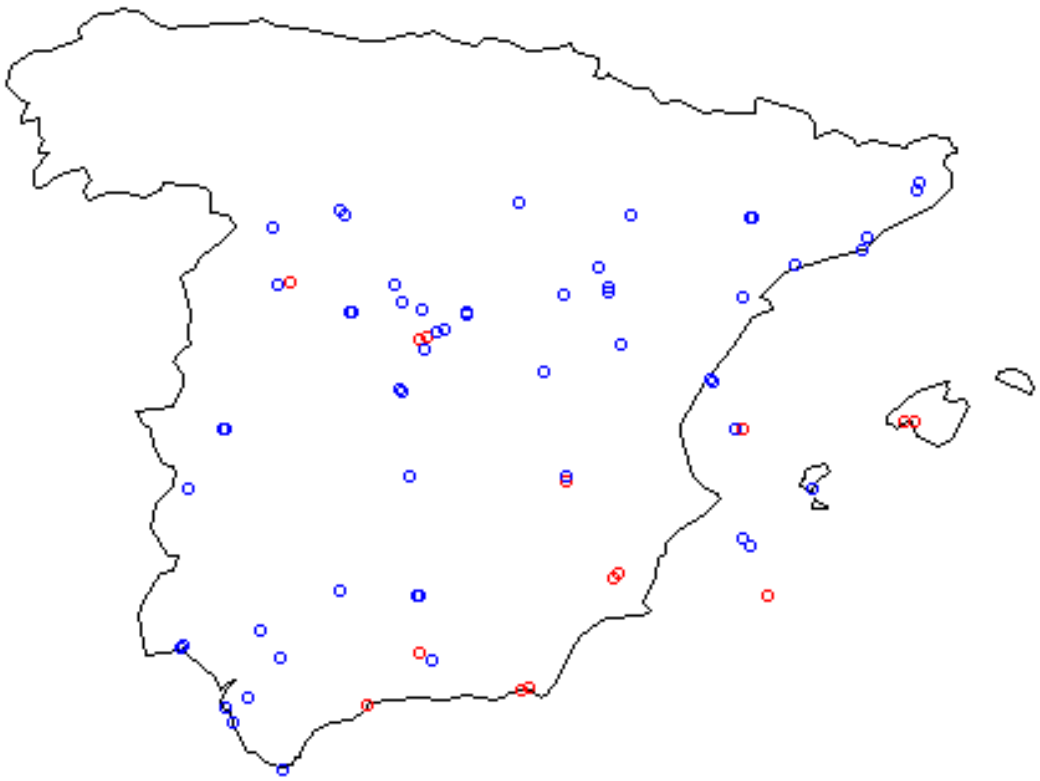

Figura 3: estaciones que nos proporcionan las observaciones.

Los datos del modelo o datos derivados provienen de un reanálisis, que es un proceso mediante el cual la información de modelos y observaciones de muchas fuentes se combinan de manera óptima para producir una estimación global de varios parámetros atmosféricos y oceanográficos. Encontraremos más información sobre los reanálisis en [17]

Nuestro reanálisis abarca desde 1979 y lo realiza la Agencia Europea de Partes Meteorológicos a Medio Plazo (ECMWF, por sus siglas en inglés) con datos que proporcionan 34 países europeos y sirve para estimar los valores iniciales de los pronósticos meteorológicos o realizar estudios climatológicos.

La figura 4 muestra la cuadrícula de los datos del modelo. La cuadrícula es gaussiana y tiene 0.7º de lado. Recibimos las siguientes variables del reanálisis: el módulo y dirección del viento a .85 bares de presión, la temperatura a dos metros de la superficie y la presión media a nivel del mar. Para más información sobre este reanálisis, un artículo sobre el tema puede

encontrarse en [18].

