Universidad de Valladolid

Facultad de Ciencias

Trabajo Fin de Grado

Grado en Estadística

Análisis de Datos Funcionales aplicado a datos de temperatura en España

Autor:

David Miguel Picón Llamas

Tutor:

Luis Ángel García Escudero

Julio 2019

Agradecimientos

A mi familia por apoyarme en todo momento,

A mis amigos por estar siempre ahí,

A los médicos y fisios por su excelente trabajo

Y a los profesores que me han ayudado.

Gracias.

Índice general

Resumen

El Análisis de Datos Funcional son técnicas estadísticas específicamente diseñadas para tratar conjuntos de datos creados mediante la observación de funciones o curvas. Estas técnicas nacen como una alternativa más eficiente al tratamiento estadístico multivariante de dichas funciones a partir de simples valores puntuales que toman estas funciones en momentos determinados. De esta forma, al cambiar del enfoque multivariante al funcional, se pueden extraer y estudiar sus características mejor.

A lo largo de este trabajo, se estudiarán una serie de procedimientos relacionadas con este enfoque, que irán desde la transformación de conjuntos de valores discretos en observaciones funcionales a la exploración de sus características. Los datos estudiados serán valores diarios de temperatura medidos por distintas estaciones meteorológicas en España en 2013, con el objetivo de extraer información sobre las características de los datos funcionales obtenidos.

Palabras clave: datos funcionales, suavizado, bases, temperaturas, funciones B-spline, componentes principales, phase-plane plot, R.

Abstract

Functional Data Analysis are statistical techniques specifically designed to treat data sets created by observing functions or curves. These techniques were born as a more efficient alternative to their multivariate statistical treatment, resulting from simple point evaluations at certain time periods. In this way, by shifting from the multivariate to the functional approach, their characteristics can be better extracted and studied.

Throughout this research work, we will study a set of techniques related to this functional approach, which involve all stages, from the transformation of sets of discrete values into functional observations, to the exploration of their main features. The data studied will be daily temperature values registered in 2013 by several Spanish meteorological stations, in order to obtain information about the characteristics of the functional data obtained.

Key words: functional data, smoothing, basis, temperature, B-spline functions, principal components, phase-plane plot, R.

Capítulo 1
Introducción

En muchas disciplinas donde se requiere Análisis de Datos, las mediciones obtenidas pueden ser susceptibles de ser consideradas más como curvas que como una respuesta escalar o un vector. Por ejemplo, este tipo de datos es común como resultado de monitorizar procesos o fenómenos en el tiempo. La toma de este tipo de datos se está realizando cada vez con más frecuencia, debido al informatizado y automatizado de las tecnologías aplicadas en la recogida de datos.

Aunque los datos subyacentes son curvas, los datos observados suelen ser una representación discretizada de estas. De esta manera, cada curva queda representada por un vector de dimensión finita. Sin embargo, es fácil ver que la aplicación de técnicas multivariantes estándar sobre estos vectores proporciona resultados imprecisos, poco informativos o de coste computacional prohibitivo, y, en determinados casos, el procedimiento es imposible de llevar a cabo. Es aquí donde aparecen al rescate una serie de adaptaciones del análisis de datos clásico que permiten trabajar con funciones.

El Análisis Funcional de Datos es una técnica relativamente novedosa, teniendo en cuenta la longevidad de la mayoría de herramientas y ramas de la Estadística. La primera vez que el término “Análisis Funcional de Datos” aparece en una publicación es en 1982, siendo el estadístico James O. Ramsay el autor. Sin embargo, Grenander (1950) y Rao (1958) ya habían sentado décadas antes las bases de este enfoque.

El hecho de que el objeto de estudio, que en este caso son funciones, sea de dimensionalidad infinita ha planteado un reto histórico desde los puntos de vista teórico y computacional. No obstante, una vez se han definidos las técnicas y herramientas, esa dimensionalidad infinita favorece un estudio más rico de las características de los datos.

Una revisión sistemática de las múltiples aplicaciones del Análisis de Datos Funcionales en diversas disciplinas puede encontrarse en Ullah and Finch (2013). Por ejemplo, estas técnicas han resultado relevantes en campos como la demografía, la medicina, las finanzas o la meteorología.

En este trabajo, estas técnicas de Análisis Funcional de Datos serán aplicadas a datos meteorológicos, concretamente relativos a las temperaturas en distintos puntos de España en el año 2013. En particular, se trabajará con curvas de temperatura tomadas con frecuencia diaria a lo largo de un año. Sobre estas curvas se usará la técnica de Componentes Principales Funcionales y otras técnicas como “phase-plane” plots.

Capítulo 2
Metodología

2.1 Análisis de datos funcionales

2.1.1 Conceptos básicos

Supongamos que partimos de n individuos para los que se cuenta con mediciones de una variable de interés, que se mide en cada individuo en J_i momentos del tiempo t_{ij}, para $j = 1, \dots, J_i$. De esta manera, para cada individuo i disponemos de J_i pares de números $(t_{ij}, y_{ij}), i = 1, \dots, n, y\ j = 1, \dots, J_i$. Este conjunto de datos proporciona una aproximación al comportamiento de una función subyacente $x_i(\cdot)$ que puede ser interesante determinar.

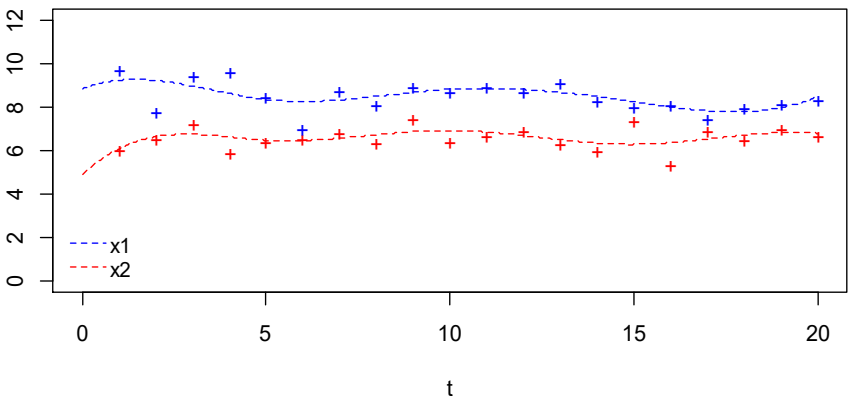

Figura 2.1: Observaciones y funciones subyacentes.

El estudio de la evolución de esos datos, en la mayoría de los casos, precisará de una herramienta más sofisticada que la simple vista de los datos como datos multivariantes en dimensión elevada. un conjunto de pares de datos. Es cierto que un conjunto grande de pares de valores (t_{ij}, y_{ij}) se corresponde con mayor información sobre el estado de la variable de interés estudiada en diferentes momentos. No obstante, al tratarse de un proceso continuo, por grande que sea el número de pares considerados, nunca se llegará a conocer del todo la forma de esas funciones. Además, nos encontraríamos con un problema de análisis multivariante de dimensiones en ocasiones inaceptables.

Parece claro entonces que el problema gira en torno a encontrar y estudiar la forma de una o varias funciones $x_i(\cdot)$ subyacentes a los datos, tantas como individuos, ajenas a alteraciones puntuales y que dan soporte a las observaciones, de manera que los pares de observaciones se asocian al estado de la función $x_i(\cdot)$ en momentos $t_1, \dots, t_{J_i}$ determinados y la posible existencia de algunos errores aleatorios en las mediciones de esta función.

Por otro lado, si lo que se pretende es estudiar y comparar las formas de evolución de estas funciones para los distintos individuos o sujetos, parece claro que podría ser útil disponer de información acerca de las derivadas de estas funciones. Que las derivadas sean de interés parece otra razón poderosa para pensar en esos conjuntos individuales de observaciones como observaciones funcionales, en vez de como observaciones vectoriales o multivariantes.

El hecho de que se disponga de registros del estado de una variable en el tiempo para un conjunto de individuos invita a la exploración de las fuentes principales de variación en los datos obtenidos. Es casi inevitable pensar que algún tipo de Análisis de Componentes Principales sería de utilidad, pero el uso de las técnicas de estudio de componentes principales multivariantes clásicas debe ser adaptado.

Por ejemplo, puede suceder que los datos hayan sido tomados en diferentes momentos del tiempo para cada individuo, o que los momentos de observación en el tiempo no estén equiespaciados. Ambos casos pueden suponer un problema de cara a ese estudio de componentes principales, por lo que habría que adaptar el procedimiento.

2.1.2 Objetivos

Los objetivos del Análisis de Datos Funcionales son esencialmente los mismos que los de cualquier otra rama de la Estadística. Se pueden destacar los siguientes:

- Representar los datos de una manera que faciliten su posterior análisis.
- Mostrar los datos de manera que se puedan resaltar varias características.
- Estudiar los más importantes patrones y fuentes de variación en los datos.
- Comparar dos o más conjuntos de datos en relación a ciertos tipos de variación, donde los conjuntos de datos son diferentes ejecuciones del mismo proceso llevadas a cabo por un mismo individuo, o la observación de la evolución del mismo proceso llevado a cabo por distintos sujetos.

Asimismo, el Análisis de Datos Funcionales comparte con otras técnicas estadísticas la característica de que, según su objetivo, puede ser exploratorio, confirmatorio o predictivo. En el primer caso, el objetivo fundamental es revelar aspectos interesantes

de los datos, sin realizar predicciones ni inferencias acerca del comportamiento de poblaciones mayores que la muestra de que se dispone. El análisis confirmatorio, por su parte, se basa en los datos de que se supone para responder cuestiones específicas y realizar contrastes sobre características de los datos. Finalmente, el análisis predictivo utiliza los datos para realizar afirmaciones acerca de otros de los que no se dispone, como pueden ser poblaciones mayores o momentos futuros.

2.1.3 Principios del Análisis de Datos Funcionales

Asumiendo que los datos discretos registrados para un sujeto *i* son un conjunto de J_i pares $(t_1, y_1), \dots, (t_{J_i}, y_{J_i})$, donde y_{ij} es el valor que toman los datos de la muestra i en el momento t_{ij}, la filosofía del análisis de datos funcionales se basa en considerar que cada observación i pasa a ser una función x_i que puede tomar valores en cualquier punto de un intervalo, de manera que a cada punto t le corresponde un valor $x_i(t)$. Es decir, los valores de y se corresponden con el estado de la función en el momento t.

El tiempo es la medida habitual de la variable independiente, aunque puede tratarse de otras, como las referidas a la posición en el espacio o el peso. Es posible que por ello en adelante, por una cuestión de brevedad, se hable de t como "el tiempo". Igualmente, salvo que se especifique lo contrario, en las siguientes explicaciones se estará trabajando con una única curva, de manera que no será necesario utilizar i para identificar la observación.

También, para simplificar la presentación, supondremos que los momentos del tiempo de evaluación son iguales para todos los individuos (y que se toma el mismo número de mediciones por individuo). Es decir, $t_{ij} = t_j$ para $j = 1, \dots, J$, independiente del individuo i.

2.2 Proceso de suavizado de una función

El suavizado es una técnica utilizada para tratar de eliminar el ruido o el error de medida de una función, y así revelar las características más importantes de una función.

2.2.1 Significado de suavizado

Si asumiéramos que los valores discretos de los que disponemos son *puros*, esto es, responden de manera estricta a un valor dado por la función x, es decir, $y_{ij} = x_i(t_j)$, el proceso elegido para hallar la función subyacente sería la interpolación. Sin embargo, este caso es extremadamente ideal, prácticamente imposible, ya que lo habitual es que los valores discretos hayan sufrido alguna clase de perturbación o "ruido". De manera analítica, siendo ε_{ij} el error, se entiende que:

$$y_{ij} = x_i(t_j) + \varepsilon_{ij}$$

En notación vectorial, más concisa y clara:

$$\boldsymbol{y} = x(\boldsymbol{t}) + \boldsymbol{\varepsilon}$$

Donde se suele asumir un modelo normal para el error; es decir, que $\boldsymbol{\varepsilon}$ sigue una distribución normal multivariante de media cero. Por ello, asumiremos que los valores observados responden a esa combinación de valor asociado a la función y al error. Se espera que este error, también llamado ruido, presente en todo momento un valor relativamente pequeño respecto al tamaño de la función.

2.2.2 Propiedades

2.2.2.1 Suavidad de una función

En muchos entornos es muy recomendable trabajar con funciones $x_i(t)$ que se suponen suaves vistas como funciones de t.

Que una función sea suave se entenderá por el hecho de que posea una o más derivadas, indicadas como $Dx, D^2x \ldots D^m x$ denota la derivada de orden m, y $D^m x(t)$ se corresponde con el valor de esa derivada en el punto t.

2.2.2.2 Resolución de los datos

La *resolución* o *tasa de muestra* de los datos discretos observados "en crudo", antes de comenzar el proceso de suavizado, es un factor decisivo a la hora de llevar a cabo un análisis funcional de los datos. Esta propiedad puede ser descrita como la densidad de los valores del argumento t en relación a la curvatura en las funciones subyacentes. Se trata de un factor importante, más concluyente que tener en cuenta simplemente el número J de evaluaciones del argumento.

La curvatura de una función x en un argumento t suele venir medida por el valor absoluto de la derivada segunda de la función en ese punto, $|D^2x(t)|$. Donde la curvatura es alta, es vital tener un número suficiente de puntos de cara a estimar la función de una manera eficaz. De no ser así, es posible elegir argumentos donde la función no refleja puntos característicos de curvatura, como puede ser el caso de picos o ascensos o descensos pronunciados. Si, por el contrario, la curva subyacente es suave, puede valer con un número más pequeño de evaluaciones.

2.3 Representación de la función mediante funciones base

2.3.1 Principios

Un sistema de funciones base es un conjunto de funciones conocidas ϕ_k matemáticamente independientes entre sí, y que tienen la propiedad de que permiten aproximar arbitrariamente bien cualquier función tomando una suma ponderada o combinación lineal de un número K suficientemente grande de estas funciones.

Así, se considera que un conjunto de K funciones base ϕ_k aproximará a una función x partiendo de la expansión lineal

$$\hat{x}(t) = \sum_{k=1}^{K} c_k \phi_k(t)$$

Siendo $\boldsymbol{c}$ el vector de longitud K con los coeficientes c_k, y $\boldsymbol{\Phi}$ el vector funcional compuesto por las funciones base ϕ_k, la fórmula anterior puede expresarse en notación compacta como

$$\hat{x} = \boldsymbol{c}' \boldsymbol{\Phi} .$$

Los métodos de expansión en bases proporcionan una representación finito-dimensional del problema infinito-dimensional tratado en el Análisis de Datos Funcionales. Siguiendo este razonamiento, uno podría pensar que el análisis de datos funcionales se reduce a un mero análisis multivariante, pero no es así, ya que gran parte de la eficacia y forma de la expansión depende de cómo se elige el sistema de bases $\boldsymbol{\Phi}$.

Si $K = n$, se daría la situación de ajuste extremo a los datos que se correspondería con la interpolación, ya que se podrían elegir los coeficientes c_k de modo que $x(t_j) = y_j$ para cada j. Por lo tanto, podemos considerar que el grado de suavizado (y en consecuencia cuánto de lejos está la expansión en bases utilizada de la interpolación) viene determinado por K, el número elegido de funciones base.

Es fácil deducir en consecuencia que la elección de K y el tipo de bases elegido dependen de las características de los datos que se quieren estimar. Elegir un número grande de bases proporciona un ajuste mayor, pero, entre otros problemas, favorece el sobreajuste y requiere de un coste computacional mayor. Sin embargo, cuanto, dentro de unos límites, menor sea K y mejor pueda reflejar el conjunto de bases elegido ciertas características de los datos:

- Se requerirá un coste computacional menor.
- Más grados de libertad estarán disponibles de cara a contrastar hipótesis y calcular intervalos de confianza al realizar inferencias.
- Será más probable que los propios coeficientes en sí proporcionen información relevante acerca de los datos.

La elección del sistema de bases puede ser importante en muchos problemas también de cara a aproximar las derivadas de la función:

$$D\hat{x}(t) = \sum_{k=1}^{K} c_k D\phi_k(t) = \boldsymbol{c'D}\Phi(t)$$

Es posible elegir bases que funcionan bien a la hora de estimar una función, pero hagan lo propio de manera pobre para la estimación de sus derivadas. Esto es debido a que una representación acertada de las observaciones puede forzar a que $\hat{x}$ tenga pequeñas oscilaciones que se repitan para ajustarse a esos datos, lo que conlleva consecuencias terribles para el ajuste a las derivadas.

Por otro lado, no hay un tipo de bases universal en el sentido de que pueda ajustarse de manera razonable a cualquier tipo de datos; al contrario, hay una serie de tipos de bases que, por sus características, son más apropiadas para conjuntos de datos que presentan un comportamiento específico u otro.

2.3.2 Sistemas de bases más comunes

Muchos Análisis de Datos Funcionales alcanzan sus mejores resultados con la presencia de bases de Fourier, en el caso de los datos periódicos, o con un sistema de bases B-spline en el caso de los datos que no presentan periodicidad. En las situaciones en las que las derivadas no van a ser tenidas en cuenta, los sistemas de bases de *wavelets* aparecen con frecuencia. Por último, las bases polinómicas carecen por lo general de protagonismo en esta clase de estudios.

2.3.2.1 Sistemas de bases de Fourier

Se trata del sistema de bases más utilizado cuando los datos presentan periodicidad, debido a su característica forma sinusoidal periódica. Esta expansión en bases se fundamenta en una serie de Fourier:

$$\hat{x}(t) = c_0 + c_1 sin(\omega t) + c_2 cos(\omega t) + c_3 sin(2\omega t) + c_4 cos(2\omega t) + \ldots$$

Esta expansión viene definida por las bases $\phi_0(t) = 1$, $\phi_{2r-1}(t) = sin(r\omega t)$, y $\phi_{2r}(t) = cos(r\omega t)$. Al tratarse de una base periódica, el parámetro ω determina el periodo $2\pi/\omega$. A continuación, en la imagen X se presentan las bases utilizadas en una expansión de tres bases de Fourier para aproximar una función de periodo 1.

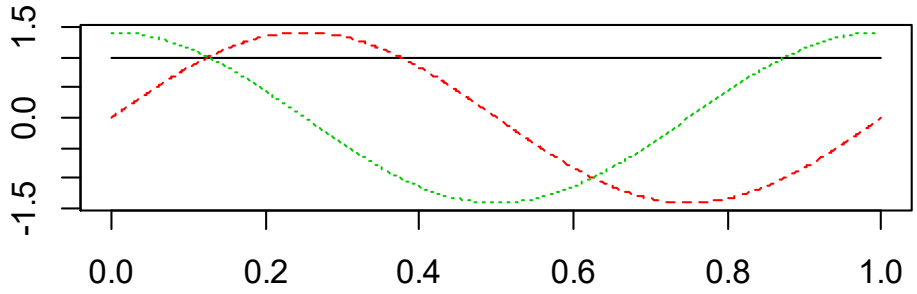

Figura 2.2: Sistema de tres bases de Fourier en el intervalo [0,1].

Cuando J es una potencia de 2 y los argumentos están igualmente espaciados, es fácil calcular todos los coeficientes c_k de una manera extremadamente eficiente. Por ello, durante mucho tiempo fueron el sistema de bases elegido en la mayoría de los casos, pero desde la aparición de los sistemas de bases de wavelets y B-splines, que igualan e incluso mejoran esta eficiencia computacional, su protagonismo ha quedado relegado a los casos en los que los datos presentan una clara periodicidad.

Otro aspecto que hace que este sistema de bases sea idóneo cuando los datos presentan un comportamiento periódico es que la estimación de las sucesivas derivadas en un sistema de bases de Fourier es simple, ya que:

$$Dsin(r\omega t) = r\omega cos(r\omega t)$$

$$Dcos(r\omega t) = -r\omega sin(r\omega t),$$

lo cual es muy útil cuando se busca aproximar el comportamiento de la velocidad o la aceleración con la que se desarrolla la función.

Una serie de Fourier es especialmente útil para funciones periódicas y estables, en el sentido de que la curvatura es del mismo orden en todo momento, y no presentan características locales que no respeten la periodicidad.

2.3.2.2 Sistemas de bases de tipo spline

Las funciones spline son el sistema de bases más utilizado para aproximar funciones o datos no periódicos, pero que sí presentan suavidad. Debido a su funcionamiento, trabajan con la eficiencia computacional de los sistemas de bases polinómicas, pero presentan una flexibilidad superior.

Funciones spline y grados de libertad

El funcionamiento de un spline queda ilustrado en la imagen X, donde se utilizan tres funciones spline para aproximar la función $sin(t)$ a lo largo del intervalo [0,2π] en los tres paneles de la izquierda, mientras que en los de la derecha se hace lo propio con su derivada, $cos(t)$.

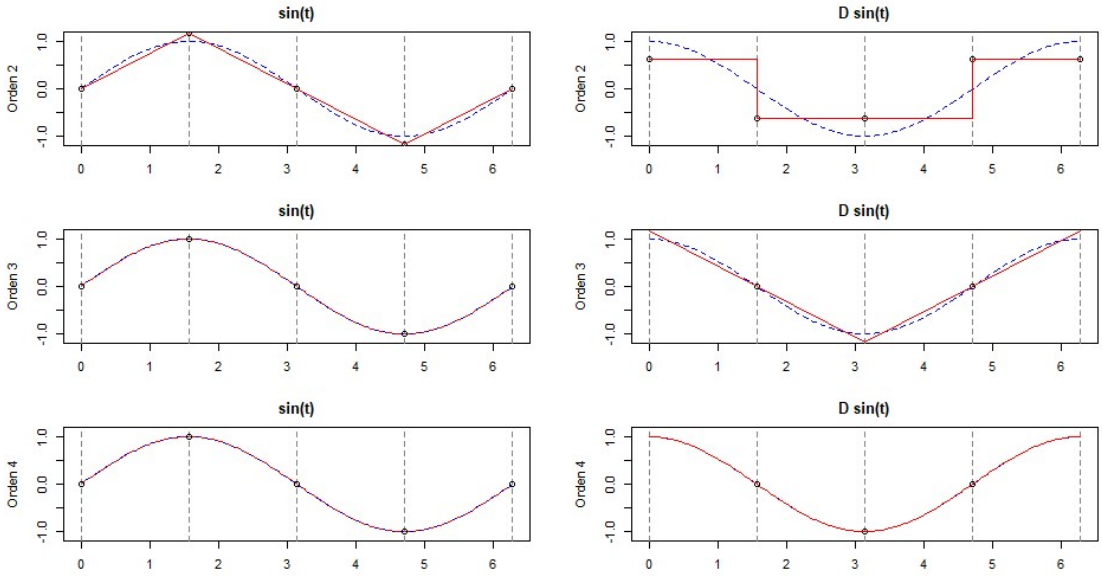

Figura 2.3: Estimaciones de la función sin(t) y su derivada mediante splines de distinto grado.

El primer paso para definir un spline es dividir el intervalo a lo largo del cual se va a aproximar la función en L subintervalos separados por valores $T_l, l = 1,\ldots, L - 1$ que reciben el nombre de *nodos* o *puntos de corte.*

Dentro de cada uno de los intervalos generados, un spline es un polinomio de un orden m especificado, que se corresponde con el número de constantes necesarias para definir el spline, que a su vez es el grado de la curva más uno.

Los polinomios de intervalos adyacentes se acoplan o unen de manera suave en un punto de la frontera que separa ambos intervalos, de manera que los valores de sus funciones encuentran la restricción de que deben coincidir en ese punto. Esta condición deberá cumplirse en todas las derivadas hasta las de orden $m - 2$.

Es por esta limitación que el número de grados de libertad del ajuste viene dado por la suma de los grados de los polinomios de sus distintos intervalos menos el total de fronteras entre intervalos.

Un spline ganará flexibilidad cuantos más nodos presente, ya que eso le permitirá adoptar un mayor número de comportamientos a lo largo del intervalo en el que opera. Es fácil deducir que, en caso de no haber nodos interiores, el spline pierde las ventajas de flexibilidad respecto a los sistemas de bases polinómicas porque, precisamente, pasa a ser un sistema de bases de este tipo.

En resumen, una función de tipo spline viene determinada por dos elementos, que son la secuencia de nodos elegida, y el orden de los segmentos polinómicos de los intervalos generados por ella.

La base B-spline para funciones de tipo spline

Una vez definida la función spline, es momento de explicar cómo se construye una de estas bases. Para ello, se especifica un conjunto de funciones base $\phi_k(t)$ que cumplen una serie de propiedades esenciales:

- Cada función base $\phi_k(t)$ es una función spline en sí, definida por un orden m y una secuencia de nodos τ.
- Dado que cualquier múltiplo de una función spline es una función de este tipo, así como la suma o diferencia de funciones spline, cualquier combinación lineal de estas funciones base es una función spline.
- Cualquier función spline definida por m y τ puede expresarse como combinación lineal de estas funciones base.

Hay muchas maneras de construir un sistema de este tipo, pero el sistema de bases B-spline, ideado por De Boor en 2001, es el más popular debido a una serie de características entre las que destacan las computacionales. Además, es un sistema disponible en muchos lenguajes de programación, entre ellos R.

La siguiente imagen muestra las trece funciones B-spline utilizadas en un sistema de bases spline de orden cuatro, definido por nueve puntos frontera o nodos equiespaciados. Es fácil ver que cada una de las siete funciones base centrales son solo estrictamente positivas en los cuatro subintervalos adyacentes. Como los splines cúbicos tienen dos derivadas continuas, cada función base hace una transición suave a las regiones donde su valor es superior a cero. Estas splines centrales tiene la misma forma debido al igual espaciado de los nodos.

Por su parte, las tres funciones base situadas más cerca del extremo izquierdo del intervalo, así como las ubicadas en el extremo derecho, mantienen la propiedad de que nunca tienen un valor por debajo de cero, pero ya no tienen valor estrictamente positivo en cuatro regiones.

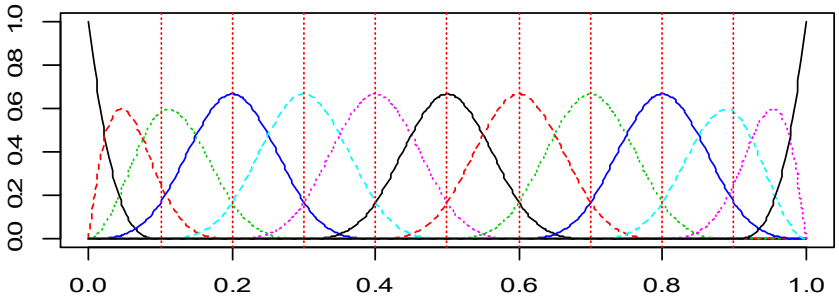

Figura 2.4: Sistema de trece bases B-spline en el intervalo [0,1].

Es más, mientras que en los puntos más interiores del intervalo los splines hacen la misma transición suave y dos veces diferenciable, las splines ubicadas en los extremos del intervalo presentarán discontinuidad, mientras que las siguientes conforme se avanza al centro del intervalo serán meramente continuas. Por último, las splines número tres y once serán una vez diferenciables. Esta pérdida de diferenciabilidad en los extremos tiene sentido, ya que la mayoría de las veces no se dispondrá de información acerca del comportamiento de la función fuera del intervalo.

La propiedad por la cual una función base B-spline de grado m nunca es estrictamente positiva en más de *m* intervalos obligatoriamente adyacentes recibe el nombre de "*propiedad de soporte compacto"*, y es una de las causas de la eficiencia computacional de este sistema spline. Si hay K funciones base de tipo B-spline, entonces la matriz de orden K de productos internos de estas funciones solo tendrá valores distintos de cero en su diagonal principal y en como mucho $m - 1$ subdiagonales por encima y debajo de esta. Esto es, por grande que sea K, el cálculo de la función spline puede organizarse de manera que aumente de manera únicamente lineal conforme lo hace el valor de K.

Para compensar la diferencia de densidad de los splines en los extremos, se añaden m nodos en al principio y al final del intervalo. Es decir, una vez ya se han calculado las B-splines, la secuencia de nodos τ se extiende en ambos extremos añadiendo m -1 réplicas del valor de la función en esos puntos.

La notación $B_k(t, \tau)$ es usada para indicar el valor en el punto t de la función generada por el sistema de funciones base B-spline con secuencia de nodos τ. Aquí, k hace referencia al número del nodo ubicado en *t*, o el más cercano a t por su izquierda. Los $m-1$ nodos añadidos en los extremos entran dentro de este cálculo, de manera que en el cálculo entran $m + L - 1$ funciones base. De acuerdo con esta notación, una función spline $S(t)$ con nodos interiores es definida por:

$$s(t) = \sum_{k=1}^{m+L-1} c_k B_k(t, \tau)$$

Respecto a la estrategia del reparto de la secuencia de nodos τ, la mayoría de las aplicaciones programadas en R y otros lenguajes utilizan por defecto el equiespaciado, pero cuando la concentración de puntos es diferentes en distintas zonas del intervalo, puede ser más conventiente colocar un nodo cada j observaciones (cuantiles). Por último, cuando la curvatura es diferente en distintas regiones del intervalo, se tiende a colocar más nodos en zonas donde la curvatura es mayor.

Una característica sorprendente de los sistemas de bases spline es que el incremento del número de bases K no siempre mejora el ajuste a los datos. Esto sucede porque, una vez el orden de un spline está fijado, el espacio de funciones definido por K B-splines no está necesariamente contenido dentro del definido por $K + 1$ B-splines. Esto se debe a que el nuevo espaciado de los nodos puede condicionar un ajuste peor a los mismos datos, de modo que, en ciertos casos, un sistema de B-splines de menor dimensión puede proporcionar un ajuste mejor que otro donde K es superior.

2.3.2.3 Otros tipos de base

Wavelets

Se puede construir una base ortogonal para cualquier curva eligiendo una función madre wavelet apropiada ψ, y luego considerando las posibles amplitudes y traslaciones de la forma

$$\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k).$$

Se trata de un tipo de funciones base que, a diferencia de las basadas en series de Fourier, funciona bien cuando los datos o funciones presentan discontinuidades o

cambios rápidos de comportamiento, pero que en otras circunstancias no es tan eficiente como la expansión en bases de funciones B-spline.

Bases exponenciales y de potencia

Como su propio nombre indica, el sistema de bases exponenciales se compone de una serie de funciones exponenciales,

$$e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_k t}, \ldots$$

Donde todos lo parámetros λ_k son diferentes, y habitualmente $\lambda_1 = 0$.

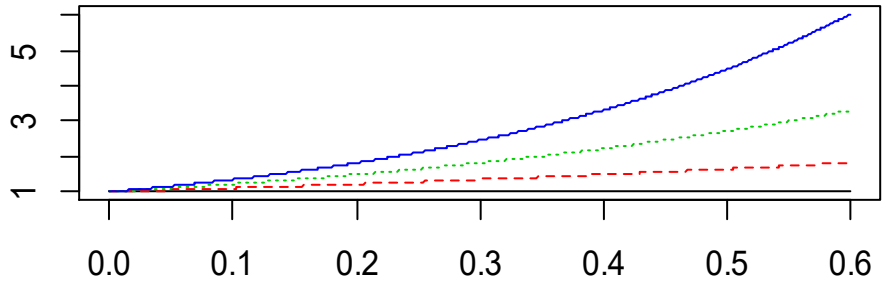

Figura 2.5: Sistema de cuatro bases exponenciales en el intervalo [0,0.6].

Este parámetro también aparece en los sistemas de base de potencias:

$$t^{\lambda_1}, t^{\lambda_2}, \ldots, t^{\lambda_k}, \ldots$$

que tienen importancia en problemas donde t es estrictamente positivo, de manera que pueden utilizarse valores de λ_k negativos.

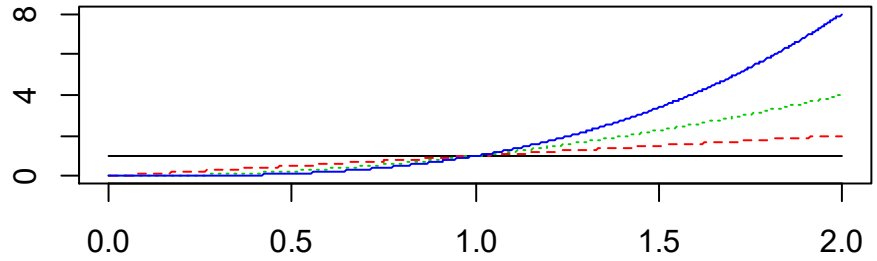

Figura 2.6: Sistema de cuatro bases de potencias en el intervalo [0,2].

Bases polinómicas (o sistemas de expansión en bases monómicas)

Los sistemas de extensión en bases monómicas $\phi_k(t) = (t - \omega)^k, k = 0, \ldots, K$ aparecen en algunos casos, donde ω es un parámetro de desplazamiento que suele elegirse de manera que coincida con el centro del intervalo de aproximación.

La siguiente imagen es un ejemplo de sistema de cinco bases monómicas simples centradas (ω = 0), donde cada curva se corresponde con el argumento t elevado al valor k = 0,…,4.

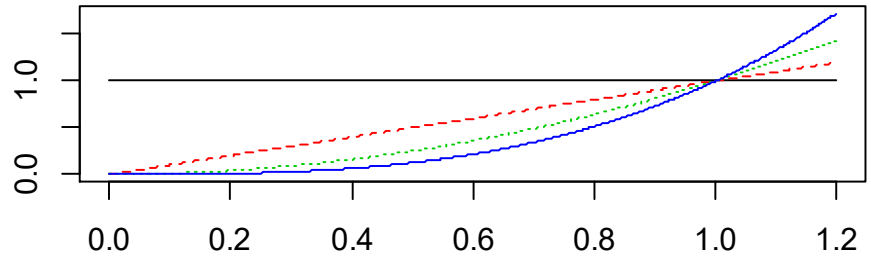

Figura 2.7: Sistema de cuatro bases monómicas en el intervalo [0,1.2].

Como sucede con las series de Fourier, las bases polinómicas no suelen poder ajustarse a características locales de los datos sin requerir de un número K de bases alto, con el coste computacional que eso conlleva. Además, en los extremos del intervalo tienden a ofrecer un funcionamiento muy pobre, y consitituyen un sistema de bases inapropiado para la extrapolación o la predicción.

2.3.3 Suavizado de los datos funcionales usando mínimos cuadrados

De entre las técnicas de suavizado de datos funcionales, la más utilizada es la basada en la estimación por mínimos cuadrados. Es la más comúnmente utilizada en los lenguajes de programación, y es la que utiliza R en la función utilizada para este fin, smooth.basis, que crea un objetio de tipo dato funcional. Este enfoque aplica principios del análisis de regresión múltiple.

Para explicar su funcionamiento, es importante recordar que el objetivo es aproximar la función x que da soporte a las observacione y_j, siguiendo el modelo

$$y_j = x(t_j) + \varepsilon_j, j = 1, \dots, J,$$

y que estamos usando una expansión en bases funcionales para $x(t)$ de la forma:

$$\hat{x}(t) = \sum_{k}^{K} c_k \phi_k(t) = \boldsymbol{c'\Phi}$$

donde el vector $\boldsymbol{c}$ de longitud K contiene los coeficientes c_k. Sea $\boldsymbol{\Phi}$ una matriz de tamaño J x K que contiene los valores $\phi_k(t_j)$.

El criterio de mínimos cuadrados que hay que minimizar es el siguiente:

$$SMSSE(\boldsymbol{y}|\boldsymbol{c}) = \sum_{j=1}^{J} [y_j - \sum_{k=1}^{K} c_k \phi_k(t_j)]^2$$

Que, expresado de una manera más concisa en notación compacta, es

$$SMSSE(\boldsymbol{y}|\boldsymbol{c}) = (\boldsymbol{y} - \boldsymbol{\Phi c})'(\boldsymbol{y} - \boldsymbol{\Phi c}).$$

Para realizar esta minimización podemos derivar e igualar a 0:

$$2\boldsymbol{\Phi\Phi}'\boldsymbol{c} - 2\boldsymbol{\Phi}'\boldsymbol{y} = 0.$$

De esta manera, se obtiene el estimador $\hat{\boldsymbol{c}}$ que minimiza el criterio de mínimos cuadrados:

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Phi\Phi})'^{-1}\boldsymbol{\Phi}'\boldsymbol{y}.$$

Y, en consecuencia, el vector $\hat{\boldsymbol{y}}$ de valores ajustados será

$$\hat{\boldsymbol{y}} = \boldsymbol{\Phi}\hat{\boldsymbol{c}} = \boldsymbol{\Phi}(\boldsymbol{\Phi\Phi})'^{-1}\boldsymbol{\Phi}'\boldsymbol{y}.$$

Recordemos que esta aproximación por mínimos cuadrados es apropiada en situaciones en las que se asume un modelo estándar para el error. Es decir, que los residuos ε_j sobre la curva real están distribuidos de manera independiente e idéntica, y que tienen media cero y varianza constante σ^2.

En caso de no cumplirse este modelo estándar, puede recurrirse a llevar a cabo un ajuste por mínimos cuadrados con pesos, que tienen en cuenta la estimación de la matriz de varianzas-covarianzas de los residuos ε_{ij}.

En este caso, el criterio de mínimos cuadrados a minimizar incoropora una matriz semidefinida positiva $\boldsymbol{W}$ que permite trabajar con estructuras de varianzas-cpvarianzas más complejas en los residuales.

$$SMSSE(\boldsymbol{y}|\boldsymbol{c}) = (\boldsymbol{y} - \boldsymbol{\Phi c})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{\Phi c}),$$

de manera que la estimación del vector $\boldsymbol{c}$ será

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Phi W\Phi})'^{-1}\boldsymbol{\Phi W y}.$$

Sin embargo, el valor y la importancia de $\boldsymbol{W}$ parten del conocimiento o la sospecha de fenómenos asociados a la distribución de los residuos en los datos que no están siempre disponibles.

2.3.4 Elección del número de funciones base

La elección del número de funciones base K, también llamado orden de la expansión en funciones base, es un factor vital en el suavizado de los datos. Un valor mayor de K proporcionará en general un mejor ajuste a los datos, hasta el punto extremo de que cuando $K = n$ la curva generada proporcionará un ajuste perfecto. Sin embargo, ello conlleva que, cuanto mayor sea K, mayor riesgo se corre de ajustar ruido o variación, que es lo que se pretende ignorar. Por otro lado, si escogemos un valor demasiado pequeño para K, es posible que se pierdan características importantes de la función suavizada que se pretende estimar.

Existen una serie de procedimientos basados en estadísticos de error y criterios que incluyen términos de penalización que ayudan a elegir el número de bases; sin embargo, requieren del conocimiento previo de valores y características de los datos, de los que,

en casos como el nuestro, no se dispone. Por otro lado, la elección del número de bases, aun cuando viene apoyada por estos métodos, es en esencia un procedimiento donde la decisión se toma comparando visualmente el ajuste de la curva estimada a los datos para una serie de valores de K, sabiendo que el objetivo es dar con una función suave que a la vez refleje las características de la curva.

2.4 Técnicas aplicadas al Análisis de Datos Funcionales

En este apartado comentaremos las técnicas de análisis de datos funcionales que han sido utilizadas en este trabajo.

2.4.1. Análisis de Componentes Principales Funcionales

2.4.1.1 Análisis de Componentes Principales clásico

La idea que fundamenta el Análisis de Componentes Principales es encontrar un conjunto de vectores ortogonales que expliquen de la manera más eficiente posible la variabilidad de los datos. Dicho de otro modo, se busca reducir la dimensión de ese conjunto de datos conservando al a vez la estructura de la variabilidad presente en ellos tanto como sea posible.

El Análisis de Componentes Principales clásico parte de la idea de buscar combinaciones lineales adecuadas de los valores de las variables medidas:

$$f_i = \sum_{j=1}^{J} \beta_j x_{ij}\,, i = 1, \dots, n,$$

donde β_j es un coeficiente de peso (*loadings* o cargas) asociado a los valores observados x_{ij} de la variable j-ésima. En notación compacta, más fácilmente interpretable, la ecuación anterior se expresa como:

$$f_i = \beta' x_i, i = 1, \dots, n,$$

donde β es el vector $(\beta_1, \dots, \beta_J)$' y x_i es el vector $(x_{i1}, \dots, x_{iJ})$'.

En el caso multivariante, el objetivo es elegir esos pesos de manera que se se destacan los tipos de variación de los datos más presentes en los datos. Con este fin se realiza el siguiente procedimiento, que define conjuntos de pesos normalizados que maximizan la variación en los f_i:

1. Calcular el vector de pesos $\boldsymbol{\xi}_1 = (\xi_{11}, \dots, \xi_{p1})'$, para el cual los valores de la combinación lineal

$$f_{i1} = \sum_{j=1}^{J} \xi_{j1} x_{ij} = \boldsymbol{\xi}_1{}' x_i$$

tienen la mayor variabilidad cuadrática media

$$n^{-1}\sum_{i=1}^{n} f_{i1}^2$$

respetando la restricción

$$\sum_{j=1}^{J} \xi_{j1}^2 = ||\boldsymbol{\xi}_1||^2 = 1$$

2. Repetir el procedimiento en sucesivos pasos. Se puede hacer hasta tantas veces como número de variables p tenga el problema. En el paso m-ésimo, se calcula un nuevo vector de pesos $\boldsymbol{\xi}_m$ con componentes ξ_{jm}, así como nuevos valores $f_{im} = \boldsymbol{\xi}_m{}' x_i$. En cada una de estas etapas, los valores f_{im} tienen media cuadrática máxima, repetando otra vez la restricción de unidad de suma de cuadrados $||\boldsymbol{\xi}_m||^2 = 1$ y las $m - 1$ restricciones adicionales de ortogonalidad:

$$\sum_{j=1}^{J} \xi_{jk}\xi_{jm} = \xi_k\xi_m = 0, k < m$$

En el primer paso, maximizando la variabilidad cuadrática media, se identifica la forma más importante de variación en las variables. La restricción de la suma unitaria de cuadrados de los pesos evita que las medias cuadráticas de los valores de la combinación lineal tomen valores arbitrariamente grandes, en cuyo caso el análisis perdería su sentido.