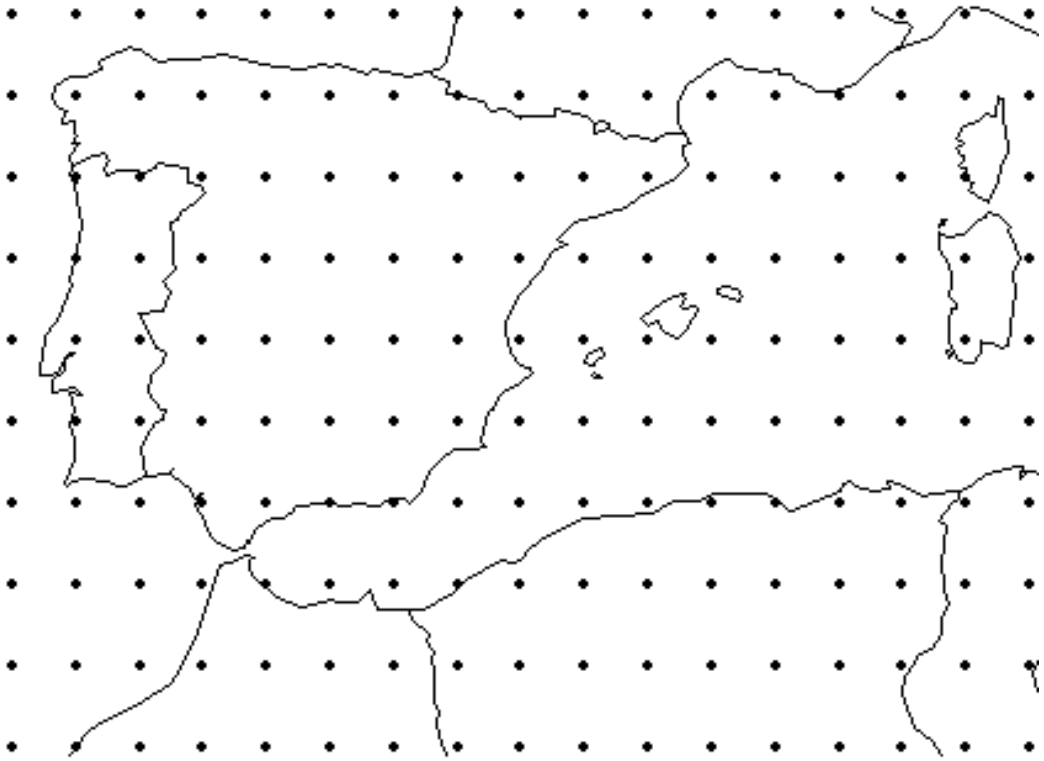

Figura 4: red de los datos generados por ordenador.

Usamos los días entre el 1 de enero de 1979 y el 31 de diciembre de 2012, 12.419 vectores como individuos con una dimensión igual al número de valores de variables meteorológicas que recibamos ese día, que en este caso son las 71 estaciones y los 680 puntos que nos da el modelo, 170 de cada una de las cuatro variables que da, pero tras procesar las variables, nos quedaremos con 39 variables en total. Por tanto tendremos una matriz de datos con 12.419 filas y 39 columnas.

4.2. Procesado de variables.

Para nuestro análisis, contamos con una serie de requisitos de calidad para estas estaciones [2] que marcamos en rojo si los cumplen. Como resumen de la discusión, esbozamos los siguientes criterios:

- Proveen un flujo de información sin interrupción lo que es fácil de comprobar pero inhabilita muchas estaciones ya que tienen fallos esporádicos de uno o dos meses de duración que las dejan fuera del análisis.
- Las estaciones deben tener una calidad mínima, que no comprometa un análisis por datos anormales que se deben a errores de medición.

Descartamos la mayor parte de las estaciones de la red para quedarnos con 14 repartidas por la Península Ibérica y Baleares, en contraste con los datos del modelo, el cual sí tiene información del resto de la zona de interés. Los datos de superficie, provenientes de las estaciones están para separar los días según una variable en superficie, la fuerza del viento, y no necesita tanta información.

Una vez seleccionadas las estaciones y los datos, se estandarizan todas las variables, de superficie y del modelo. Esto busca restar importancia a las unidades de las variables, que de otro modo influirían en la determinación de los componentes principales.

Las variables son tantas que no condensarlas provocaría un innecesario tiempo de espera en los cálculos que vamos a resumir la información del modelo usando Componentes Principales. Nos quedamos con 25 componentes Principales para evitar que se pierda información relevante, ya que con ellos tenemos un 98 por ciento de la varianza explicada.

4.3. Agrupación de datos.

Combinamos ambas fuentes de datos dándole una importancia distinta a cada una usando un parámetro de peso sobre los datos experimentales que empieza sin darle importancia y va aumentando con paso de .04 cada vez.

[2] Cuáles son los criterios que aplicar a esta red es un debate en sí mismo, para el lector interesado recomendamos la siguiente referencia [16].

Como el parámetro de peso carece de un límite natural podemos aumentarlo hasta que la agrupación sea esencialmente los datos de las observaciones. El análisis que hacemos, una vez se han obtenido las agrupaciones, consiste en observar la calidad de las mismas aplicadas a cada una de las dos fuentes de datos que tenemos lo que nos da dos valores de sus índices de validez: uno para los datos del modelo y otro para los experimentales. La variación en sus índices la explicamos usando el peso que le damos a los datos experimentales.

Y por eso necesitamos una explicación a por qué terminar en un momento dado, que depende de qué queremos hacer con el estudio. A modo ilustrativo el experimento finaliza en el punto de corte entre los índices de calidad, cuando agrupamos con calidad parecida ambas fuentes de datos.

Tenemos que especificar el número de clústeres al iniciar el algoritmo y elegimos el número de grupos sin considerar argumentos matemáticos pues los tipos de días de la atmósfera existen independientemente de su separabilidad, así que tomamos 20 clústeres porque lo usan otros estudios. Para indagar sobre dichos estudios o cómo elegir el número de días, proporciono la siguiente bibliografía: [19, 20]

4.4. Análisis de las agrupaciones.

Las gráficas que ilustran el comportamiento de las agrupaciones en los datos observados son la 5, 6, 7 y 8 por medio de un índice en cada una, que utiliza dichos datos y la agrupación para medir la calidad de esta última. Una línea azul es la regresión robusta de los valores de los índices.

Antes de empezar a analizar los índices, aclaro qué comportamiento se espera de los índices al mejorar la calidad de los datos: el RS debería aumentar, Davies-Bouldin bajar, Dunn aumentar y SD bajar.

Además, como se explica en la sección 2, el índice Dunn y el SD son sensibles a los datos atípicos. Esto ayuda a entender las gráficas que luego vemos: mejora la calidad en ambos índices pero su comportamiento es errático por su sensibilidad.

En la figura 5, el índice Davies-Bouldin baja, mostrando una mejora de los datos. En la figura 6 el índice Dunn tiene un comportamiento errático y muestra mejora global; En la figura 7 el índice RS no hace más que mejorar, con poca o ninguna dispersión de los datos; en la figura 8 le índice SD está

afectado por los mismos datos extremos que el Dunn y también la calidad de las observaciones mejora.

El comportamiento de los índices es simple: todos atestiguan que los grupos separan los datos experimentales cada vez mejor y los del modelo cada vez peor lo que es razonable: empezamos el experimento sólo con los datos del modelo y después vamos aumentando la importancia de los observados.

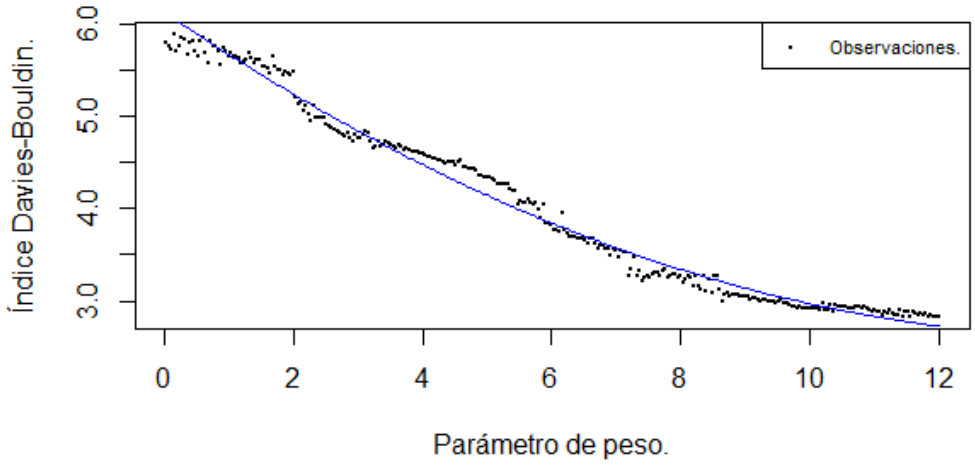

Figura 5: El índice Davies-Bouldin.

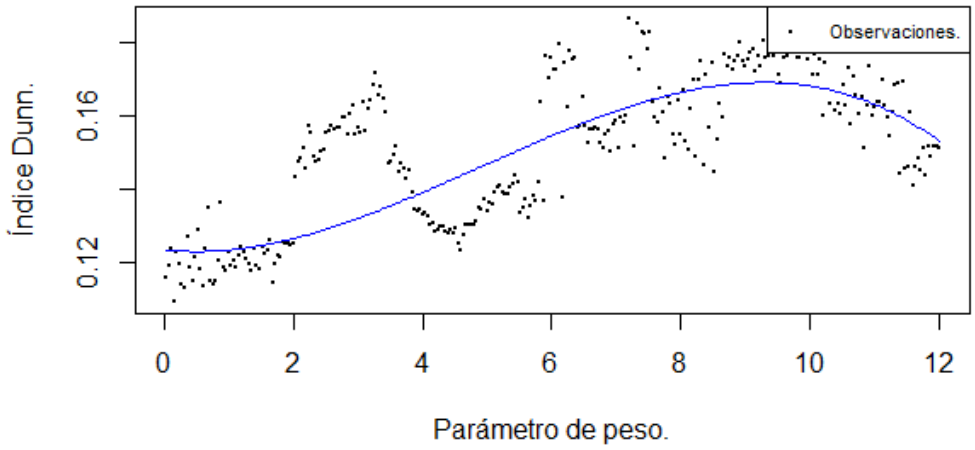

Figura 6: El índice Dunn.