En los sucesivos pasos se vuelven a buscan los modos más importantes de variación, pero los respectivos pesos tienen que ser ortogonales a los identificados previamente, para así indicar modos de variación completamete nuevos, sin ninguna relación con los anteriormente encontrados.

Obviamente, la cantidad de variación medida en términos de la media cuadrática decrecerá con cada paso, lo que significa que la importancia de los modos de variación encontrados será cada vez menor, hasta llegar a un punto en el que los modos obtenidos ya no sean de interés, que como mucho se corresponderá con el paso p.

Los valores de las combinaciones lineales f_{im} reciben el nombre de *scores de las componentes principales*, y son importantes a la hora de decribir cuáles de los modos de variación son importantes y hasta qué punto en cada caso específico.

En el caso de los datos funcionales podría recurrirse con este fin a estudiar las matrices de varianzas-covarianzas y correlaciones en los datos “en crudo”, pero, al igual que en el caso multivariante en dimensiones altas, no se obtienen resultados tan esclarecedores o fácilmente interpretables. El Análisis de Componentes Principales funcionales resulta ser una herramienta mucho más eficiente e interpretable.

2.4.1.2 Extensión a datos funcionales

En el caso funcional, el Análisis de Componentes Principales va a funcionar de manera análoga. Las observaciones objeto de estudio pasan a ser valores funcionales $x_i(t)$, de manera que el índice continuo t sustituye al discreto j utilizado en el contexto multivariante. Partiendo de que β y x ahora son funciones $\beta(s)$ y $x(s)$, el producto interno queda sustituido por la integración:

$$\int \beta x = \int \beta(s)x(s)\, ds.$$

Análogamente, el cálculo de los scores correspondientes al peso β pasa a ser

$$f_i = \int \beta(s)x_i(s)\, ds.$$

de ahora en adelante abreviado como $\int \beta x_i$.

En el primer paso del Análisis de Componentes Principales funcionales, se elegirá la función de peso $\xi_1(s)$ de manera que maximice la respectiva media cuadrática

$$\frac{\sum_i f_{i1}^2}{n} = \frac{\sum_i (\int \xi_1 x_i)^2}{n}$$

a la vez que se respete la versión continua de la restricción de la suma de cuadrados unitaria, $\int \xi_1(s)^2 ds = 1$, abreviada de ahora en adelante como $||\xi_1||^2 = 1$.

Al igual que en el ACP multivariante, también se deberá cumplir que, en los pasos sucesivos, la función de peso ξ_m cumpla la(s) restricción(es) de ortogonalidad $\int \xi_k \xi_m$ = 0, $k < m$. Cada función de peso tiene la tarea de definir el modo más importante de variación en las curvas dentro del subespacio de los ortogonales a los modos definidos en los pasos previos.

La solución en este problema, desde este enfoque funcional, puede consultarse en Ramsay y Silverman (2005). El problema se vuelve a reducir a una búsqueda de autovectores y autovalores donde los productos internos de los elementos de la base funcional considerada son también tendiso en cuenta. Por tanto, la solución del Análisis de Componentes Principales Funcional depende de la base funcional fijada. Se ha usado la implementación en R de este procedimiento mediente la librería fda.

Como ejemplo, a continuación se muestran los pesos correspondientes a las tres primeras componentes principales, una vez aplicado todo el proceso, para el conjunto de datos "*gait".* Se trata de un dataset incluido en la librería fda de R, que registra la evolución del ángulo de la cadera de 39 chicos a lo largo de 20 puntos temporales equiespaciados.

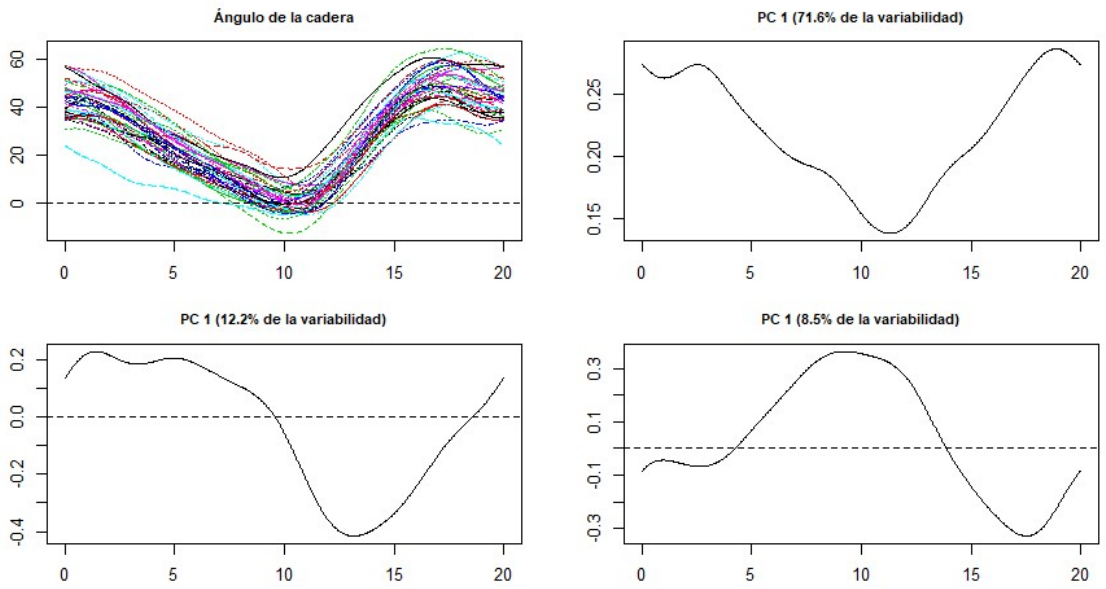

Figura 2.8: Representación de las curvas de ángulo del movimiento de la cadera (arriba a la izquierda) y de las funciones de peso de las tres primeras componenes principales.

2.4.1.3 Visualización de los resultados

Representación de las componentes como una perturbación sobre la media

Es posible que la representación anterior, en muchos casos, no sea fácilmente interpretable, aún conociendo el conjunto de datos con el que se trabaja. Un método que favorece la interpretación es representar tantas componentes que se estimen oportunas como perturbaciones de la media; esto es, como la función resultante de sumar o restar a la función media del conjunto un múltiplo conveniente de la componente.

La imagen 2.9 muestra esta manera de visualizar el efecto de las componentes para el ejemplo anterior. En cada caso, la curva continua representa la función media, y otras curvas son el resultado de sumar (+) o restar (-) a la función media el múltiplo de la función componente principal respectiva.

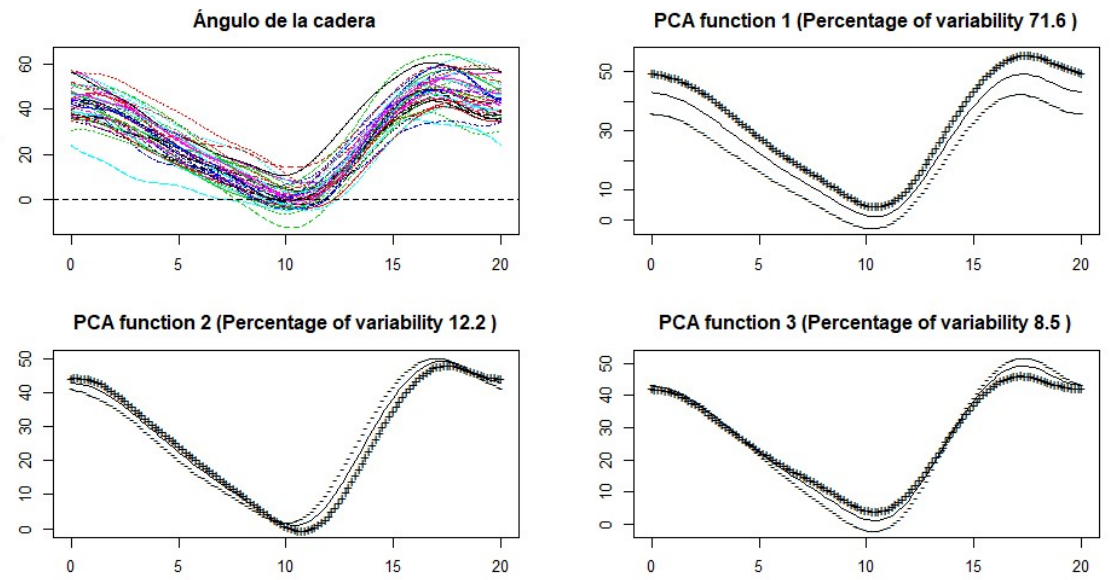

Figura 2.9: Representación de las curvas de ángulo del movimiento de la cadera (arriba a la izquierda) y de las componentes principales como perturbaciones de la media.

En el caso de este ejemplo, es fácil ver que la primera componente se corresponde con un incremento global en el ángulo del movimiento, mientras que el efecto de la segunda es un retraso en el proceso. Un mayor valor positivo de la tercera componente implica una mayor estabilidad en la curva.

Mapa de scores de las observaciones

Un aspecto importante del Análisis de Componentes Principales, tanto en el caso multivariante como en el funcional, es examinar las posiciones de los scores en cada componente.

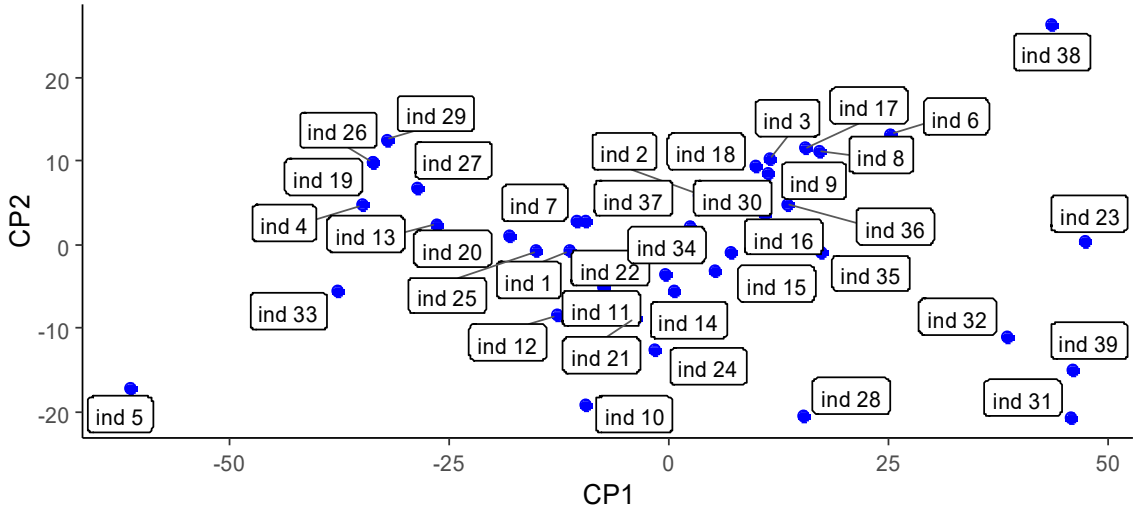

Figura 2.10: Mapa de scores del conjunto de funciones de ángulo de la cadera.

Estas scores, también llamadas "*puntuaciones en las componentes principales*", representan cómo de fuerte es en cada función observada la variación característica correspondiente a cada componente. Por ejemplo, se puede observar que los individuos 23, 31, 32, 38 y 39 presentan un valor positivo fuerte en el score de la primera componente principal, mientras que el sujeto número 5 tiene un valor altamente negativo. Los primeros individuos presentarán una mayor movilidad o flexibilidad, mientras que el sujeto número cinco tendrá una mayor rigidez de movimientos.

A continuación se representan las curvas, destacando, por un lado, la correspondiente al individuo 5 (en rojo), y por otro las de los sujetos 23, 31, 32, 38 y 39. De esta manera, se confirma la información del mapa de scores sobre las funciones.

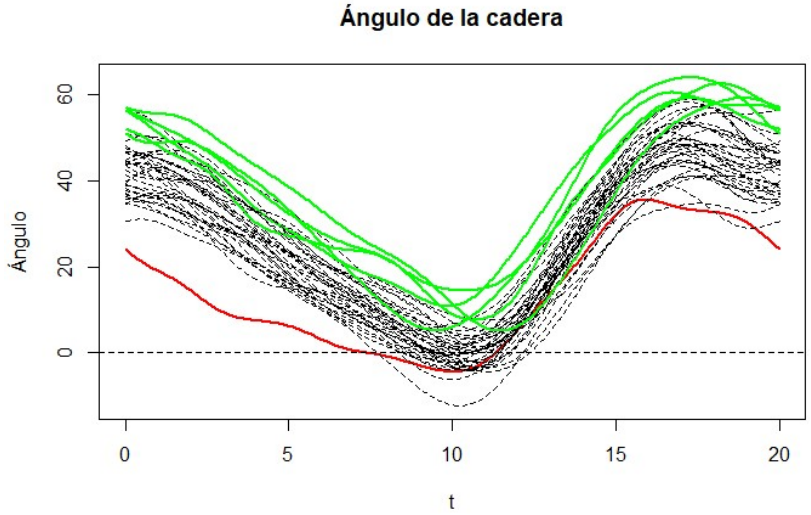

Figura 2.11: Representación de las curvas 5 (en rojo) y 23,31,32, 38 y 39 (en verde).

Las scores funcionan de manera análoga para la otra componente, y se puede interpretar el mapa como un plano de coordenadas dividido en cuatro cuadrantes, en cada uno de los cuales la combinación de valores positivos y negativos para las componentes es distinta, y con ello su comportmiento respecto a un modo de variabilidad o comportamiento específico de la curva.

2.4.2 Phase-plane plot

En muchos de los casos en los que el estudio de unos datos recurre al Análisis de Datos Funcionales, suele ser importante estudiar derivadas de la función subyacente obtenida. En concreto, las derivadas más importantes suelen ser la primera y la segunda. La primera derivada de la función representa la velocidad con la que cambia en cada momento, mientras que la segunda derivada se corresponde con la aceleración o energía acumulada.

Es interesante ver cómo se relacionan ambas derivadas, y en qué momento se producen los cambios en las tendencias. Esto se ve representado de manera gráfica en el *phase-plane* plot, que se corresponde con la curva que enfrenta ambas variables. Este gráfico

estudia la interrelación entre ambas derivadas. Si se trata de datos periódicos, se puede valorar tanto el *phase-plane plot medio* como ver cómo evoluciona la forma del mismo cada vez que se cumple el periodo.

Es importante ser consciente de que hay factores que afectan a la curva en diferentes momentos y de diferentes maneras, en cuanto a cadencia e intensidad. El phase-plane plot, en este tipo de casos, puede ayudar a entender o profundizar sobre estas características u otras que a priori podrían pasar desapercibidas.

Para entender el funcionamiento de este gráfico, es vital saber entender la relación entre ambas derivadas, y cómo se puede entender el proceso como un "intercambio de energía". Para ello, vamos a partir del phase-plane plot de la función $\sin(2\pi t)$, mostrado en la figura 2.12..

Se trata de un ejemplo bastante recurrido y muy esclarecedor acerca de las características y el significado de este gráfico. Esta función describe un *proceso armónico básico*, que describiría el movimiento de la posición vertical de un objeto suspendido al final de un muelle, empezando en la posición cero cuando *t* = 0. Es conveniente recordar que este experimento tendría que tener lugar en condiciones ideales, así que no se considera que intervengan más fuerzas o factores a mayores de los ya nombrados.

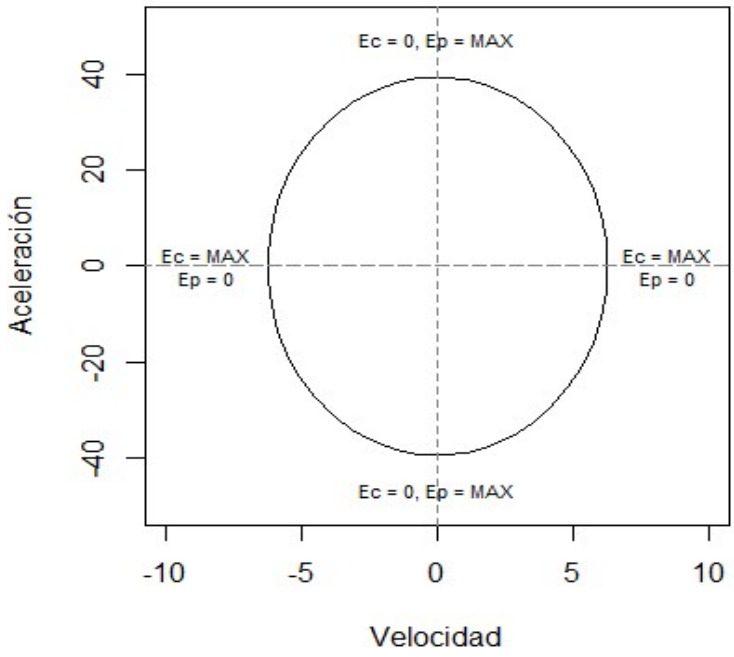

Figura 2.12: Phase-plane plot del movimiento armónico simple unitario.

En este caso, el muelle oscila porque hay un intercambio de energía entre dos estados: *potencial* y *cinético*. Cuando $t = \pi \pm 2\pi i,$ donde i es un número natural, el muelle está en uno u otro extremo de su trayectoria, donde su energía cinética es 0 y su energía potencial es máxima. Como la fuerza es proporcional a la aceleración, la segunda derivada de la posición del muelle, $-(2\pi)^2 \sin(2\pi t),$ alcanza su máximo. Por otro lado,cuando el muelle pasa por la posición 0, su velocidad es máxima, y con ello su energía cinética, mientras que la aceleración es 0.

El procedimiento se mueve entre momentos de gran energía cinética y otros de gran energía potencial. Este razonamiento es fácilmente extrapolable a otros contextos, como la temperatura o muchos tipos de indicadores sociales o demográficos. El significado de ambas energías dependerá del contexto.

En resumen, los aspectos que se observan en este gráfico son los siguientes:

- Ciclo general: en los caso periódicos, es lógico que el grafo generado tenga una forma más o menos cíclica.
- Tamaño del radio: cuanto mayor sea, más intercambio de energía hay.
- Presencia de subciclos: a veces, el gráfico presenta otros ciclos más pequeños, que se corresponden con intervalos de tiempo donde se mantienen más o menos los valores de velocidad y aceleración.
- Localización horizontal del centro del ciclo: si está a la derecha del 0, hay una velocidad subyacente positiva; si, por el contrario, está a la izquierda, la velocidad resultante en ese intervalo es negativa.
- Localización vertical del ecntro del ciclo: explicación análoga a la anterior, pero en el eje vertical, y con referencia a la aceleración.
- Cambios en la forma de los ciclos en las distintas repeticiones del periodo (años, en la mayoría de los casos).

Se utiliza a continuación un ejemplo basado en los valores que toma el *Índice de bienes perecederos* americano entre 1920 y 2000 (*Nondurable goods index)*. Se trata de un conjunto de datos incluido en la librería `fda` de R, que incorpora técnicas de análisis de datos funcionales, así como este y otros conjuntos de datos.

El índice, como muchos indicadores económicos, presenta una evolución exponencial, de manera que se ha preferido trabajar con la evolución del logaritmo en base 10 de los datos, de manera que se obtiene una tendencia general positiva y lineal. Como todos los índices, experimenta cambios tanto de tipo estacional, como los derivados de eventos que repercuten en la sociedad y la economía, como las guerras.

Se ha procedido a suavizar la función, y después de eso se reflejan los phase-plane plot de dos años muy diferentes en cuanto a tendencia: 1929, cuando estalla la Gran Depresión, y 1967, uno de los años en los que el país experimenta un auge económico en la década de los sesenta, debido al auge en el consumismo protagonizado por la generación del *baby boom.*

En indicadores económicos como el de nuestro ejemplo, lo que considerábamos energía potencial corresponde a los recursos humanos, energéticos y materiales de los que se dispone en el momento para realizar una actividad económica concreta, que en nuestro caso sería la fabricación de bienes perecederos. La energía cinética, por su parte, se corresponde con el procedimiento de fabricación en sí, cuando los bienes están siendo elaborados y puestos a disposición del mercado.

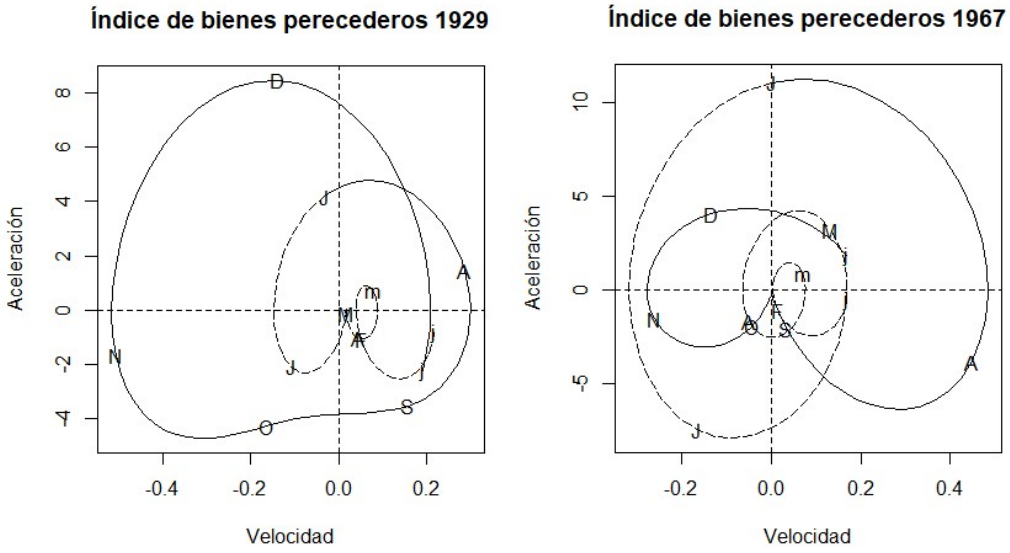

Figura 2.13: Phase-plane plots para el nondurable goods index en 1929 (izquierda) y 1967 (derecha).