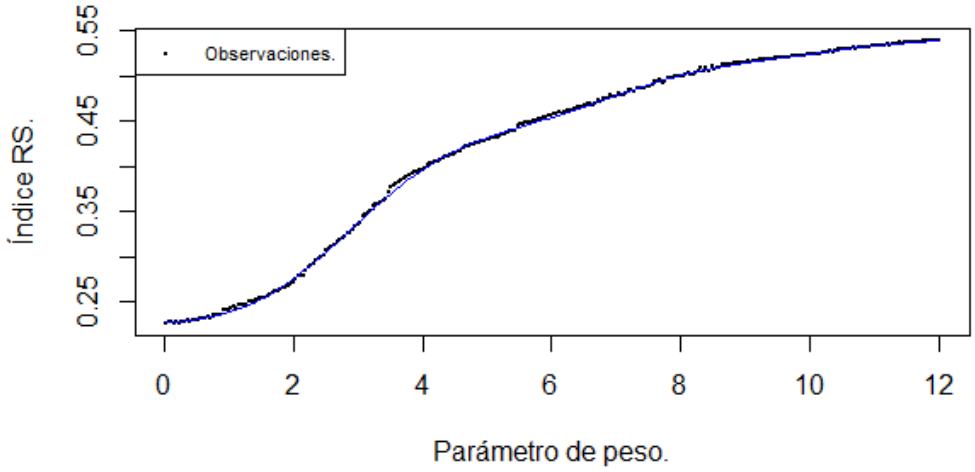

Figura 7: El índice RS.

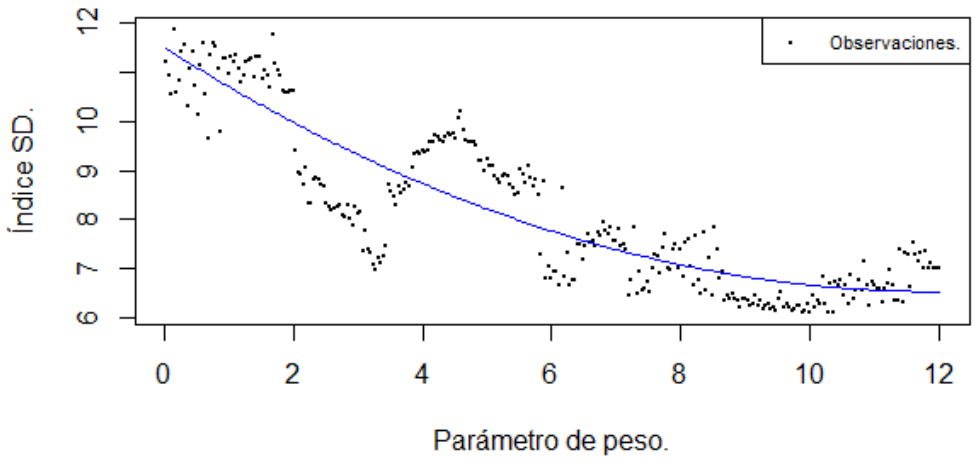

Figura 8: el índice SD.

Ahora buscamos un punto en el que los índices de calidad de una agrupación sea igual en los datos observados y en los derivados. Para eso calculamos los índices de calidad de ambos grupos de datos, y en cada gráfica hay dos grupos de puntos: el mismo índice de calidad calculado en un grupo de datos diferente. La línea roja aproxima los índices de la agrupación sobre los datos del modelo, la línea azul los de las observaciones. Como podemos ver en las figuras 9, 10, 11 y 12 Este punto en los dos índices más robustos a datos atípicos, el Davies-Bouldin y el RS, se encuentra cerca del 9, siendo los otros dos de poca utilidad en esta tarea.

El índice Dunn tan solo un empeora para los datos de la atmósfera; el índice sd no se beneficia ni empeora con el experimento por lo que no lo usamos para entender qué sucede con la agrupación. Esto hace interesante usar varios índices: Así, podemos ver cómo los datos pueden separarse mejor según su óptica, lo que ofrece una perspectiva mejor.

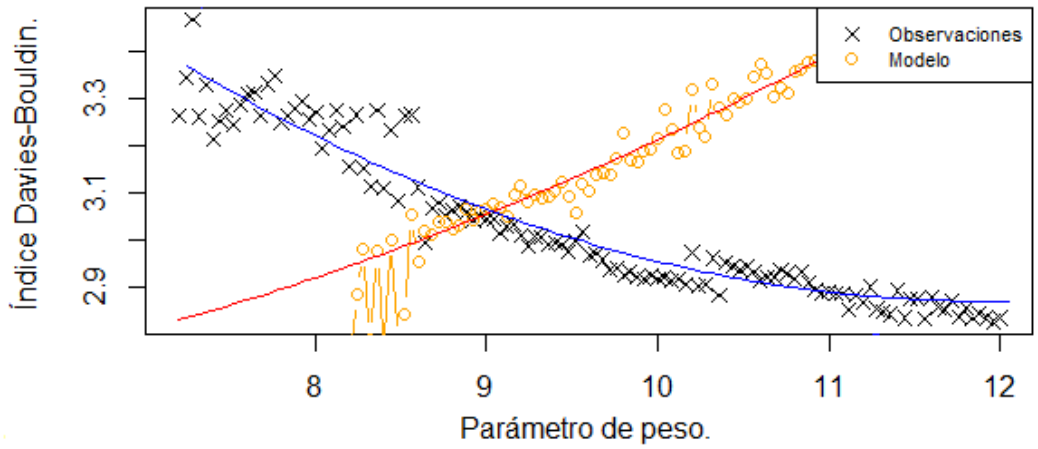

Figura 9: el índice Davies-Bouldin.

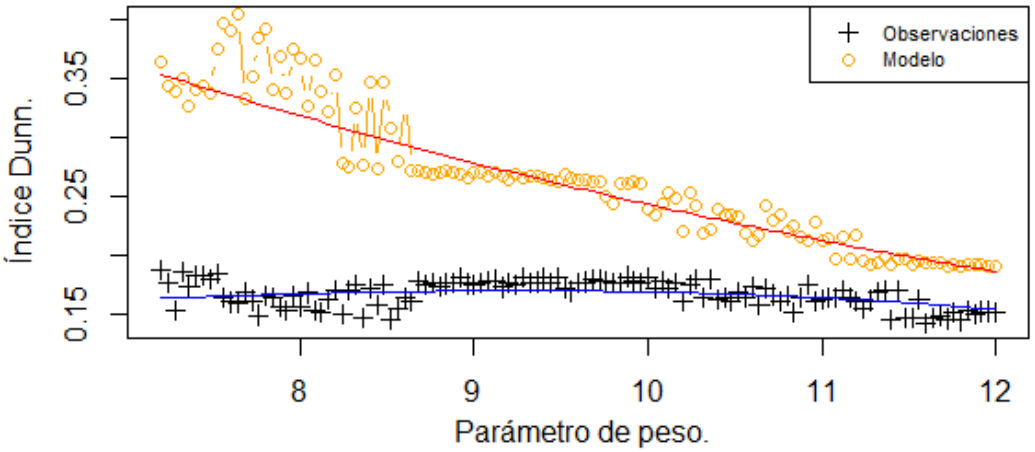

Figura 10: el índice Dunn.

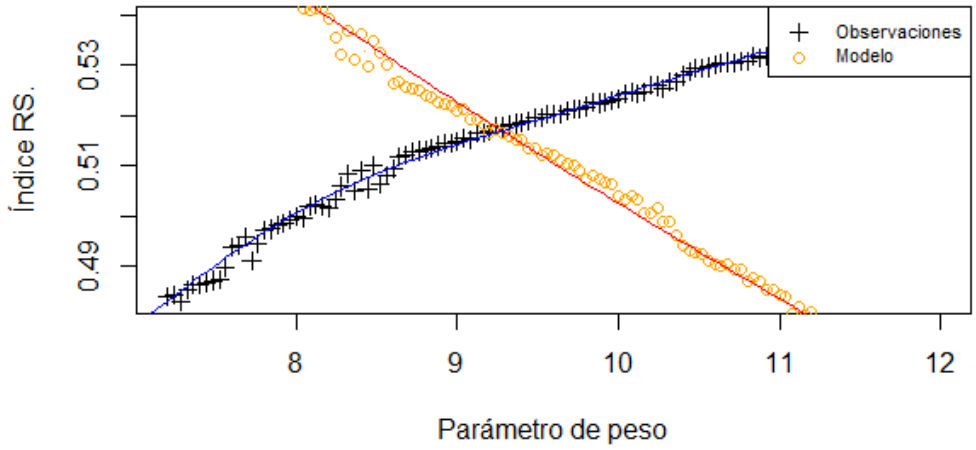

Figura 11: el índice RS.

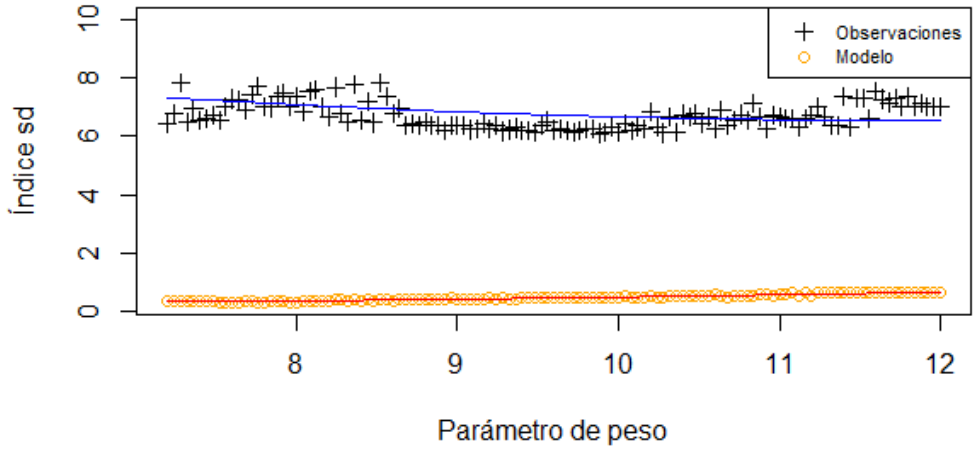

Figura 12: el índice SD.

Visto que un valor cercano al 9 es un compromiso para ambas fuentes de datos, examinamos el resultado de dicha elección.