Una vez se conoce el contexto, se puede utilizar este ejemplo para ilustrar las caracteristicas de este gráfico. Uno de los elementos que se consideran es la ubicación del centro del grafo. Como era de esperar, en el primer ciclo el ecntro está a la izuquiera del cero, lo que representa undescenso en la velocidad, explicable por la crisis. En el ciclo de 1967 sucede lo contrario, correspondiendo esto a un auge económico. Otro elemento destacable es el cambio de escala, no tan evidente en la aceleración como en la velocidad, lo que sugiere un mayor intercambio de energía en la economía en 1967.

Como suele suceder con los indicadores económicos, cada uno de los grafos en realidad está compuesto por dos ciclos. Sin embargo, el comportamiento no es el mismo. En 1929 el verano no es el momento de año donde el índice económico presenta una velocidad de crecimiento más negativa, mientras que en 1967 sí desbanca a las navidades como el momento del año donde el índice experimenta más dificultades para crecer.

Como se puede ver, este gráfico sirve para confirmar sospechas sobre el aspecto o comportamiento de los gráficos basadas en el conocimiento previo del contexto, pero también puede revelar comportamientos que a priori no se consideraban.

Capítulo 3
Aplicación a los datos meteorológicos

3.1 Datos: recolección y estaciones meteorológicas

Una vez explicado el funcionamiento del Análisis Funcional de Datos mediante suavizado, vamos a ver qué conclusiones se pueden extraer de la aplicación de esta técnica a los datos de temperatura recogidos en el estudio.

En este estudio se han considerado datos de temperatura de 90 estaciones meteorológicas repartidas por España. Estas estaciones se han elegido de manera que cada provincia, isla y ciudad autónoma tenga representación en los datos estudiados.

Los datos han sido extraidos de los sitios web de la Asociación Estatal de Meteorología (AEMET), de la Consejería de Agricultura y Pesca de la Junta de Andalucía y de Mateoclimatc. Las estaciones meteorológicas estudiadas se distribuyen por el territorio de la siguiente manera:

En primer lugar, es importante conocer cómo se reparten las 90 estaciones meteorológicas estudiadas a lo largo del territorio de España.

Figura 3.1: Ubicación de las estaciones meteorológicas de la Península, Islas Baleares y ciudades autónomas.

Figura 3.2: Ubicación de las estaciones meteorológicas de las Islas Canarias.

En esta tabla se detalla la localización y alititud de cada una de esas estaciones meteorolóicas, así como el identificador asignado a cada una.

COD	NOMBRE	PROVINCIA	LONGITUD	LATITUD	ALTITUD (metros)
CRN1	A Coruña	A CORUÑA	43°21’57” N	8°25’14” O	58
CRN2	A Coruña Aeropuerto	A CORUÑA	43°18’25” N	8°22’19” O	98
CRN3	Santiago de Compostela Aeropuerto	A CORUÑA	42°53’17” N	8°24’38” O	370
ALA	Foronda: Txokiza	ÁLAVA	42°22’55” N	2°44’6” O	513
ALB1	Albacete	ALBACETE	39°0’20” N	1°51’44” O	676
ALB2	Albacete: Base Aérea	ALBACETE	38°27’15” N	1°51’23” O	702
ALI1	Alicante Aeropuerto	ALICANTE	38°16’58” N	0°34’15” E	43
ALI2	Alicante	ALICANTE	38°22’21” N	0°29’39” E	81
ALM1	Almería	ALMERÍA	36°49’52” N	2°27’20” O	22
ALM2	Almería Aeropuerto	ALMERÍA	36°50’47” N	2°21’25” O	21
AST1	Gijón Puerto	ASTURIAS	43°33’36” N	5°42’3” O	5
AST2	Asturias Aeropuerto	ASTURIAS	43°34’1” N	6°2’39” O	127
AST3	Oviedo	ASTURIAS	43°21’12” N	5°52’27” O	336
AVI1	Ávila	ÁVILA	40°39’33” N	4°40’48” O	1103
AVI2	Sotillo de la Adrada	ÁVILA	40°17’33”	4°34’50” O	603
BAD	Badajoz Aeropuerto	BADAJOZ	38°53’0” N	6°48’50” O	185
BAR1	Barcelona Aeropuerto	BARCELONA	41°17’34” N	2º4’12” E	4
BAR2	Barcelona: Fabra	BARCELONA	41°25’6” N	2°7’27” E	408
BUR	Burgos Aeropiuerto	BURGOS	42°21’25” N	3°37’13” O	891
CAC1	Cáceres	CÁCERES	39°28’17” N	6°20’20” O	394
CAC2	Plasencia	CÁCERES	40°4’25” N	6°8’47” O	420
CAD1	Jerez de la Frontera	CÁDIZ	36°45’2” N	6°3’21” O	27
CAD2	Tarifa	CÁDIZ	36°0’50” N	5°35’56” O	32
CAD3	Vejer de la Frontera	CÁDIZ	36°17’6” N	5°50’24” O	24
CANT1	Santander Aeropuerto	CANTABRIA	43°25’26” N	3°49’32” O	33
CANT2	Santander	CANTABRIA	43°29’28” N	3°48’2” O	52
CAS	Castellón de la Plana: Almazora	CASTELLÓN	39°57’26”” N	0°24’19” E	43
CEU	Ceuta	CEUTA	35°53’17” N	5°20’49” O	27
CR	Ciudad Real	CIUDAD REAL	38°59’21” N	3°55’13” O	628
CORD	Córdoba Aeropuerto	CÓRDOBA	37°50’56” N	4°50’48” O	90
CUE	Cuenca	CUENCA	40°4’2” N	2°7’55” O	948
GER	Gerona Aeropuerto	GERONA	41°54’42” N	2°45°48” E	143

COD	NOMBRE	PROVINCIA	LONGITUD	LATITUD	ALTITUD (metros
GRA2	Granada Aeropuerto	GRANADA	37°11’23” N	3°47’22” O	567
GUA1	Guadalajara: El Serranillo	GUADALAJARA	40°39’33” N	3°10’24” O	639
GUA2	Molina de Aragón	GUADALAJARA	40°50’30” N	1°52’44” O	1062
GUI1	Hondarribia: Malkarroa	GUIPÚZCOA	43°21’25” N	1°47’32” O	4
GUI2	San Sebastián: Igueldo	GUIPÚZCOA	43°18’23” N	2°2’28” O	251
GRA1	Granada: Base Aérea	GRANADA	37°8’14” N	3°37’53” O	687
HUEL1	Huelva: Ronda Este	HUELVA	37°16’42” N	6°54’42” O	19
HUEL2	Moguer	HUELVA	37°8’47” N	6°47’33” O	87
HUES	Huesca Aeropuerto	HUESCA	42°5’4” N	0°19’32” O	546
IB1	Ibiza Aeropuerto	I. BALEARES	38°52’35” N	2°7’55” E	6
IB2	Menorca Aeropuerto	I.BALEARES	39°51’17” N	4°12’56” E	91
IB3	Palma Puerto	I.BALEARES	39°33’12” N	2°37’31” E	3
JAE	Jaén	JAÉN	37°46’39” N	3°48’32” O	580
LPA1	Fuerteventura Aeropuerto	LAS PALMAS	28°26’41” N	13°51’47” O	25
LPA2	Gran Canaria Aeropuerto	LAS PALMAS	27°55’4” N	15°23’43” O	32
LPA3	Lanzarote Aeropuerto	LAS PALMAS	28°57’7” N	13°36’1” O	14
LEO1	León Aeropuerto	LEÓN	42°35’58” N	5°38’27” O	912
LEO2	Ponferrada	LEÓN	42°33’50” N	6°36’0” O	534
LER	Lérida	LÉRIDA	41°37’34” N	0°35’53” E	185
LUG	Lugo Aeropuerto	LUGO	43°36’41” N	7°27’27” O	445
MAD1	Madrid	MADRID	40°24’43” N	3°40’41” O	667
MAD2	Madrid Aeropuerto	MADRID	40°28’0” N	3°33’20” O	609
MAD3	Puerto de Navacerrada	MADRID	40°37’35” N	4°0’38” O	1894
MAD4	Torrejón de Ardoz	MADRID	40°29’19” N	3°26’37” O	607
MAD5	Madrid: Cuatro Vientos	MADRID	40°22’32” N	3°47’10” O	690
MLG	Málaga Aeropuerto	MÁLAGA	36°59’58” N	4°28’56” O	5
MEL	Melilla	MELILLA	35°16’35” N	2°57’23” O	52
MUR1	Murcia	MURCIA	39°0’7” N	1°15’15” O	61
MUR2	Alcantarilla: Base Aérea	MURCIA	37°57’28” N	1°43’43” O	75
MUR3	Murcia Aeropuerto	MURCIA	37°47’20” N	0°48’12” O	4
NAV1	Pamplona	NAVARRA	42°49’4” N	1°38’18” O	450
NAV2	Pamplona Aeropuerto	NAVARRA	42°46’37” N	1°39’0” O	459
ORE	Ourense	OURENSE	42°19’31” N	7°51’35” O	143

COD	NOMBRE	PROVINCIA	LONGITUD	LATITUD	ALTITUD (metros
PAL	Carrión de los Condes	PALENCIA	42°21’3” N	4°37’22” O	830
PON1	Pontevedra	PONTEVEDRA	42°26’18” N	8°36’57” O	108
PON2	Vigo Aeropuerto	PONTEVEDRA	42°14’19” N	8°37’26” O	261
SAL1	Salamanca Aeropuerto	SALAMANCA	40°57’34” N	5°29’54” O	790
SAL2	Salamanca	SALAMANCA	40°57’25” N	5°39’44” O	775
SEG	Segovia	SEGOVIA	40°56’34” N	4°7’35” O	1005
SOR	Morón de Almazán	SORIA	41°24’53” N	2°24’47” O	1006
SEV1	Sevilla Aeropuerto	SEVILLA	37°25’0” N	5°52’45” O	87
SEV2	Morón de la Frontera	SEVILLA	37°9’32” N	5°36’41” O	34
TAR1	Reus Aeropuerto	TARRAGONA	41°8’24” N	1°9’49” E	71
TAR2	Tortosa	TARRAGONA	40°39’13” N	0°29’36” E	50
TEN1	El Hierro Aeropuerto	S. C. TENERIFE	27°49’8” N	17°53’20” O	32
TEN2	La Palma Aeropuerto	S. C. TENERIFE	28°37’59” N	17°45’18” O	33
TEN3	Sta. Cruz de Tenerife	S. C. TENERIFE	28°27’48” N	16°15’19” O	35
TER1	Teruel	TERUEL	40°21’2” N	1°7’27” O	900
TER2	Calamocha	TERUEL	40°55’34” N	1°17’36” O	890
TOL	Toledo	TOLEDO	39°53’5” N	4°2’43” O	515
VALE1	Valencia	VALENCIA	39°27’34” N	0°22’9” O	12
VALE2	Valencia: Viveros	VALENCIA	39°28’50” N	0°21’59” O	11
VALL1	Valladolid Aeropuerto	VALLADOLID	41°42’43” N	4°51’20” O	846
VALL2	Valladolid	VALLADOLID	41°38’27” N	4°45’16” O	735
VIZ	Bilbao Aeropuerto	VIZCAYA	43°17’53” N	2°54’23” O	42
ZAM	Zamora	ZAMORA	41°30’56” N	5°44’7” O	656
ZAR1	Zaragoza Aeropuerto	ZARAGOZA	41°39’38” N	1°0’15” O	249
ZAR2	Daroca	ZARAGOZA	41°6’52” N	1°24’36” O	779

Tabla 3.1: Listado de estaciones meteorológicas.

Es sabido que la altitud es un factor que puede tomar bastante importancia en un estudio como el realizado. Por ello, se presenta un gráfico que incluye información sobre la altitud a la que se encuentran las estaciones meteorológicas de la Península, así como las de las Islas Baleares.

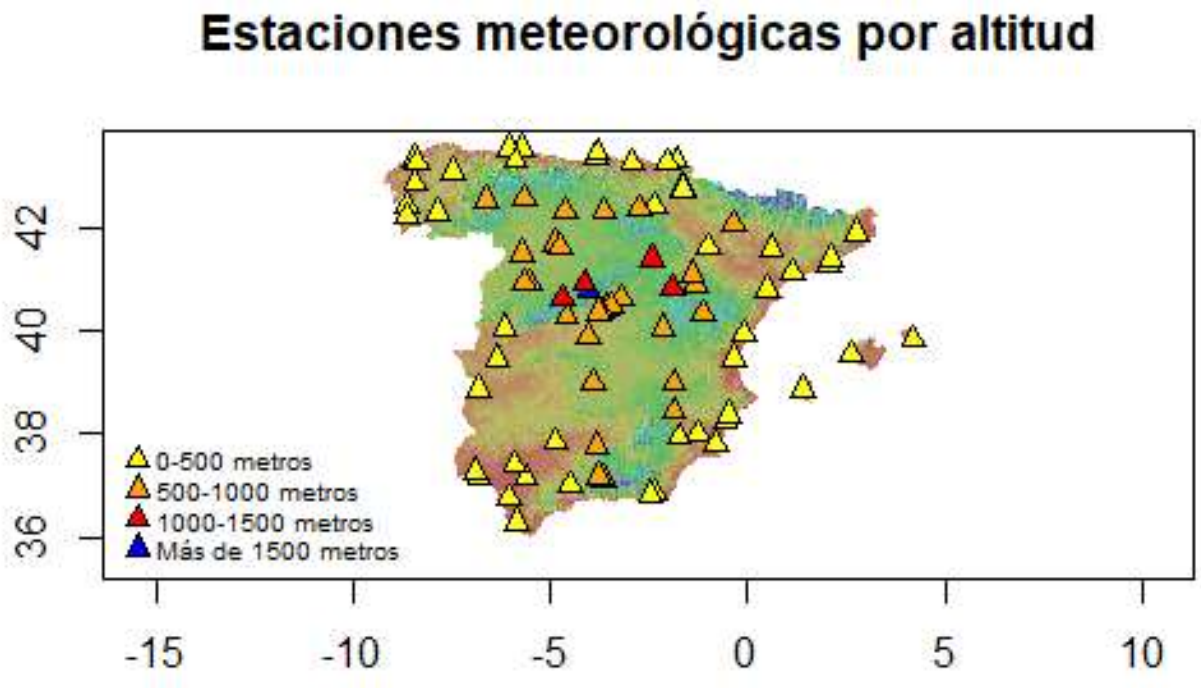

Figura 3.3: Ubicación y altitud de las estaciones meteorológicas de la Península.

La altitud media a la que se encuentran las estaciones estudiadas es 358,3 metros. Es destacable el hecho de que que cuatro de las estaciones estudiadas se encuentran ligeramente a más de mil metros sobre el nivel del mar: Segovia (1005 metros), Morón de Almazán (1006 metros), Molina de Aragón (1062 metros) y Ávila (1103 metros), todas ellas ubicadas en la Submeseta Norte. Destaca por encima de las demás la situada en un puerto de montaña como es el Puerto de Navacerrada (1894 metros), que, como se verá más adelante, tendrá un comportamientio anual diferente al resto. Ninguna de las estaciones meteorológicas ubicadas en Ceuta, Melilla o las Islas Canarias presenta una altitud por encima de los 100 metros, y no han sido reflejadas en el gráfico por motivos de software.

3.2 Suavizado de los datos

Los siguientes gráficos sirven para ilustrar la importancia del suavizado de los datos obtenidos de cara a estudiar la tendencia, como ya se ha expuesto antes.

A continuación figuran las gráficas que representan el comportamiento diario de cinco estaciones representativas de cinco zonas climáticas en las que se podría dividir, a priori, el territorio español: Islas Canarias (LPA1), costa mediterránea (VALE1), zona central (MAD1), costa atlántica (AST1) y sur (CAD1). En ninguno de los gráficos se ha aplicado ninguna clase de suavizado.

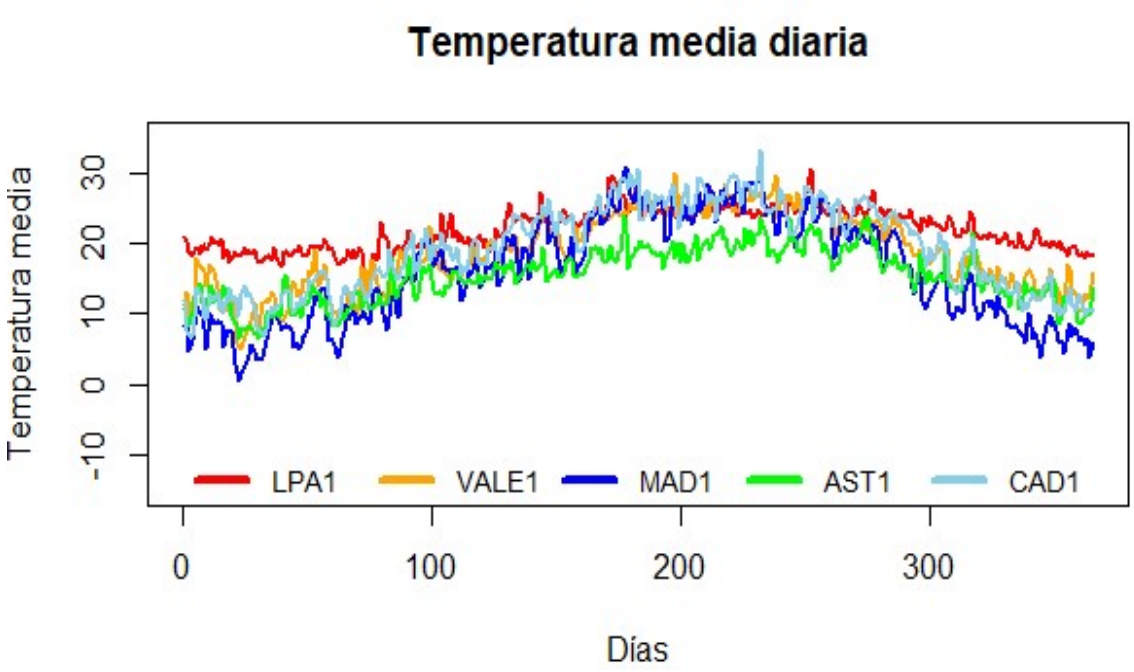

Figura 3.4: Evolución de la temperatura media diaria en las estaciones destacadas.

Como se esperaba, se puede adivinar cierta tendencia tras la unión de puntos (interpolación) realizada, pero no sin dificultad debido a la alta variabilidad entre días, y la existencia de rachas cortas de aumento o disminución de la temperatura en periodos cortos de días consecutivos. Por otro lado, la consideración de los datos "en crudo", esto es, sin suavizar, hace casi imposible el desarrollo de los estudios que se quieren llevar a cabo, como el Análisis de Componentes Principales sobre las funciones subyacentes.

Por ello, es necesario llevar a cabo un suavizado de los datos, de cara a trabajar con unas tendencias que expliquen la evolución anual de las temperaturas, eliminando en la medida de lo posible alteraciones puntuales sin que ello haga desaparecer características importantes de esa tendencia o función subyacente.

Ese "compromiso" entre ambas intenciones se corresponde con la elección del número de bases sobre las que se suavizará, que, en caso de ser demasiado grande, no eliminará las variaciones puntuales, y si, por el contrario, se escoge un número muy pequeño de bases, las funciones se suavizarán demasiado, hasta el punto de aproximar su comportamiento al de una recta, de manera que desaparece su comportamiento característico.

La elección del número de bases se ha realizado de una manera subjetiva, teniendo en cuenta la vocación puramente descriptiva de la muestra. Se han utilizado siete bases de tipo B-spline, representadas en el siguiente gráfico:

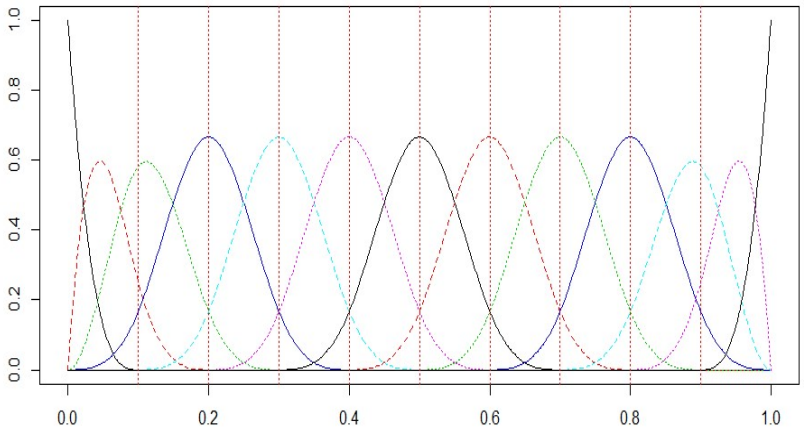

Figura 3.5: Sistema de bases B-spline utilizado para suavizar las funciones de temperatura.