4.5. Descripción de una agrupación.

El examen de una agrupación concreta clarifica nuestros resultados. Los datos los extraemos del modelo porque son más detallados. En cuanto a las imágenes, todas son grupos de tres mapas con su respectiva leyenda. De izquierda a derecha están: las mediana de la dirección[3] y módulo del viento, las temperaturas y la presión, medidos en quilómetros por hora, grados celsius y bares, respectivamente, siendo estos aspectos clave del clima, pretenden resumir la situación que analizamos. Primero, es importante saber qué esperar del clima en la zona, en la figura 13 sus valores medianos y en la 14 la dispersión esperada de las variables que calculamos con la mediana de las desviaciones absolutas de cada cuadrícula en el período. Los mapas con los valores de las variables son las medianas del grupo que describen: vemos ahora la mediana del periodo, luego las medianas de algunos grupos interesantes.

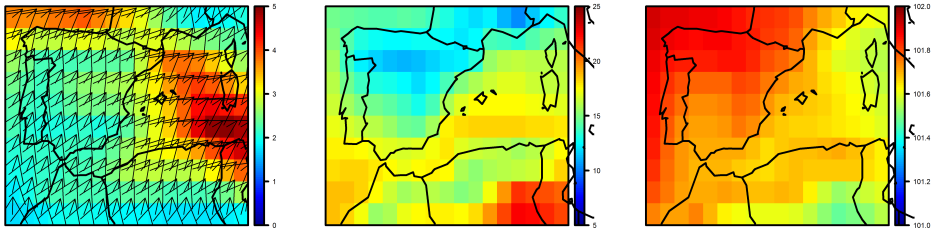

Figura 13: medianas del viento, la temperatura y la presión.

Las imágenes muestran las características del clima en la mediana del período de estudio, de 1970 a 2012: el Mediterráneo más cálido que el Atlántico, mayores temperaturas cuanto más al sur, poco o ningún viento aunque la componente atlántica de la costa cantábrica se ve claramente, al entrar el viento de esa parte.

El rango entre 101 y 102 bares de presión es el que hace propicia la lluvia o no ya que con mayor presión que 101.6 bares y hace que lo esperable sea que no llueva, siendo las bajas presiones necesarias para la borrasca lo que

[3]La dirección se calcula por sus componentes u y v, u paralelo al ecuador y v perpendicular a aquél. Están orientados de manera que u sea positivo cuando el viento sopla hacia el este y v lo sea cuando sopla hacia el norte. Su mediana se calcula calculando la de estos dos componentes

divide el mapa por la costa mediterránea que forma la diagonal donde se dividen las altas y las bajas presiones, haciendo que se esperen menos lluvias en el norte que en el sur de la Península ibérica.

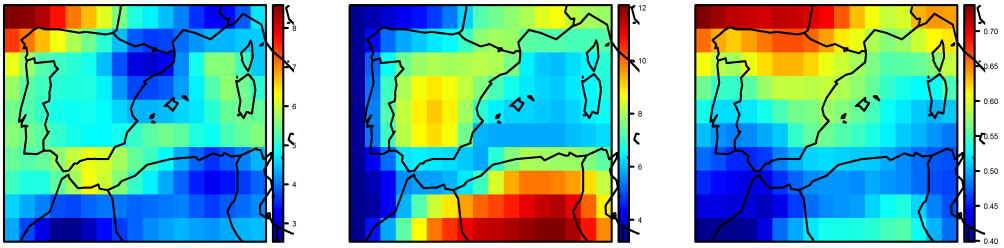

Figura 14: mediana de desviaciones absolutas del módulo del viento, la temperatura y la presión.

La figura 14 es la mediana de las desviaciones absolutas de los tres parámetros del modelo en el periodo que nos indica las zonas importantes para distinguir los tipos de días como la costa noroeste de España en la que la fuerza del viento es más variable o las zonas continentales con temperaturas más cambiantes, especialmente la africana.

La presión es más cambiante en el tercio superior, lo que enriquece el mapa anterior: con mayor variabilidad en el norte de la Península ibérica, las presiones altas de la figura 13 en la misma zona ya no implican menor posibilidad de lluvias durante el periodo.

Una vez que tenemos una imagen de cómo es un día típico y qué zonas son más volátiles, explicamos nuestras agrupaciones según los meses del año en la figura 15 con una tabla que muestra qué meses son más habituales para los 20 días que hemos elegido como típicos, más oscuro el color cuanto más habituales sean en un mes. Los días típicos los ordenamos por la estación en la que son más comunes Por ejemplo, el 1 y el 7 están concentrados en verano, el 18 es típico de septiembre y así sucesivamente, cuanto más oscuro es el color del recuadro más frecuente que ese grupo esté en ese mes. Es fácil ver que la incidencia de la mayoría de los grupos está en una o dos estaciones.

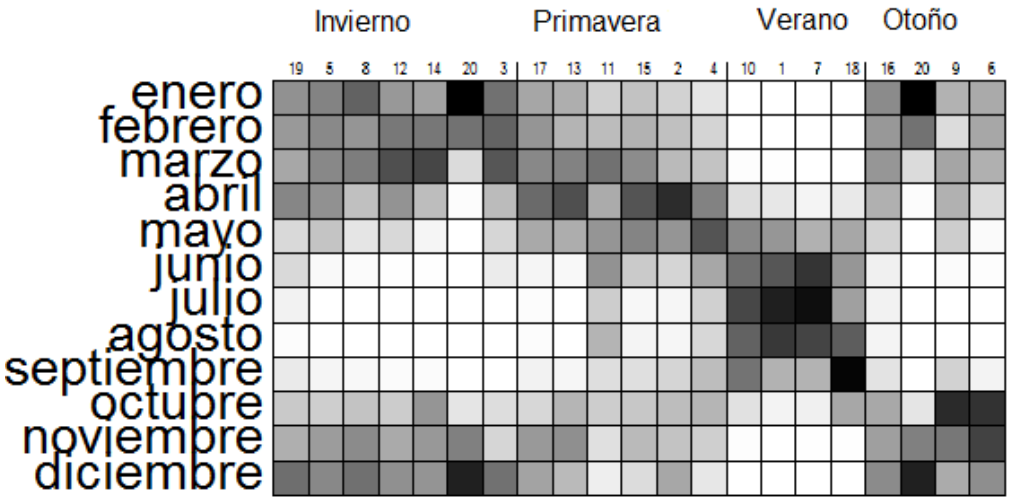

Figura 15: días típicos agrupados según la estación en la que son más comunes.

Elijo un día por estación para mostrar las distintas situaciones de la región. El grupo 1 representa el verano porque de los más frecuentes en el estío es el más cálido, teniendo sus mapas en la figura 16. En la figura 17 el grupo 9 ejemplifica el otoño porque es el que más veces está en esa estación. El invierno en la figura 18 lo representa el grupo 20 por la misma razón. La primavera tiene el grupo 15, en la figura 19. Como en los mapas anteriores, son las medianas de los 4 valores que nos da el modelo: dirección y módulo del viento en la primera, la temperatura en la segunda y la presión en la última.

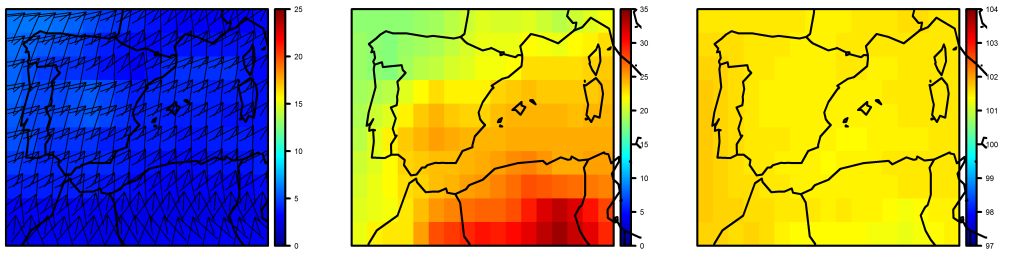

Figura 16: las medianas del grupo 1, representando el verano.

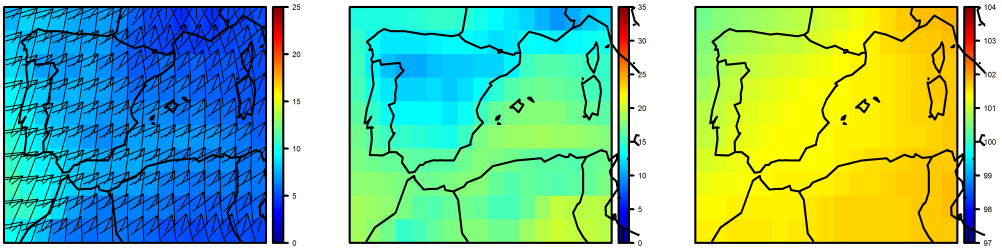

Figura 17: las medianas del grupo 9, representando el otoño.

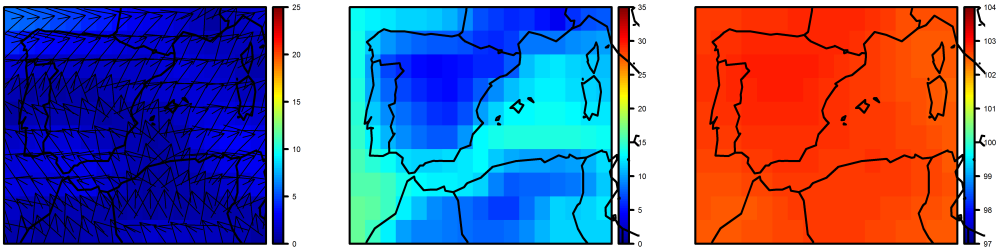

Figura 18: las medianas del grupo 20, representando el invierno.

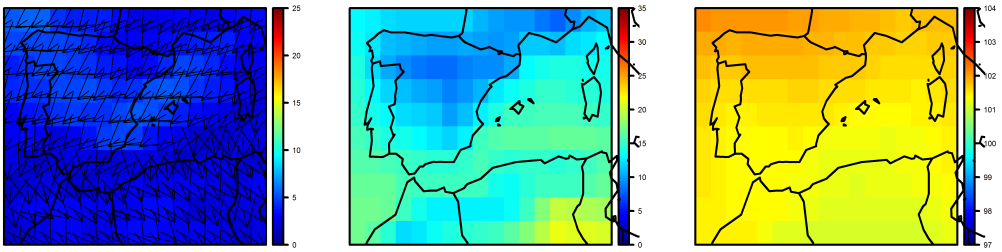

Figura 19: las medianas del grupo 15, representando la primavera.