Tras realizar el suavizado con respecto a estas bases, se obtienen los siguientes resultados para las cinco estaciones antes mostradas:

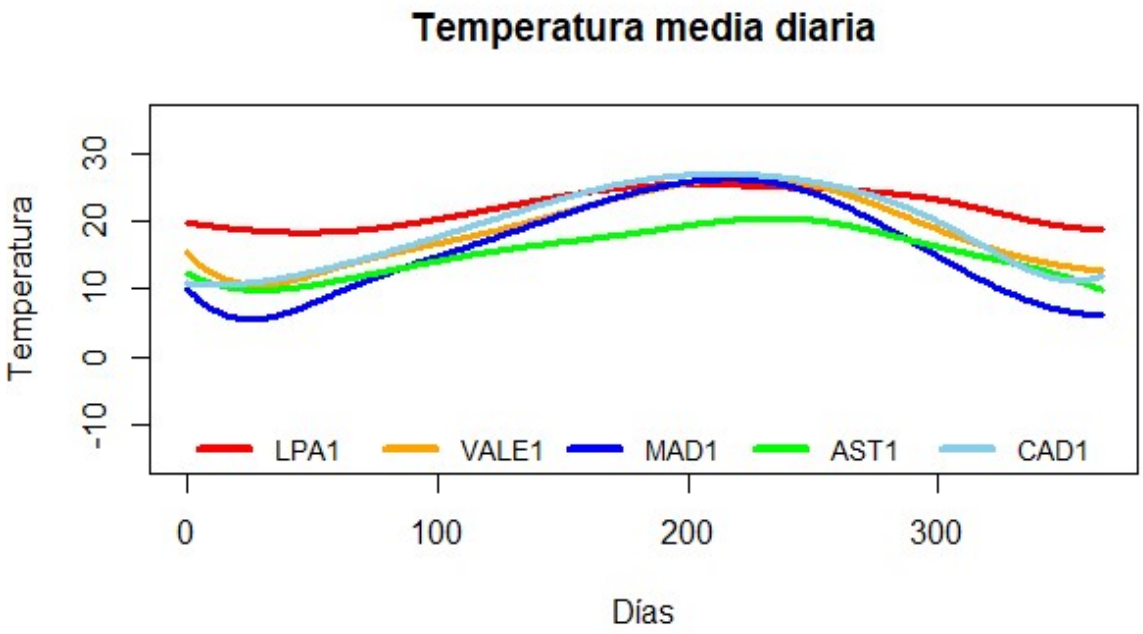

Figura 3.6: Curvas suavizadas de la evolución de la temperatura media diaria en las estaciones destacadas.

Es evidente que, tras este proceso, se dispone de funciones que permiten una interpretación mucho más sencilla. Desde el punto de vista intuitivo o humano, no hay que hacer esfuerzos mentales para intentar eliminar variaciones puntuales de cara a adivinar la función subyacente.

Se pueden realizar observaciones más o menos esperadas, pero que antes del suavizado de los datos eran más complicadas de extraer. La estación ubicada en las Islas Canarias presenta una variación térmica anual mínima, con temperaturas cálidas muy constantes a lo largo de todo el año. La estación de Madrid (MAD1) se encuentra en el interior de la Península, por lo que presenta una tendencia propia de un clima más continental: veranos cálidos e inviernos fríos. La estación ubicada en Gijón (AST1) no muestra una amplitud térmica grande, situándose su función siempre en valores que por lo general no sobresalen por debajo de los diez grados ni por encima de los veinte; por su parte, CAD1 y VAL1 registran, por lo general, temperaturas más altas durante todo el año y mayores diferencias entre los meses de invierno y los de verano.

Por otro lado, ahora que se dispone de funciones que representan un comportamiento o tendencia general en vez de datos discretos unidos por líneas, los posteriores análisis funcionales arrojarán resultados que no tendrán en cuenta esas alteraciones puntuales, y, como ya se ha dicho, las funciones de R relativas al análisis funcional devolverán resultados más concluyentes que si no hubiera habido un suavizado de los datos.

3.3 Análisis de los datos

A continuación se procede a analizar carácterísticas de los datos utilizando las técnicas explicadas en el capítulo anterior.

3.3.1 Análisis de Componentes Principales Funcional

Se procede, una vez suavizados los datos, a realizar el Análisis de Componentes Principales de las funciones obtenidas, con el objetivo de estudiar los modos de variación más importantes en los datos. A la vez, se obtiene una idea de la complejidad de las curvas, en el sentido de ver en base a cuántos comportamientos de variabilidad esencial se pueden establecer.

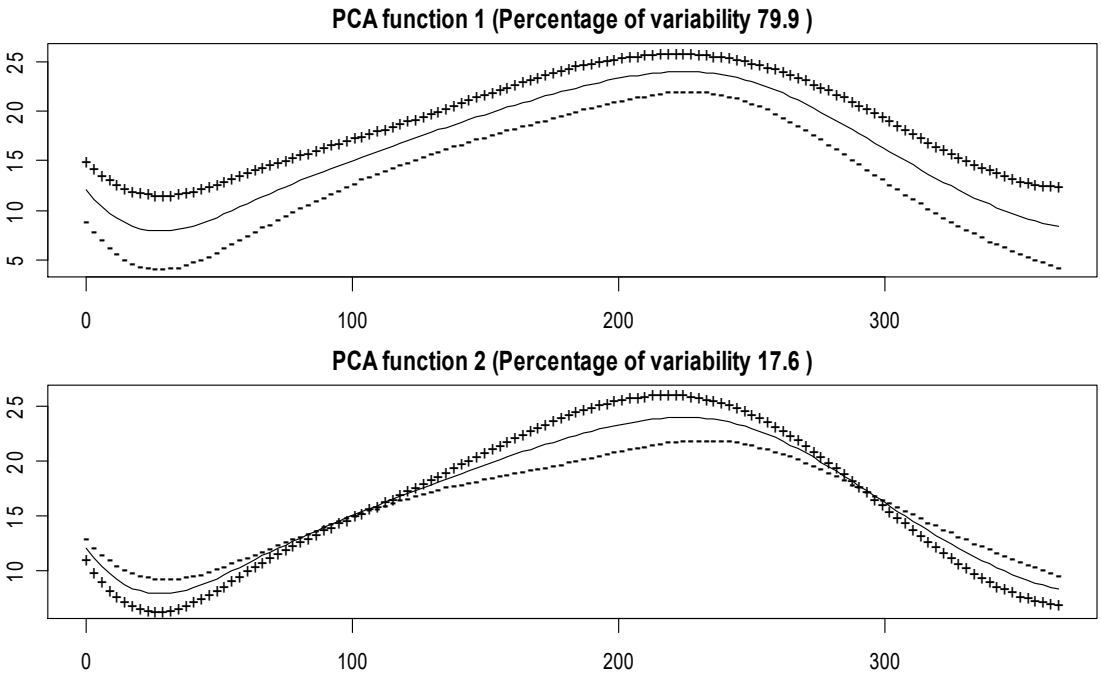

Figura 3.7: Representación de las dos componentes principales funcionales como perturbaciones de la función media.

La imagen muestra la perturbación que resulta de sumar o restar a la función media un múltiplo conveniente (en este caso ha parecido bastante oportuno que sea 1) de cada una de las dos componentes principales, para así tener una idea de la variabilidad que explica cada componente. Ambas explican un 97.6% de la variabilidad, así que no parece oportuno ni necesario estudiar la existencia de una tercera componente principal que, como mucho, tendría un peso del 2,4% en ese aspecto.

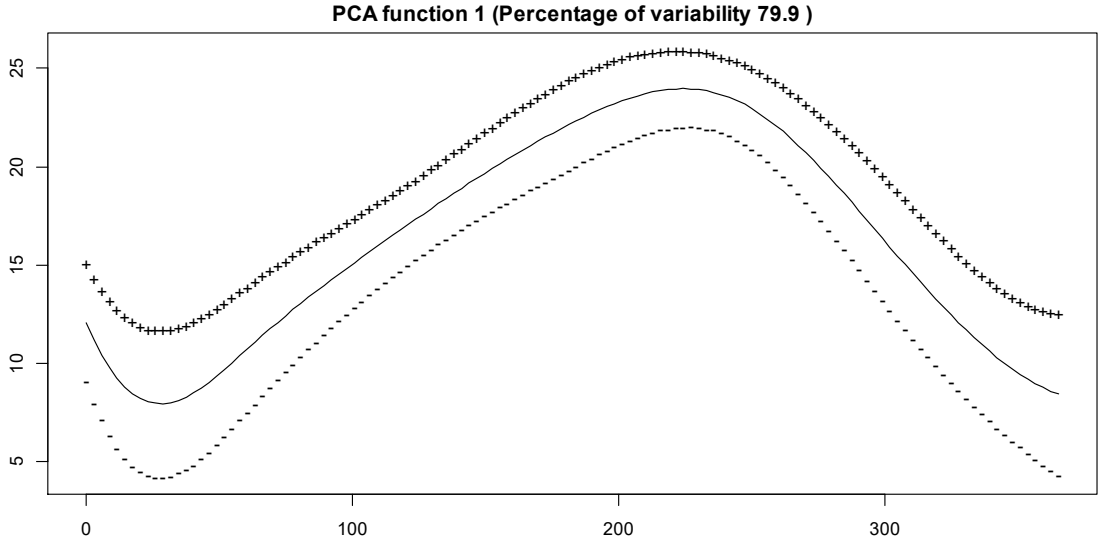

Figura 3.8: Representación de la primera componente principal funcional como una perturbación de la función media.

La primera componente funcional explica por sí misma prácticamente un 80% de la variabilidad. Claramente, el efecto de su presencia es una componente de tamaño:: cuanto mayor sea el valor asociado a esta componente de una de las funciones estudiadas, mayor temperatura se registrará globalmente a lo largo de todo el año. Por el contrario, cuanto menor sea, se tratará de una estación ubicada en un lugar más frío.

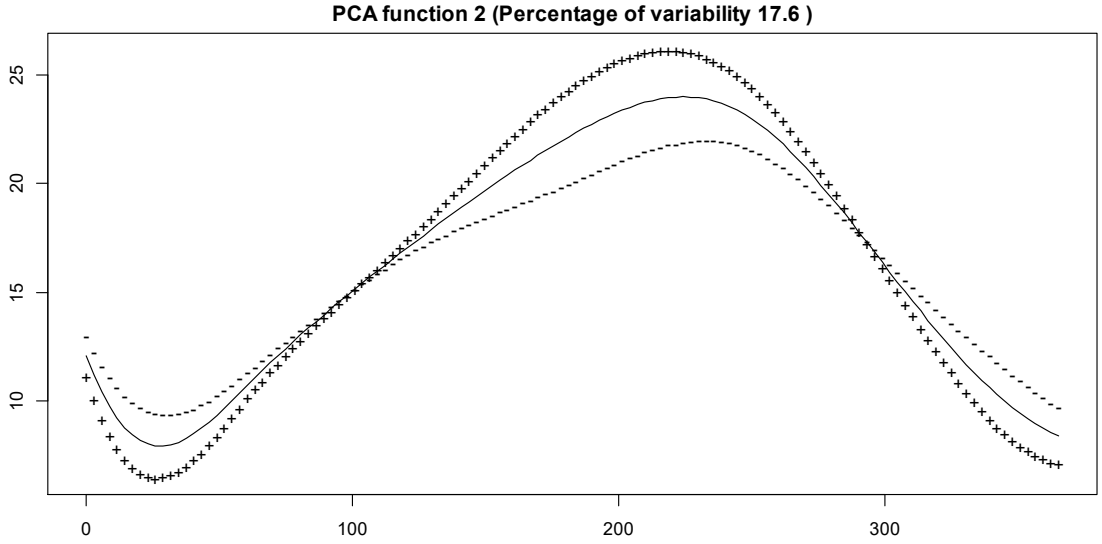

Figura 3.9: Representación de la segunda componente principal funcional como una perturbación de la función media.

Por el otro lado, la segunda componente se corresponde con un efecto algo más complejo sobre el comportamiento de la temperatura anual, pero perfectamente esperable: en base a ella, podríamos clasificar las funciones de temperatura observadas según la estabilidad de las temperaturas a lo largo del año. Una curva con un valor mayor en esta componente presentará valores más altos en el centro, y más bajos en los extremos; dicho de otro modo, representará a una región con veranos más cálidos e inviernos más fríos. Por otro lado, cuanto menor sea el valor correspondiente a esta componente, mayor suavidad o estabilidad presentará la curva.

Fundamentalmente, los comportamientos de todas las curvas estudiadas podrán clasificarse respecto a la mayor o menor fuerza de estas dos componentes. Se trata de una conclusión que coincide bastante con los resultados que uno podría esperar si tiene conocimiento suficiente de la climatología de nuestro país.

Al igual que en el ACP clásico, la “fuerza” o coordenadas de cada función observada en cada una de las componentes se representa con los scores o coordenadas de las curvas en la representación de las componentes principales.

Se ha realizado el mismo estudio llevando a cabo una rotación VARIMAX, pero se ha estimado que aportaba una información extra escasa y más difícil de interpretar, a la vez que se perdían características importantes de los datos, por lo que no se ha desarrollado más en esta memoria.

A continuación, se procede a ver cómo se distribuyen las diferentes estaciones en el territorio nacional en función de ambos scores con la ayuda visual del mapa. Se utilizan colores más cálidos para representar valores más positivos de los scores, de manera

que cuanto más positivo sea el valor de un score más se acercará su tono a un rojo vivo, mientras que un valor altamente negativo se representará con un color azul intenso.

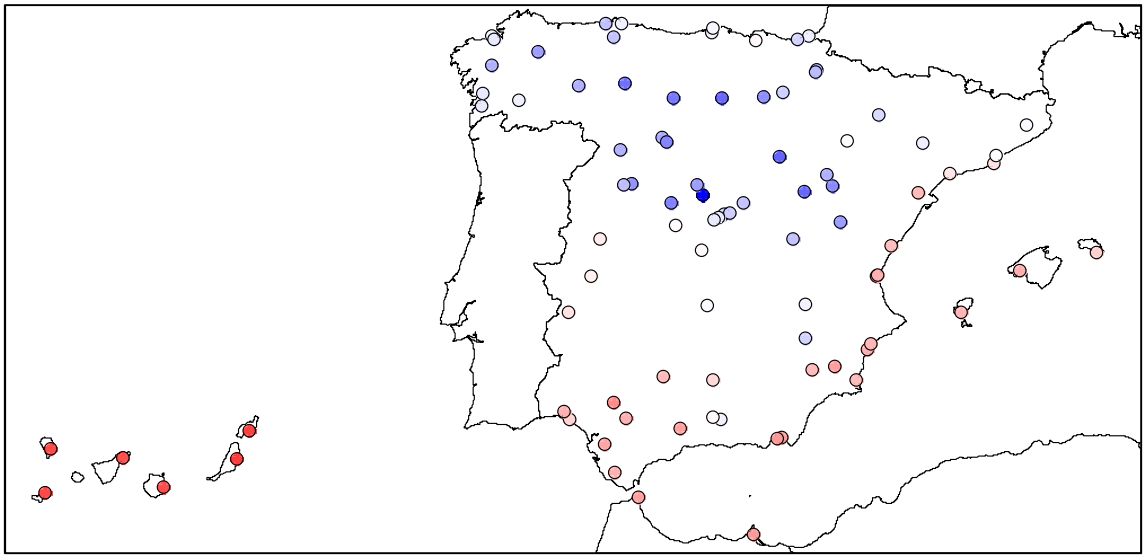

Figura 3.10: Reparto sobre el territorio del score de la primera componente principal.

En primer lugar se observa el mapa relativo a los scores en la primera componente, que se correspondía con un valor muy positivo para una curva con valores altos, y un valor muy negativo para zonas más frías a lo largo del año. Se aprecia de manera clara que, en general, la zona noroeste de la península, así como la Submeseta Norte, es donde se encuentran los puntos con menor score para esta componente; por otro lado, salvo alguna excepción, la costa mediterránea, incluyendo las Islas Baleares, presenta un valor claramente positivo, que se acentúa en el sur peninsular. Los valores más altos se dan en las Islas Canarias, que presentan una temperatura cálida durante todo el año.

Destaca de manera remarcable la estación de Puerto de Navacerrada (MAD3), que, como se decía al principio, era con diferencia la que más altitud presentaba; en consecuencia, se trata de la estación con un valor más marcadamente negativo en el score de la primera componente, porque las temperaturas en la zona son más bajas a lo largo del año.

El siguiente mapa muestra el gráfico de scores en la segunda componente principal.

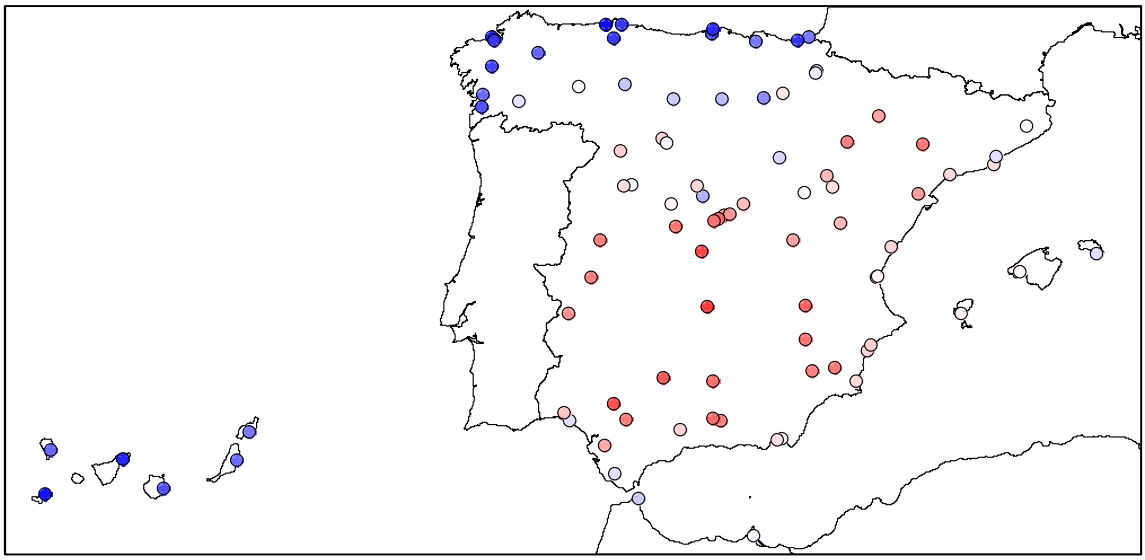

Figura 3.11: Reparto sobre el territorio del score de la segunda componente principal.

Recordemos que la segunda componente principal toma fuerza en lugares con gran amplitud térmica a lo largo del año, mientras que en el mapa aparecen en azul las zonas donde la temperatura presenta una menor variación a lo largo del año. Hay dos zonas del territorio nacional con un score marcadamente negativo, las Islas Canarias y el norte peninsular. Pese a presentar temperaturas muy diferentes entre sí durante todo el año, comparten la característica de que la temperatura se mueve en unos rangos relativamente pequeños.

Justo lo contrario sucede en el centro de la Meseta, gran parte de la Submeseta Sur y la zona más interior de Cataluña, Andalucía y Aragón, donde sí hay una gran diferencia entre las temperaturas de los meses de verano e invierno; en la zona mediterránea, posiblemente debido a la influencia del mar como regulador térmico, no destaca tanto esta característica en ninguna estación, pero sí hay una amplitud térmica moderada a lo largo del año.

Se procede ahora a comentar el gráfico de scores, que representa en un espacio de coordenadas bidimensional las coordenadas de cada una de las curvas (estacione) en las dos componentes principales, de manera que pueden relacionarse fácilmente por sus proximidades en este gráfico bidimensional.

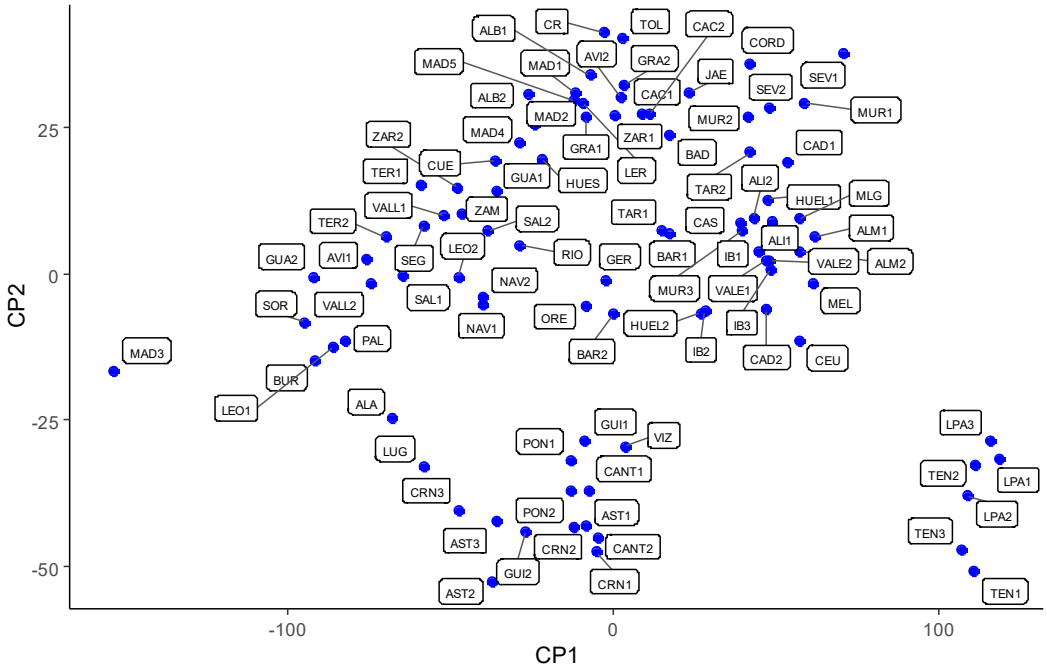

Figura 3.12: Mapa de scores de las curvas de temperatura.

Si dividiéramos el gráfico en cuatro cuadrantes iguales usando los ejes horizontal y vertical para ello, consideraríamos que los elementos del primer cuadrante (valor negativo para el primer score y positivo para el segundo) se corresponden con lugares fríos donde la diferencia entre los meses de más calor y los de más frío es grande. Por otro lado, arriba a la derecha se encontrarían las estaciones meteorológicas donde se dan las temperaturas más altas, y a la vez éstas presentan una variación anual grande.

Asimismo, los puntos ubicados por debajo del valor cero en la segunda componente presentan temperaturas que podríamos considerar que registran una mayor regularidad a lo largo del año, siendo los que quedan a la izquierda los más fríos, y a la derecha los correspondientes a zonas más cálidas.

Posiblemente, lo que más llame la atención desde el punto de vista visual, aunque se trata de algo más o menos esperado por lo comentado en el punto anterior, es la presencia de un punto atípicamente aislado del resto, que se corresponde con la estación ubicada en el Puerto de Navacerrada, en Madrid. Es importante recordar que se trata de un puerto de montaña ubicado a casi 1900 metros de altitud, por lo que registra temperaturas mucho más bajas a lo largo del año que cualquier otra de las estaciones meteorológicas estudiadas.

En segundo lugar, destaca la posición de las estaciones meteorológicas de las Islas Canarias, muy juntas entre sí y bastante aisladas de cualquier otro punto. Esto se debe a que, a la vez que destacan positivamente en la primera componente principal debido a las altas temperaturas generales, estas temperaturas presentan una variación muy pequeña a lo largo del año, por lo que presentan unos scores asociados a la segunda componente principal altamente negativos. Esta peculiaridad climática viene influida, sobre todo, por su posición marcadamente separada del resto de estaciones estudiadas.

Dentro de lo que parece la nube de puntos principal, es fácil ver que en la parte inferior del gráfico se concentran, por sus características comunes, las estaciones meteorológicas correspondientes a la zona atlántica.

Por otra parte, en su mayoría con valores positivos para ambas componentes se encuentran las estaciones ubicadas en la costa mediterránea que va desde la Comunidad Valenciana hasta Andalucía, donde en general hay temperaturas más altas a lo largo del año y una importante amplitud térmica.

Otro grupo a destacar son las estaciones centradas en lo que a la primera componente se refiere, que presentan score alto para la segunda. Aquí se encuentra gran parte de las estaciones que se encuentran en la Submeseta Sur, así como las de Cataluña y algunas de las estaciones ubicadas en Aragón. En general, el comportamiento medio anual de sus temperaturas no destaca, pero están ubicadas en zonas donde hay mucha diferencia entre los meses cálidos y los fríos.

Por último, hay un grupo de estaciones donde las temperaturas son algo más frías y no se presenta una amplitud térmica tan extrema. En su mayoría son las ubicadas en Castilla y León y Navarra.

Se ha llevado a cabo un procedimiento cluster con el método de k-medias con k=6, que arroja unos resultados bastante similares a las conclusiones a las que se ha llegado anteriormente. Las estaciones meteorológicas quedan divididas en seis grupos que, salvo algún caso aislado, están compuestos por puntos situados en ubicaciones próximas geográficamente.

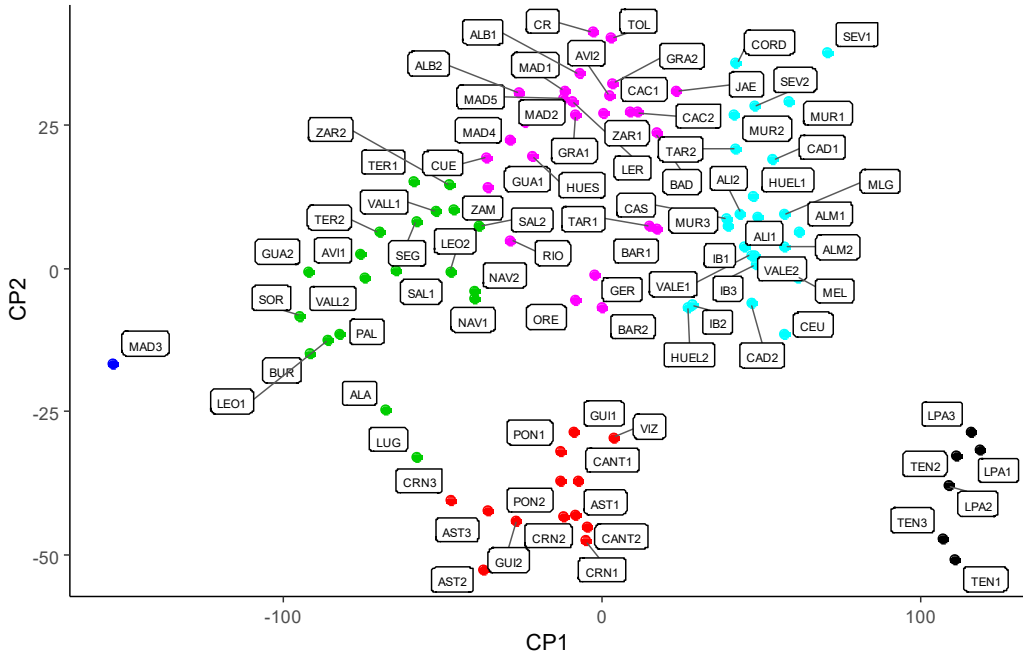

Figura 3.13: Mapa de scores con el resultado de las 6-medias representado con distintos colores.

Es interesante apoyarse en la ubicación geográfica de las estaciones para apreciar un patrón distributivo bastante claro. Exceptuando la estación ubicada en el Puerto de Navacerrada, las características de las estaciones las dividen en cinco grupos bastante diferenciables en el territorio: Islas Canarias, costa atlántica, costa mediterránea, Submeseta Norte y Submeseta Sur más Noreste peninsular.

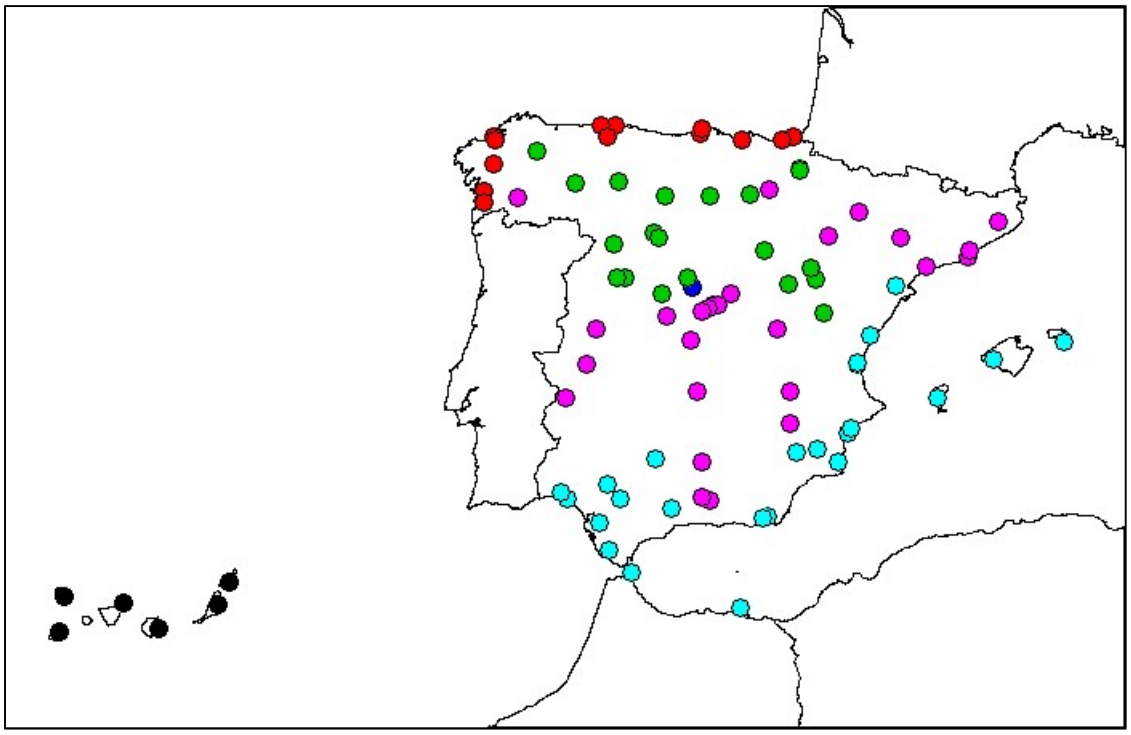

Figura 3.14: Reparto sobre el territorio de los clusters generados a partir del método de 6-medias sobre los scores.

3.3.2 Phase-plane plot

Este gráfico consiste en la representación de las dos primeras derivadas, Dx y D^2x, que se corresponden con la evolución de la velocidad y de la aceleración, respectivamente, en los cambios de temperatura. Al enfrentarlas se puede estudiar la estructura cíclica de la evolución de la temperatura a lo largo del año, la cual permite comprender o entender en mayor profundidad el comportamiento de la curva.

A continuación, se analiza este gráfico para cada una de las estaciones representativas elegidas al principio. En cada caso, se acompaña de su curva correspondiente, y se han usado los mismos límites horizontales y verticales, de manera que se pueden comparar entre sí más fácilmente. Por convención, el eje horizontal corresponde a la velocidad, y el vertical coincide con la aceleración.

Es importante recordar que, en este gráfico, se buscan una seria de características:

- Grado de parecido del ciclo con un ciclo cerrado (ciclo perfecto).
- Tamaño del radio del ciclo: cuanto más grande sea, más transferencia de energía hay; en nuestro caso, mayor transformación o cambio en las temperaturas.
- Localización horizontal del ciclo: si el centro está hacia la derecha del 0, hay una velocidad neta positiva; en caso contrario, será negativa. En nuestro caso esto no requerirá nuestra atención, ya que, al observar un único año, no esperamos ver una evolución creciente ni decreciente en este aspecto.

- Localización vertical del centro del ciclo: si está sobre el cero, hay un incremento neto de la velocidad, y lo contrario si está por debajo. Pasa lo mismo que con el punto anterior.
- La concentración de puntos de estudio equiespaciados (meses, en este caso), que reflejan que, en mayor o menor medida, se mantiene una tendencia.
- El momento del año en el que empieza a aumentar o disminuir la velocidad (momento en el que se atraviesa el eje vertical).

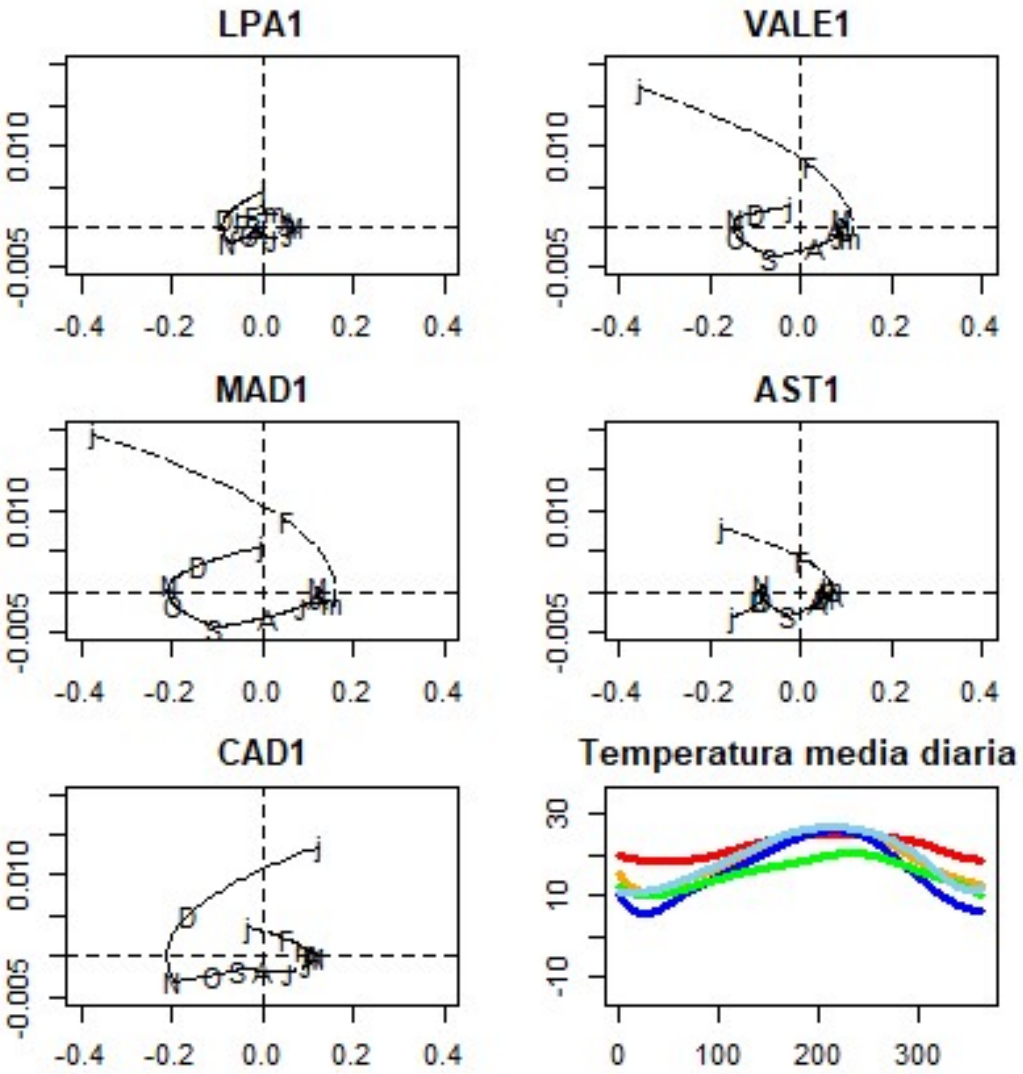

Figura 3.15: Phase-plane plots de las curvas de las estaciones destacadas.

Gráfico 1: Fuerteventura Aeropuerto (LPA1)

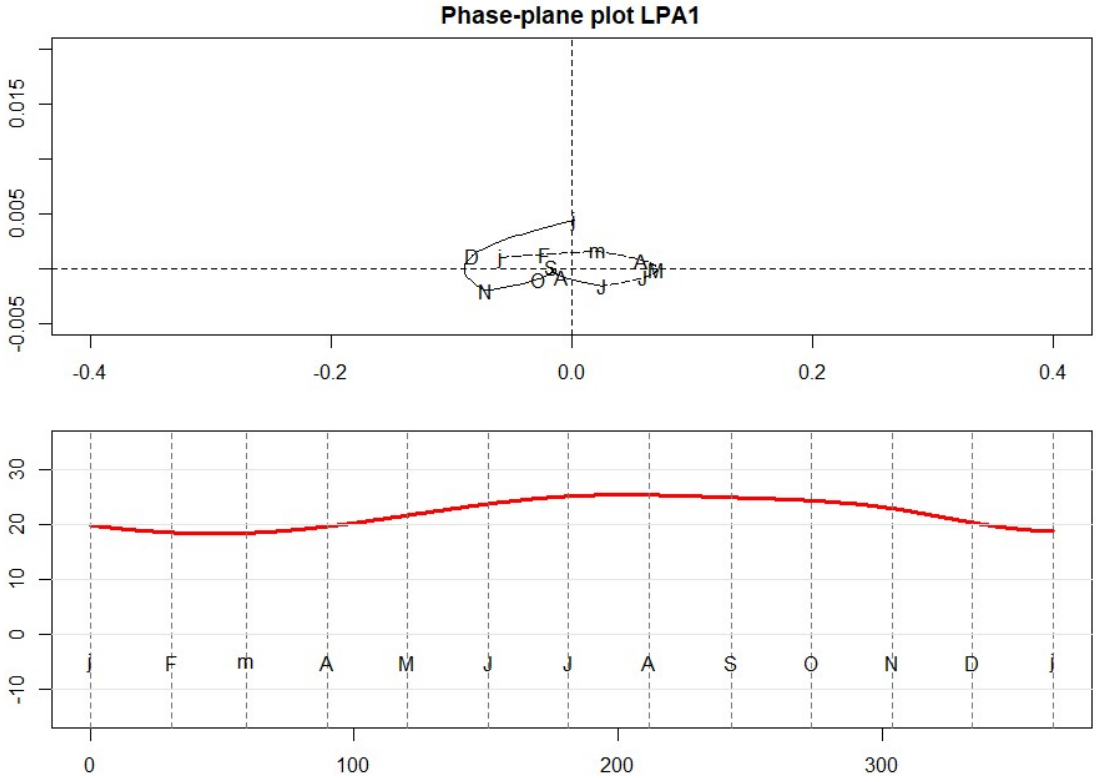

Figura 3.16: Phase-plane plot y curva suavizada de la temperatura de la estación Fuerteventura Aeropuerto (LPA1).

En este caso, nos encontramos con el ciclo con el radio más pequeño de entre los destacados. Tiene sentido, ya que se trata de una zona en la que no hay apenas variación en la temperatura a lo largo del año. Es también por esto que el ciclo es el que parece más cerca de estar cerrado, aunque el hecho de que el año termine a una temperatura similar a aquella con la que comenzó podría considerarse un hecho que parte de lo anterior, pero es algo más circunstancial, ya que sólo estamos estudiando un año.

Gráfico 2: Valencia (VALE1)

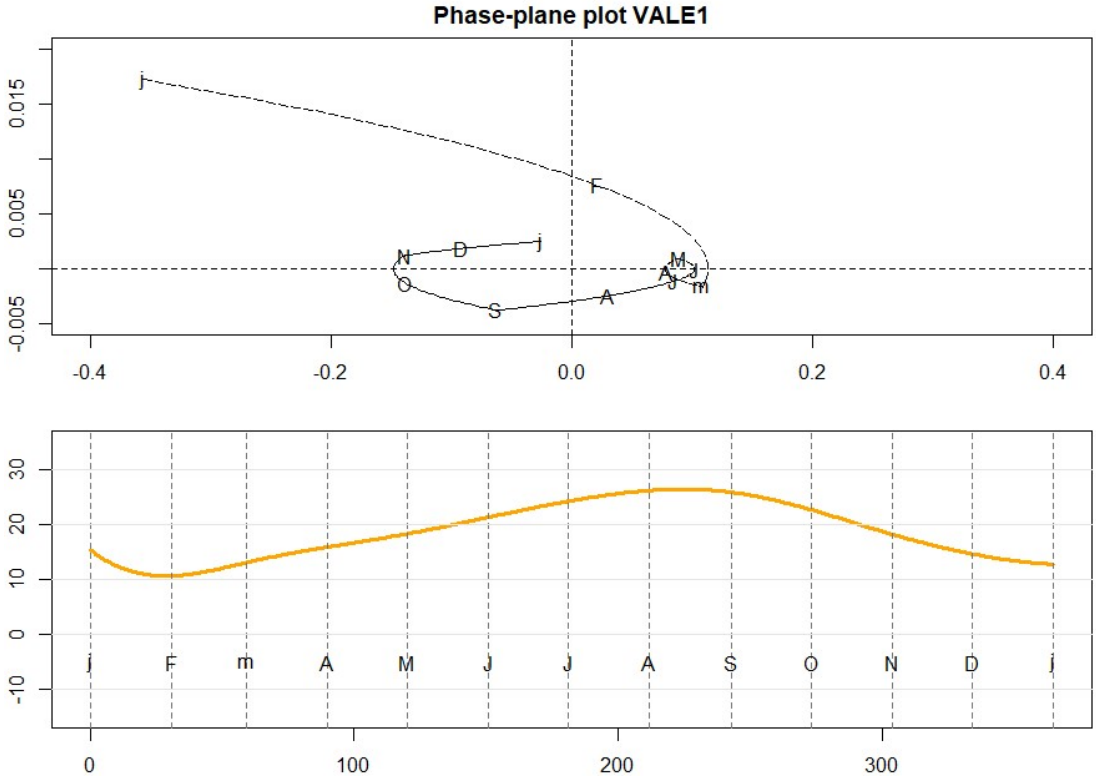

Figura 3.17: Phase-plane plot y curva suavizada de la temperatura de la estación Valencia (VALE1).