5. Conclusiones.

Este trabajo sirve para comprender el rol del módulo del viento, en una región de la atmósfera y saco de su realización las siguientes conclusiones:

- Tenemos suficiente información meteorológica para comprender cómo funciona la atmósfera pues el acceso a datos de gran calidad en un período amplio permite ver la infraestructura a nuestra disposición, casi siempre invisible u olvidada.
- El módulo del viento es un buen predictor pues aún teniendo mucho peso en el resultado de la agrupación los grupos generados encajan con nuestro conocimiento del clima.
- Los índices de calidad de las agrupaciones tienen un rol importante en decisiones como elegir el valor de un parámetro, el número de clústeres o más generalmente, permiten entender procesos de difícil visualización.
- Mostrar resultados es complejo y de una profundidad que no aparente y me ha sorprendido por la variedad de mapas y procedimientos para informar así como su importancia para comprender la situación.
- La opinión de un experto modela un trabajo de matemática aplicada, lo cual es positivo ya que permite utilizar información de calidad sin tener que añadirla explícitamente a tu modelo.
- La dimensión de los datos, más grande de lo que se ve en la carrera incentiva el estudio de la programación concurrente y nuevas técnicas apropiadas a ese volumen de datos.

Referencias

[1] MC. Peel, BL. Finlayson y TA. McMahon (2007). Updated world map of the Köppen-Geiger climate classification, Hydrol. E. Syst. Sci., 11, 1633-1644.

[2] M. Palomares (2012). 'AEMET a Lo Largo De Su Historia.' Discurso, Día Meteorológico Mundial, Madrid.

[3] M. Palomares (2015). Breve Historia De La Agencia Estatal De Meteorología, El Servicio Climatológico Español.

[4] http://www.aemet.es/es/eltiempo/prediccion/incendios/ayuda, a 30 de Septiembre de 2016

[5] Análisis Multidimensional. En Wikipedia, The Free Encyclopedia. Recuperado de https://en.wikipedia.org/w/index.php?title=Multidimensional_analysis&oldid=704896498 el (2016, Febrero 14).

[6] J.A. Cuesta Albertos (2012). Análisis Multivariante. Universidad de Cantabria

[7] J. Gutiérrez, R. Cano, A. Cofiño y R. Sordo (2004). Redes Probabilísticas Y Neuronales En Las Ciencias Atmosféricas. Series Monográficas Del Instituto Nacional De Meteorología.

[8] I. Jolliffe (2002). Principal Component Analysis. John Wiley and Sons, Ltd.

[9] G. Gan, C. Ma y J. Wu (2007). Data clustering: theory, algorithms, and applications (Vol. 20). Siam.

[10] Cluster analysis. En Wikipedia, The Free Encyclopedia. recuperado de https://en.wikipedia.org/w/index.php?title=Cluster_analysis&oldid=739555540 el (2016, Septiembre 15).

[11] F. Kovács, C. Legány, y A. Babos (2005). Cluster validity measurement techniques. In 6th International Symposium of Hungarian Researchers on Computational Intelligence.

[12] M. Halkidi, Y. Batistakis, y M. Vazirgiannis (2001). On clustering validation techniques. Journal of intelligent information systems, 17(2-3), 107-145.

[13] Q.Zhao (2012). Cluster Validity in Clustering Methods. Publications of the University of Eastern Finland

[14] Coeficiente de correlación de Pearson. En Wikipedia, The Free Encyclopedia. recuperado de https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient el (2016, septiembre 26).

[15] R. F. Phillips (2002). Least absolute deviations estimation via the EM algorithm. Statistics and Computing 12, 3 (July 2002). 281-285. DOI=http://dx.doi.org/10.1023/A:1020759012226

[16] S. Herrera, J.M. Gutiérrez, R. Ancell, M.R. Pons, M.D. Frías y J. Fernández (2010). Development and Analysis of a 50 Year High-Resolution Daily Gridded Precipitation Dataset over Spain (Spain02). International Journal of Climatology.

[17] Uppala, S., et al. (2005). The ERA-40 Re-Analysis. Quart. J. Roy. Meteor. Soc., 131, 2961–3012. DOI

[18] P. Berrisford et al. (2011). The ERA-Interim archive Version 2.0, ERA Report Series, ECMWF

[19] J.M. Gutiérrez, A.S. Cofiño, R. Cano, y M.A. Rodríguez (2004). Clustering methods for statistical downscaling in short-range weather forecasts. Monthly Weather Review.123(9),2169-2183

[20] Ed. J.M. Cuadrat y J.M. Vide (2007). La Climatología española. Pasado, presente y futuro. Zaragoza: Prensas Universitarias de Zaragoza