El phase-plane plot, en este caso, dibuja un ciclo de tamaño intermedio, ya que las temperaturas cambian algo más que en el caso anterior. Se aprecia un ciclo secundario de radio muy pequeño, que concentra los meses de marzo a julio, de lo que se deduce que, durante ese intervalo, la temperatura aumenta de una manera constante, ya que se presenta prácticamente la misma velocidad.

Gráfico 3: Madrid (MAD1)

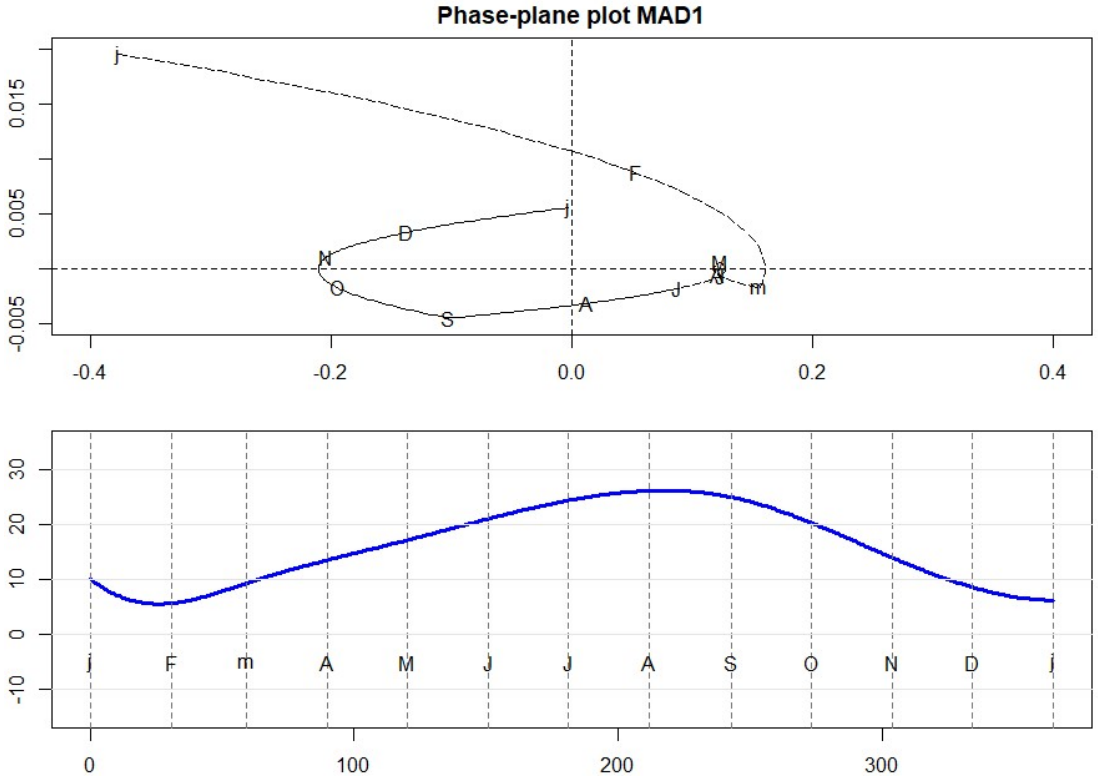

Figura 3.18: Phase-plane plot y curva suavizada de la temperatura de la estación Madrid (MAD1).

De los cinco estudiados, se trata del plot que presenta una mayor amplitud, lo que ayuda a hacerse una idea de que las temperaturas cambian bastante a lo largo del año; es lo esperado, al tratarse de la parte más continental de la Península, donde la amplitud térmica es mayor. Hay un pequeño subciclo que representa que en los meses de abril y mayo la evolución de la temperatura es constante, pero por lo demás la longitud del ciclo sugiere que durante el resto del año hay un cambio de tendencias constante.

Gráfico 4: Gijón Puerto (AST1)

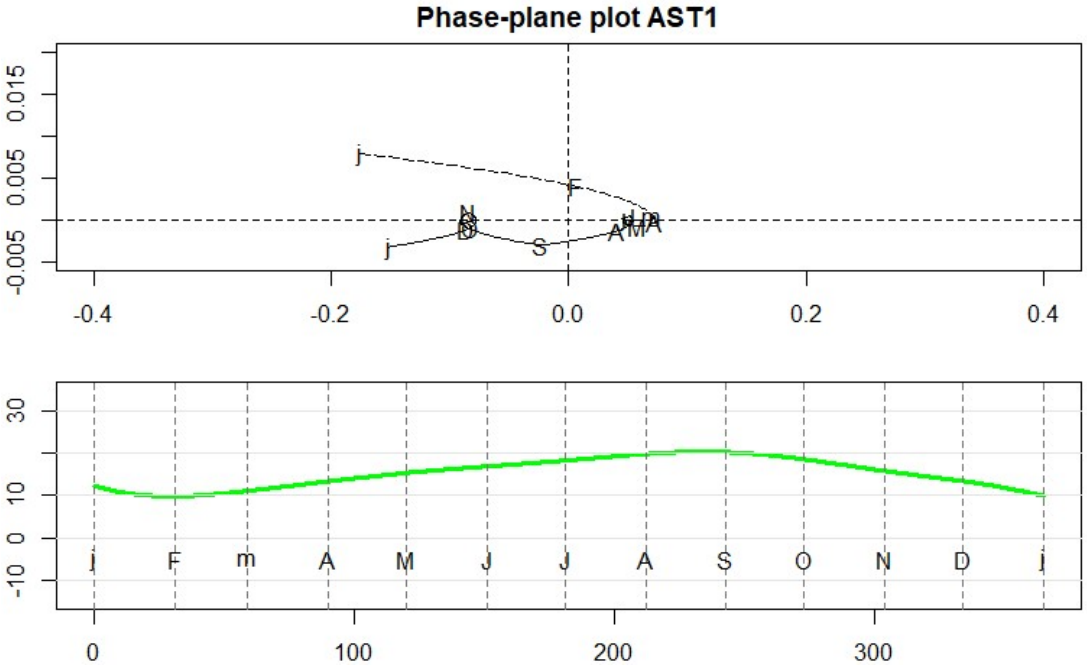

Figura 3.19: Phase-plane plot y curva suavizada de la temperatura de la estación Gijón Puerto (AST1).

Se trata del ciclo más pequeño después del correspondiente a la estación meteorológica del Aeropuerto de Fuerteventura (CAN1). De varias formas, ya se ha visto que la curva que representa la temperatura en las estaciones de la zona norte es bastante suave, por lo que no es un hecho sorprendente. Tampoco lo es la aparición de dos subciclos con un radio muy pequeño, que van de marzo a agosto y de octubre a diciembre, respectivamente.

Gráfico 5: Jerez de la Frontera (CAD1)

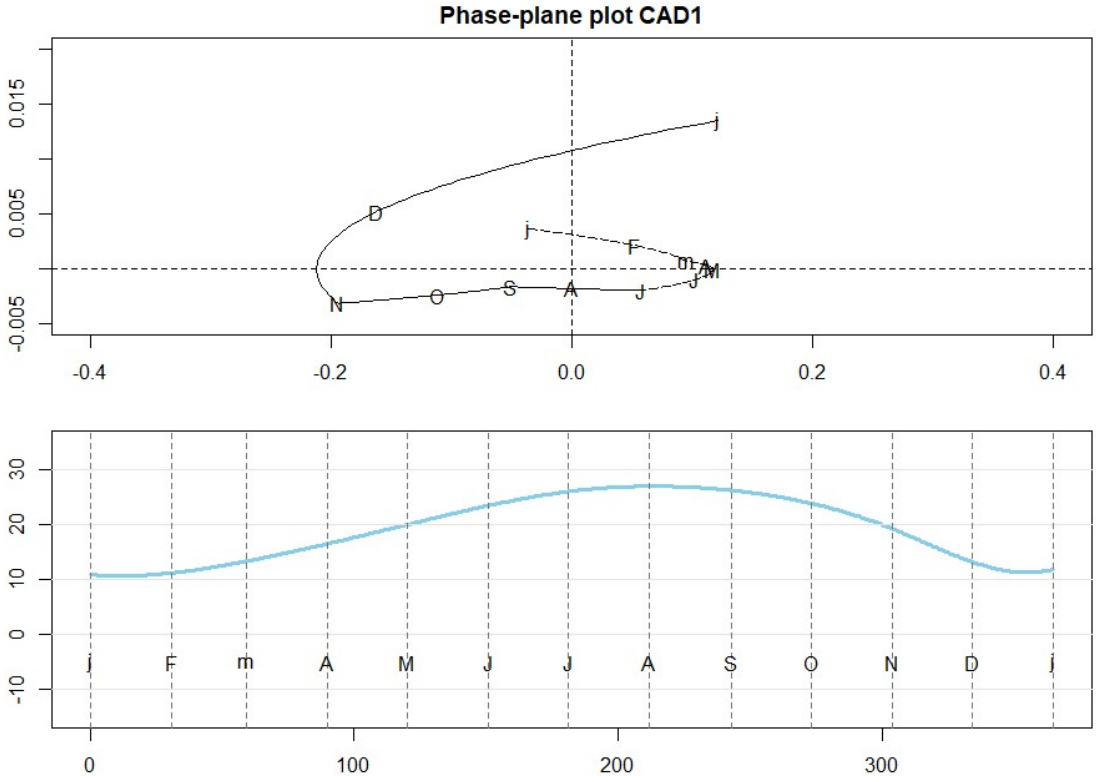

Figura 3.20: Phase-plane plot y curva suavizada de la temperatura de la estación Jerez de la Frontera (CAD1).

El gráfico evidencia que los cambios en la temperatura a lo largo de año son bastante grandes, aunque no tanto como en el interior peninsular; los puntos que representan el inicio de los meses están bastante espaciados, a excepción de los meses de marzo a junio, donde la temperatura crece de una manera constante. A final de año, la temperatura decrece de manera fuerte.

Conclusión de los phase-plane plot

En base a estos gráficos, también es posible hacer un comentario sobre cuándo comienzan las temperaturas a crecer o a decrecer en el año 2013, a partir de la observación de los momentos del año en los que el ciclo atraviesa el eje vertical.

En Valencia, Madrid y Jerez de la Frontera, la temperatura empieza a aumentar antes del comienzo del mes de febrero, mientras que en Gijón sucede casi en el momento en el que empieza el mes. En Canarias lo hace a mediados del mes, aunque el cambio de temperatura en esta región es suave en todos los momentos del año.

En consecuencia, Fuerteventura es el lugar donde más tarde empieza a descender la temperatura, bien entrado agosto, mientras que en el resto de estaciones esto sucede, como muy tarde, al comienzo del mes. Puede decirse entonces que el verano empieza más tarde que en las otras regiones observadas.

Conclusiones

En este TFG se han revisado diferentes técnicas de Análisis Funcional de Datos, como el suavizado y el Análisis de Componentes Principales Funcionales. Se ha utilizado fundamentalmente como referencia el libro *Functional Data Analysis* de Ramsay y Silverman.

Estas metodologías se han aplicado exitosamente en un conjunto de datos de temperatura en España en 2013, recogidos en 90 estaciones meteorológicas repartidas por todo el territorio. Así se han detectado cinco zonas climáticas y alguna estación claramente atípica.

Por su parte, los phase-plane plot nos han permitido detectar diferentes patrones bien diferenciados de evolución de la temperatura en el tiempo.

El hecho de trabajar con datos, de cuyo contexto se conocen bien sus características, ha permitido en todo momento trabajar con unos resultados en parte esperados que han servido de guía. No obstante, obviamente también se han descubierto características de los datos que han aportado una información con la que a priori no se contaba.

En base a este trabajo se pueden marcar unas líneas futuras de investigación, como por ejemplo partir de datos que abarquen más años, de cara a poder estudiar el funcionamiento de la periodicidad y si hay algún tipo de evolución a largo plazo en los datos. Sin embargo, el procedimiento de recogida de los datos para solo un año (2013) ya ha sido arduo, así que es muy posible que se trate de un proceso costoso, y en el que posiblemente falten datos para muchas estaciones.

Por otro lado, también podría estudiarse la posibilidad de estudiar la relación entre los datos de temperatura y otros fenómenos climatológicos, como las precipitaciones.

Bibliografía

Revistas, libros y artículos

Chávez-Chong, C.O., Sánchez-García, J.E., De la Cerda-Gastélum, J. (2015). Análisis de Componentes Principales Funcionales en series de tiempo económicas. *GECONTEC: Revista Internacional de Gestión del Conocimiento y la Tecnología. Volumen 3* (2), 13-25.

De Boor, C. (1978), *A Practical Guide to Splines.* Springer.

Grenander, U. (1950). Stochastic processes and stadistical inference. *Arkiv för matematik. Volumen 1* (3), 195-277.

Rao, C.R. (1958). Bengal anthropometric survey 1945: A statistical study. *Sankhya, Volumen 19,* 201-408.

Ramsay, J.O., Silverman, B.W. (1989). *Functional Data Analysis*. Springer.

Ramsay, J.O. (1982). When the data are functional. *Psychometrika*. *Volumen 47*, 379-396

Valderrama Bonnet, M.J., Aguilera del Pino, A.M. (2000). *Predicción Dinámica mediante Análisis de Datos Funcionales.* La Muralla.

Recursos web

https://cran.r-project.org/web/packages/fda/fda.pdf

http://faculty.bscb.cornell.edu/~hooker/ShortCourseHandout.pdf

http://www.psych.mcgill.ca/misc/fda/index.html

https://www.juntadeandalucia.es/agriculturaypesca/ifapa/ria/servlet/FrontController?action=Init

https://datosclima.es/Aemethistorico/Meteostation.php

https://foro.tiempo.com/2750-estaciones-meteorologicas-incluye-emaacutes-de-aemet-para-googleearth-t120238.0.html;msg2441206#msg2441206

https://www.meteoclimatic.net

Lista de figuras

Lista de tablas

Anexos: Código R

Anexo 1. Parte teórica

```
# - - - - - - - - - - - - - - - - - -  - - - - - -
# ----- Instalación de paquetes y librerías ----
# - - - - - - - - - - - - - - - - - -  - - - - - -
install.packages("fda")
install.packages("ggplot2")
install.packages("ggrepel")
install.packages("zoo")
library(readxl)
library(leaflet)
library(RColorBrewer)
library(fda)
library(ggplot2)
library(ggrepel)
library(zoo)

# - - - - - - - - - - - - - - - - - -  - - - - - -
# -------- Plot de funciones subyacentes --------
# - - - - - - - - - - - - - - - - - -  - - - - - -

set.seed(2)
curva1 <- c(rep(4.5, 20)+  rnorm(20, mean=2, sd=0.5))
curva2 <- c(rep(7, 20) + rnorm(20, mean=1.5, sd=0.6))

argvals1 <- 1:20
bks = 5
nba1 = bks+2
breaks1=seq(0,20,length=bks)


base1 <- create.bspline.basis(rangeval=c(0,20), nbasis=nba1,
breaks=breaks1)
c1fd <- smooth.basis(argvals=argvals1, y=curva1, fdParobj=base1)
plot.fd(c1fd$fd, col="red", ylim=c(0,12), lwd=1.5, lty=2, xlab="t",
ylab="", main="")
points(curva1, col="red", pch="+")

c2fd <- smooth.basis(argvals=argvals1, y=curva2, fdParobj=base1)
plot.fd(c2fd$fd, col="blue", ylim=c(0,12), lwd=1.5, add=TRUE, lty=2)
points(curva2, col="blue", pch="+")

legend("bottomleft", x.intersp = 0.2, col=c("blue","red"),
legend=c("x1","x2"), cex=0.9, bty="n", lwd=1, lty=2)

# - - - - - - - - - - - - - - - - - -  - - - - - -
# -------------- Sistemas de bases --------------
# - - - - - - - - - - - - - - - - - -  - - - - - -

# -- Fourier
base_fourier <- create.fourier.basis(rangeval=c(0, 1), nbasis=3)
plot(base_fourier)
# -- B-spline
par(mar=c(2,2,2,2))
breaks1 <- seq(0, 1, length=11)
```

```r
base_sp1 <- create.bspline.basis(rangeval=c(0,1), nbasis=13,
breaks=breaks1)
plot(base_sp1, lty=1)
# -- Exponencial
base_exp <- create.exponential.basis(rangeval=c(0,0.6), nbasis=4)
plot(base_exp)
# -- Potencia
base_pot <- create.power.basis(rangeval = c(0,2), nbasis=4)
plot(base_pot)
# -- Polinómicaa (expansiónn bases monómicas)
base_pol <- create.monomial.basis(rangeval=c(0,1.2), nbasis=4)
plot(base_pol)

# - - - - - - - - - - - - - - - - - - -  - - - -  -
# ---------- ACP: Trabajo con gait data ---------
# - - - - - - - - - - - - - - - - - - -  - - - -  -
gait

gaitrange = c(0,1)
gaitbasis <- create.fourier.basis(gaitrange, nbasis=21)
gaittime <- as.numeric(dimnames(gait)[[1]])*20
gaitrange <- c(0,20)

#  Operador diferencial de aceleración armónica
harmaccelLfd <- vec2Lfd(c(0, (2*pi/20)^2, 0), rangeval=gaitrange)
#  Creaci□e las bases
gaitbasis <- create.fourier.basis(gaitrange, nbasis=21)
#  Suavizado de los datos
gaitfd <- smooth.basisPar(gaittime, gait,
                          gaitbasis, Lfdobj=harmaccelLfd, lambda=1e-
2)$fd

names(gaitfd$fdnames) <- c("Normalized time", "Child", "Angle")
gaitfdPar  <- fdPar(gaitbasis, harmaccelLfd, lambda=1e-2)

# Se trata de datos bivariantes (ángulo de la cadera y de la cintura),
usamos los de la cadera.
hipfd  <- gaitfd[,1]
hipgaitpca.fd <- pca.fd(hipfd, nharm=4, gaitfdPar)
plot(hipgaitpca.fd$harmonics)
plot.pca.fd(hipgaitpca.fd)
hipfd$rangeval=c(0,1)

# Armónicos (funciones de peso)
par(mfrow=c(2,2), mar=c(2.5, 2.5, 2.5, 2.5))
plot.fd(hipfd, main="Ángulo de la cadera",cex.main=.8)
plot(hipgaitpca.fd$harmonics[1], main="PC 1 (71.6% de la
variabilidad)", cex.main=0.8)
abline(h=0, lty=2)
plot(hipgaitpca.fd$harmonics[2], main="PC 1 (12.2% de la
variabilidad)", cex.main=0.8)
abline(h=0, lty=2)
plot(hipgaitpca.fd$harmonics[3], main="PC 1 (8.5% de la
variabilidad)", cex.main=0.8)
abline(h=0, lty=2)

# CP como perturbaciones de la función media
par(mfrow=c(2,2), mar=c(3, 3, 3, 3))
plot.fd(hipfd, main="Ángulo de la cadera")
plot.pca.fd(hipgaitpca.fd, nharm=3, cex.main=c(0.8, 0.8, 0.8, 0.8))
```

```
# Mapa de scores
str(gaitfd)
par(mfrow=c(1,1))

scplot <-
as.data.frame(cbind(hipgaitpca.fd$scores[,1],hipgaitpca.fd$scores[,2])
)
names(scplot) <- c("CP1", "CP2")
plotscores <- ggplot(scplot, aes(x=CP1,
y=CP2))+geom_point(colour="blue", size=2)
plotscores + geom_label_repel(aes(label=ns, size=4),box.padding=0.15,
                              point.padding=0.15,
                              segment.colour="grey35",size=3)+
  theme_classic()

# Representación de las curvas destacadas
colgait <- c(rep("black", 39))
colgait[5] <- "red"
colgait [c(23,31,32,38,39)] <- "green"
ltyp <- c(rep(2,39))
ltyp[c(5, 23, 31, 32, 38, 39)] <- 1
lwd <- c(rep(1,39))
lwd[c(5, 23, 31, 32, 38, 39)] <- 2

plot.fd(hipfd, main="Ángulo de la cadera", col=colgait, lwd=lwd,
lty=ltyp, xlab="t", ylab="ngulo")

# - - - - - - - - - - - - - - - - - - -  - - - -  -
# ------------ Phase-plane-plots ------------
# - - - - - - - - - - - - - - - - - - -  - - - -  -

# -- Movimiento Armónico Simple sin(2*pi*t)
par(mar=c(4,4,4,4))
sin. <- expression(sin(2*pi*x))
D.sin <- D(sin., "x")
D2.sin <- D(D.sin, "x")

op <- par(pty="s")
par(mar=c(4,4,4,4))

# En este caso es más sencillo evaluar directamente las derivadas que
usar la función
with(data.frame(x=seq(0, 1, length=46)),
     plot(eval(D.sin), eval(D2.sin), type="l",
          xlim=c(-10, 10), ylim=c(-50, 50),
          xlab="Velocidad", ylab="Aceleraci□/span>) )

# Función
pi.2 <- (2*pi)
pi.2.2 <- pi.2^2

# Dibujo de los ejes e información sobre el estado de las energías
abline(h=0, col="gray55", lty=2)
abline(v=0, col="gray55", lty=2)
text(c(0,0), c(-47, 47), rep("Ec = 0, Ep = MAX", 2), cex=0.7)
text(c(-8.5,8.5), c(0,0), rep("Ec = MAX\nEp = 0", 2), cex=0.7)

par(op)

# ---
# -- log-nondurables
```

```
nondurables

# Creación de la base
lognonbases <- create.bspline.basis(rangeval=c(1919,2000),
                                    nbasis=979, norder=8)
# Objeto de acel.
LfdobjNonDur = int2Lfd(4)

# Obeto funcional
logNondurSm <- smooth.basisPar(argvals=index(nondurables),
                               y=log10(coredata(nondurables)),
fdobj=lognonbases,
                               Lfdobj=LfdobjNonDur, lambda=1e-11)

# Phase-plane-lots de los años 1929 y 1967
phaseplanePlot(1929, logNondurSm$fd, main="Índice de bienes
perecederos 1929", xlab="Velocidad", ylab="Aceleración")
phaseplanePlot(1967, logNondurSm$fd, main="Índice de bienes
perecederos 1967", xlab="Velocidad", ylab="Aceleración")
```

Anexo 2. Parte práctica

```
# - - - - - - - - - - - - - - - - - - -  - - - - -  -
# ----- Instalación de paquetes y librerías -----
# - - - - - - - - - - - - - - - - - - -  - - - - -  -
install.packages("fda")
install.packages("leaflet")
install.packages("maps")
install.packages("mapdata")
install.packages("readxl")
install.packages("ggplot2")
install.packages("ggrepel")
library(readxl)
library(leaflet)
library(RColorBrewer)
library(fda)
library(maps)
library(mapdata)
library(ggplot2)
library(ggrepel)

# - - - - - - - - - - - - - - - - - - -  - - - - -  -
# ------- Lectura y manipulación de datos -------
# - - - - - - - - - - - - - - - - - - -  - - - - -  -
tmedias <-  read_excel("E:/tfg/meteo.xlsx",
                       range = "G1:G32851")
tmedias <- t(tmedias)
temperaturas <- vector()
nombres <-
c("TAR1","BAR1","BAR2","GER","GUI1","GUI2","VIZ","CANT1","CANT2","AST1
",

"AST2","AST3","CRN1","CRN2","CRN3","PON1","PON2","LUG","LEO2","ORE",

"SOR","BUR","VALL1","AVI1","MAD3","SEG","VALL2","ZAM","LEO1",
             "SAL1","SAL2","GUA2","MAD2","GUA1","MAD4","MAD1","MAD5",
```

```
"TOL","CAC1","CR","CAD1","HUEL2","HUEL1","SEV2","SEV1","CORD",

"JAE","MLG","GRA1","GRA2","ALM2","CAD2","ALM1","CAC2","BAD","AVI2",

"PAL","RIO","ALA","ALI1","ALI2","CAS","VALE1","VALE2","NAV1","NAV2",

"LER","TAR2","MUR1","MUR2","MUR3","HUES","TER1","TER2","ZAR2","ZAR1",

"ALB1","ALB2","CUE","IB1","IB2","IB3","LPA1","LPA2","LPA3","TEN1",
             "TEN2","TEN3","CEU","MEL")

coordenadas <- read_excel("E:/tfg/coordenadas.xlsx",
                          range="A1:N91")
coordenadas <- coordenadas[,c(1,2,9,13,14)]

# - - - - - - - - - - - - - - - - - - -  - - - -  -
# ----- Mapas de estaciones meteorológicas ------
# - - - - - - - - - - - - - - - - - - -  - - - -  -

## MAPS

color <- c(rep("red", 90), rep("blue",90))
df <- data.frame(coordenadas$lonporc, coordenadas$latporc)
leaflet(df) %>% addTiles()%>%

  addCircles(lng = ~coordenadas$lonporc, lat = ~coordenadas$latporc,
weight=1,
             color=~color,#"red",
             radius=10000,
             popup=~coordenadas$NOMBRE
  )

## Mapa de altitudes
elevation <- getData('alt', country='ESP')

x <- terrain(elevation, opt = c("slope", "aspect"), unit = "degrees")
slope <- terrain(elevation, opt = "slope")
aspect <- terrain(elevation, opt = "aspect")
plot(x)
hill <-hillShade(slope, aspect, 40, 270)

par(mfrow=c(1,1))
plot(hill, col = grey(0:100/100), legend = FALSE, main = "Estaciones
meteorológicas por altitud")
plot(elevation, col = rainbow(25, alpha = 0.35),legend = FALSE, add =
TRUE)
for(i in 1:88){
  if(coordenadas$ALTITUD[i]<500){
    points(coordenadas$lonporc[i], coordenadas$latporc[i], pch=24,
col="black", bg="yellow")
  }else{
    if(coordenadas$ALTITUD[i]<1000){
      points(coordenadas$lonporc[i], coordenadas$latporc[i], pch=24,
col="black", bg="orange")
    }else{
      if(coordenadas$ALTITUD[i]<1500){
        points(coordenadas$lonporc[i], coordenadas$latporc[i], pch=24,
col="black", bg="red")
      }else{
```

```
        points(coordenadas$lonporc[i], coordenadas$latporc[i], pch=24,
col="black", bg="blue")
      }
    }
  }
}
legend("bottomleft", legend=c("0-500 metros", "500-1000 metros",
"1000-1500 metros", "Más de 1500 metros"), pch=24, cex=0.7,
col="black", bty="n",
pt.bg=c("yellow","orange","red","blue"),fill="white", border="white",
pt.cex=1)


# - - - - - - - - - - - - - - - - - - -  - - - -  -
# ----- Separación de los datos por estaciones: Matriz 90x365 -----
# - - - - - - - - - - - - - - - - - - -  - - - -  -

for(i in 1:90){
  previo <- 365*(i-1)+1
  fin <- 365*i
  temperaturas <- cbind(temperaturas, tmedias[previo:fin])
}
colnames(temperaturas) <- c(nombres)

# -- Plots representativos de distintas zonas
temperaturas <- as.data.frame(temperaturas)
ylim1 <- c(-15,35)
attach(temperaturas)

par(mfrow=c(1,1))
colores <- c("red","orange","blue","green","skyblue")
plot(LPA1, type="l", col=colores[1], ylim= ylim1, xlab = "D", ylab =
"Temperatura media", lwd=2, main="Temperatura media diaria")
points(VALE1, type="l", col=colores[2], lwd=2)
points(MAD1, type="l", col=colores[3], lwd=2)
points(AST1, type="l", col=colores[4], lwd=2)
points(CAD1, type="l", col=colores[5], lwd=2)

estaciones <- c("LPA1", "VALE1","MAD1","AST1","CAD1")
legend("bottom", y.intersp=0.6, col=colores,legend=estaciones,
cex=0.9, bty="n", lwd=4, horiz=TRUE)

# - - - - - - - - - - - - - - - - - - -  - - - -  -
# ---- Suavizado de las curvas de temperatura ----
# - - - - - - - - - - - - - - - - - - -  - - - -  -
# NÚMERO DE BASES: 9

# --- Parte 1: plots
## Nmero de breaks y bases
# -> Hay que respetar: (num. breaks) + (grado) = (num. bases) + 2
bks <- 7
nbases <- bks+2
breaks1=seq(0,365,length=bks)

## Creacióe la base
base_prueba_1 <- create.bspline.basis(rangeval=c(0,365),
nbasis=nbases, breaks=breaks1)
plot(base_prueba_1)
argvals_1 <- 1:365

## Representación de las curvas
```

```
daytempfd_media <- smooth.basis(argvals=argvals_1, y=media_diaria,
fdParobj=base_prueba_1)
plot.fd(daytempfd_media$fd, col=colores[1], ylim=ylim1, lwd=3,
xlab="D", ylab="Temperatura", main="Temperatura media diaria")

daytempfd_LPA1 <- smooth.basis(argvals=argvals_1, y=LPA1,
fdParobj=base_prueba_1)
plot.fd(daytempfd_LPA1$fd, col=colores[1], ylim=ylim1, lwd=3,
xlab="D", ylab="Temperatura", main="Temperatura media diaria")
daytempfd_VALE1 <- smooth.basis(argvals=argvals_1, y=VALE1,
fdParobj=base_prueba_1)
plot.fd(daytempfd_VALE1$fd, col=colores[2], ylim=ylim1, lwd=3,
add=TRUE)
daytempfd_MAD1 <- smooth.basis(argvals=argvals_1, y=MAD1,
fdParobj=base_prueba_1)
plot.fd(daytempfd_MAD1$fd, col=colores[3], ylim=ylim1, lwd=3,
add=TRUE)
daytempfd_AST1 <- smooth.basis(argvals=argvals_1, y=AST1,
fdParobj=base_prueba_1)
plot.fd(daytempfd_AST1$fd, col=colores[4], ylim=ylim1, lwd=3,
add=TRUE)
daytempfd_CAD1 <- smooth.basis(argvals=argvals_1, y=CAD1,
fdParobj=base_prueba_1)
plot.fd(daytempfd_CAD1$fd, col=colores[5], ylim=ylim1, lwd=3,
add=TRUE)

legend("bottom", y.intersp=0.6, col=colores,legend=estaciones,
cex=0.9, bty="n", lwd=4, horiz=TRUE)

## Representaci□e todas las curvas
mis.colores <- colorRampPalette(c("yellow", "green", "blue"))

temp <- as.matrix(temperaturas)
daytempfd_2013 <- smooth.basis(argvals=argvals_1, y=temp,
fdParobj=base_prueba_1)
plot.fd(daytempfd_2013$fd, col=rainbow(6), ylim=ylim1, lwd=2,
main="Temperatura media diaria", xlab="D", ylab="Temperatura")

# - - - - - - - - - - - - - - - - - - -  - - - -  -
# ------ Análisis de componentes principales ----
# - - - - - - - - - - - - - - - - - - -  - - - -  -
PCA2013 <- pca.fd(daytempfd_2013$fd, nharm=2, centerfns=TRUE)
par(mfrow=c(2,1), mar=c(2, 2, 2, 2))#)
plot.pca.fd(PCA2013, nharm=2)
par(mfrow=c(1,1), mar=c(2, 2, 2, 2))#)
plot.pca.fd(PCA2013, nharm=2)
par(mfrow=c(1,1))
plot(PCA2013$harmonics)
scores <- PCA2013$scores

## MAPA DE SCORES
scplot <- as.data.frame(cbind(scores[,1],scores[,2]))
names(scplot) <- c("CP1", "CP2")
plotscores <- ggplot(scplot, aes(x=CP1,
y=CP2))+geom_point(colour="blue", size=2)
plotscores + geom_label_repel(aes(label=nombres),box.padding=0.15,
                              point.padding=0.15,
                              segment.colour="grey35",size=2)+
  theme_classic()

## Ahora con clusters:
```

```
set.seed(4)
fit <- kmeans(scores,6)
clust_pca <- as.factor(fit$cluster)

scplot <- as.data.frame(cbind(scores[,1],scores[,2]))
names(scplot) <- c("CP1", "CP2")
plotscores <- ggplot(scplot, aes(x=CP1,
y=CP2))+geom_point(colour=as.factor(clust_pca), size=2)
plotscores + geom_label_repel(aes(label=nombres),box.padding=0.15,
                              point.padding=0.15,
                              segment.colour="grey35",size=2)+
  theme_classic()

## Representación de los clusters en el mapa
par(mfrow=c(2,1))
par(mar=c(0, 0, 0, 0))
layout(matrix(c(1,2,3,2), 2, 2, byrow = TRUE),
       widths=c(1,3), heights=c(2,1))
plot(1,1,type="n",axes=FALSE)
par(mar=c(0, 0, 0, 0))
map("worldHires",col="white",fill=TRUE,  xlim=c(-
11.72388889,4.215556), ylim=c(35.27778,43.56694), main="Score 1")
points(coordenadas$lonporc,coordenadas$latporc,col="black",
pch=21,xlab="Long",ylab="Lat",cex=2, bg=as.factor(clust_pca))

par(mar=c(0, 0, 0, 0))
map("worldHires",col="white",fill=TRUE,  xlim=c(-17.88889,-
13.60027778), ylim=c(27.81889,28.95194))
points(coordenadas$lonporc,coordenadas$latporc,col="black",
pch=21,xlab="Long",ylab="Lat",cex=2, bg=as.factor(clust_pca))

pcasc1 <- PCA2013$scores[,1]
# mínimo=-152.676, máximo=118.432
pcasc2 <- PCA2013$scores[,2]
# mínimo=-52.13074, máximo=41.569

## Creación de códigoo de colores para los scores

grad <-  scale_colour_gradient(colours=c("blue", "red"))
color.gradient <- function(x, colors=c("blue", "white", "red"),
colsteps=10000) {
  return( colorRampPalette(colors) (colsteps) [ findInterval(x,
seq(min(x),max(x), length.out=colsteps)) ] )
}

## Score 1
scbase1 <- seq(from=-155,to=155, by=0.01)
colpasc1 <- color.gradient(scbase1)
cal <- c(rep(0,90))

for(i in 1:90){
  a <- 5
  for(j in 1:length(scbase1)){

    if(abs(scbase1[j]-pcasc1[i])<a){
      a <- abs(scbase1[j]-pcasc1[i])
      cal[i] <- j
    }
  }
}
```

```
## Score 2
pcasc2 <- PCA2013$scores[,2]

grad2 <-  scale_colour_gradient(colours=c("blue", "red"))

scbase2 <- seq(from=-55,to=55, by=0.01)
colpasc2 <- color.gradient(scbase2)
cal2 <- c(rep(0,90))

for(i in 1:90){
  a <- 5
  for(j in 1:length(scbase2)){

    if(abs(scbase2[j]-pcasc2[i])<a){
      a <- abs(scbase2[j]-pcasc2[i])
      cal2[i] <- j
    }
  }
}


#######


# REPRESENTACIÓN EN MAPA: PCA 1
lata <- min(coordenadas$latporc)
latb <- max(coordenadas$latporc)
lona <- min(coordenadas$lonporc)
lonb <- max(coordenadas$lonporc)


par(mfrow=c(2,1))
par(mar=c(0, 0, 0, 0))
layout(matrix(c(1,2,3,2), 2, 2, byrow = TRUE),
       widths=c(1,3), heights=c(2,1))
plot(1,1,type="n",axes=FALSE)
par(mar=c(0, 0, 0, 0))
map("worldHires",col="white",fill=TRUE,  xlim=c(-
11.72388889,4.215556), ylim=c(35.27778,43.56694), main="Score 1")
points(coordenadas$lonporc,coordenadas$latporc,col="black",
pch=21,xlab="Long",ylab="Lat",cex=2, bg=colpasc1[cal])
par(mar=c(0, 0, 0, 0))
map("worldHires",col="white",fill=TRUE,  xlim=c(-17.88889,-
13.60027778), ylim=c(27.81889,28.95194))
points(coordenadas$lonporc,coordenadas$latporc,col="black",
pch=21,xlab="Long",ylab="Lat",cex=2, bg=colpasc1[cal])

# REPRESENTACIÓN EN MAPA: PCA 2
par(mfrow=c(2,1))
par(mar=c(0, 0, 0, 0))
layout(matrix(c(1,2,3,2), 2, 2, byrow = TRUE),
       widths=c(1,3), heights=c(2,1))
plot(1,1,type="n",axes=FALSE)
par(mar=c(0, 0, 0, 0))
map("worldHires",col="white",fill=TRUE,  xlim=c(-
11.72388889,4.215556), ylim=c(35.27778,43.56694))
points(coordenadas$lonporc,coordenadas$latporc,col="black",
pch=21,xlab="Long",ylab="Lat",cex=2, bg=colpasc2[cal2])
par(mar=c(0, 0, 0, 0))
map("worldHires",col="white",fill=TRUE,  xlim=c(-17.88889,-
13.60027778), ylim=c(27.81889,28.95194))
```

```
points(coordenadas$lonporc,coordenadas$latporc,col="black",
pch=21,xlab="Long",ylab="Lat",cex=2, bg=colpasc2[cal2])

# - - - - - - - - - - - - - - - - - - - - - - - - - - -
# ----- Phase-plane plot -----
# - - - - - - - - - - - - - - - - - - - - - - - - - - -
meses <- c(0,31,59,90,120,151,181,212,243,273,304,334,365)
par(mfrow=c(2,1), mar=c(2,2,2,2))
i_mes <- c("j","F","m","A","M","J","J","A","S","O","N","D")

#
phaseplanePlot(argvals_1, daytempfd_LPA1$fd, main="Phase-plane plot
LPA1",xlim=c(-0.4,0.4), ylim=c(-0.005, 0.02))
plot.fd(daytempfd_LPA1$fd, col=colores[1], ylim=ylim1, lwd=3,
xlab="D", ylab="Temperatura")
abline(v=meses, lty=2, col="gray47")
abline(h=c(-10,0,10,20,30), col="gray90")
text(x=meses, y=-5,i_mes)

#
phaseplanePlot(argvals_1, daytempfd_VALE1$fd,main="Phase-plane plot
VALE1",xlim=c(-0.4,0.4), ylim=c(-0.005, 0.02))
plot.fd(daytempfd_VALE1$fd, col=colores[2], ylim=ylim1, lwd=3,
xlab="D", ylab="Temperatura")
abline(v=meses, lty=2, col="gray47")
abline(h=c(-10,0,10,20,30), col="gray90")
text(x=meses, y=-5,i_mes)

#
phaseplanePlot(argvals_1, daytempfd_MAD1$fd,main="Phase-plane plot
MAD1",xlim=c(-0.4,0.4), ylim=c(-0.005, 0.02))
plot.fd(daytempfd_MAD1$fd, col=colores[3], ylim=ylim1, lwd=3,
xlab="D", ylab="Temperatura")
abline(v=meses, lty=2, col="gray47")
abline(h=c(-10,0,10,20,30), col="gray90")
text(x=meses, y=-5,i_mes)

#
phaseplanePlot(argvals_1, daytempfd_AST1$fd,main="Phase-plane plot
AST1",xlim=c(-0.4,0.4), ylim=c(-0.005, 0.02))
plot.fd(daytempfd_AST1$fd, col=colores[4], ylim=ylim1, lwd=3,
xlab="D", ylab="Temperatura")
abline(h=c(-10,0,10,20,30), col="gray90")
abline(v=meses, lty=2, col="gray47")
text(x=meses, y=-5,i_mes)

#
phaseplanePlot(argvals_1, daytempfd_CAD1$fd, main="Phase-plane plot
CAD1",xlim=c(-0.4,0.4), ylim=c(-0.005, 0.02))
plot.fd(daytempfd_CAD1$fd, col=colores[5], ylim=ylim1, lwd=3,
xlab="D", ylab="Temperatura")
abline(h=c(-10,0,10,20,30), col="gray90")
abline(v=meses, lty=2, col="gray47")
text(x=meses, y=-5,i_mes)

# Todos en un mismo plot:
par(mfrow=c(3,2), mar=c(2.2, 2.2, 2.2, 2.2))
phaseplanePlot(argvals_1, daytempfd_LPA1$fd, xlim=c(-0.4,0.4),
ylim=c(-0.005, 0.02),main=estaciones[1])
phaseplanePlot(argvals_1, daytempfd_VALE1$fd, xlim=c(-0.4,0.4),
ylim=c(-0.005, 0.02),main=estaciones[2])
```

```
phaseplanePlot(argvals_1, daytempfd_MAD1$fd, xlim=c(-0.4,0.4),
ylim=c(-0.005, 0.02),main=estaciones[3])
phaseplanePlot(argvals_1, daytempfd_AST1$fd, xlim=c(-0.4,0.4),
ylim=c(-0.005, 0.02),main=estaciones[4])
phaseplanePlot(argvals_1, daytempfd_CAD1$fd, xlim=c(-0.4,0.4),
ylim=c(-0.005, 0.02),main=estaciones[5])
plot.fd(daytempfd_LPA1$fd, col=colores[1], ylim=ylim1, lwd=3,
xlab="D", ylab="Temperatura", main="Temperatura media diaria")
plot.fd(daytempfd_VALE1$fd, col=colores[2], ylim=ylim1, lwd=3,
add=TRUE)
plot.fd(daytempfd_MAD1$fd, col=colores[3], ylim=ylim1, lwd=3,
add=TRUE)
plot.fd(daytempfd_AST1$fd, col=colores[4], ylim=ylim1, lwd=3,
add=TRUE)
plot.fd(daytempfd_CAD1$fd, col=colores[5], ylim=ylim1, lwd=3,
add=TRUE)

legend("bottom", y.intersp=0.6, x.intersp
=0.2,col=colores,legend=estaciones, cex=0.9, bty="n", lwd=4,
horiz=TRUE)
```