Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Estadística

ANÁLISIS FUNCIONAL DE DATOS DE INGRESOS URGENTES EN PSIQUIATRÍA

Autora: Paula García Alonso
Tutoras: Yolanda Larriba González y Andrea Pérez Álvarez
Año: 2024/2025

A mi familia, por su apoyo incondicional.

A Miguel, por su paciencia y por acompañarme en cada paso.

Y muy especialmente a mis tutoras, Yolanda y Andrea,

por su orientación y dedicación.

Índice

Resumen 1

Abstract 1

Índice de figuras 2

Índice de tablas 3

1 Introducción **4**

1.1 Contextualización 4

1.2 Objetivo 7

1.3 Asignaturas relacionadas 8

2 Metodología estadística para el FDA **9**

2.1 Notación y conceptos básicos 9

2.2 Suavizado de una función 11

2.2.1 Bases de Fourier en FDA 12

2.2.2 Bases de *spline* en FDA 13

2.2.3 Imputación 17

2.3 Estadística descriptiva funcional 18

2.4 Análisis de Componentes Principales Funcional 22

2.4.1 PCA: revisión del enfoque clásico 22

2.4.2 FPCA: Extensión del PCA al contexto funcional 24

2.4.3 MFPCA: Análisis de Componentes Principales Funcional Multivariante . 26

3 Análisis de datos y resultados **29**

3.1 Descripción de los datos 29

3.2 Procesado e imputación de datos 32

3.2.1 Tratamiento para valores ausentes 32

3.2.2 Tratamiento para datos censurados 34

3.3 Suavizado de los datos mediante bases *B-splines* 35

3.4 Análisis descriptivo funcional de las variables funcionales observadas 35

3.4.1 Análisis descriptivo para *tMedia* 36
3.4.2 Análisis descriptivo para *prec* 37
3.4.3 Análisis descriptivo para O_3 38
3.4.4 Análisis descriptivo para PM_{10} 39
3.5 Implementación del FPCA y MFPCA 39
3.5.1 FPCA en variables climatológicas y de contaminación 40
3.5.2 MFPCA aplicado a las variables seleccionadas 47

4 Conclusiones. Trabajo futuro. 51

Referencias 54
Anexo 56
1. Conjunto de datos growth: descripción para ejemplificar la metodología del TFG 56
2. Metodología modelo Tobit 56
3. Estadísticos descriptivos funcionales 58
4. FPCA 64
5: Código R 69

Resumen

En este Trabajo Fin de Grado se analiza la relación entre condiciones ambientales y los ingresos urgentes psiquiátricos desde la perspectivca del Análisis de Datos Funcionales.

Se analiza un conjunto de datos formado por 4104 pacientes ingresados en la Unidad de Psiquiatría del Hospital de León (2011-2022), con registro de temperaturas, humedad y contaminantes desde el día del ingreso hasta 20 días previos. Estas series temporales se modelizan como funcionaes continuas, capturando la dinámica completa en lugar de valores aislados.

Se realiza un análisis descriptivo funcional de las variables permitiendo identificar tendencias temporales, zonas de mayor o menor dispersión y patrones de dependencia entre los días previos a la hospitalización. A continuación, se aplican técnicas de reducción de la dimensionalidad para extraer los principales patrones de variación. En esta fase exploratoria, se identifican perfiles funcionales diferenciados por nivel, variabilidad local y momentos de mayor exposición. En conjunto, estos patrones explican una gran parte de la variabilidad observada y sientan las bases para futuras estrategias de modelización predictiva más elaboradas.

Palabras clave: datos funcionales, suavizado, imputación, *B-splines*, componentes principales funcionales, variables meteorológicas, contaminación atmosférica.

Abstract

This Final Degree Project analyzes the relationship between environmental conditions and emergency psychiatric admissions from the perspective of Functional Data Analysis.

A dataset consisting of 4104 patients admitted to the Psychiatric Unit of León Hospital (2011-2022) is analyzed, with records of temperature, humidity, and pollutants from the day of admission up to 20 days prior. These time series are modeled as continuous functions, capturing the complete dynamics rather than isolated values.

A descriptive functional analysis of the variables is performed, allowing for the identification of temporal trends, areas of greater or lesser dispersion, and patterns of dependence between days prior to hospitalization. Dimensionality reduction techniques are then applied to extract the main patterns of variation. In this exploratory phase, functional profiles are identified, differentiated by level, local variability, and times of greatest exposure. Together, these patterns explain a large part of the observed variability and lay the groundwork for more sophisticated predictive modeling strategies in the future.

Key words: functional data, smoothing, imputation, *B-splines*, functional principal components, meteorological variables, air pollution.

Índice de figuras

Figura 2.1 *Gráfico de funciones subyacentes.* 10
Figura 2.2 *Representación gráfica del sistema de tres funciones base de Fourier en [0,1].* 13
Figura 2.3 *Gráfico de la función base B-spline.* 16
Figura 2.4 *Gráfico de la función base B-spline periódicas* 17
Figura 2.5 *Gráfico media funcional niñas.* 19
Figura 2.6 *Gráfico varianza funcional niñas.* 20
Figura 2.7 *Representaciones de la covarianza funcional de la altura en niñas* 21
Figura 2.8 *FPCA para las curvas de altura de las niñas. Creado a partir del conjunto de datos growth de la librería fda en R* 26

Figura 3.1 *Curvas anuales de $PM_{2,5}$ por año.* 33
Figura 3.2 *Representación de los estadísticos descriptivos funcionales para la variable tMedia* 36
Figura 3.3 *Representación de los estadísticos descriptivos funcionales para la variable prec* 37
Figura 3.4 *Representación de los estadísticos descriptivos funcionales para la variable O_3* 38
Figura 3.5 *Representación de los estadísticos descriptivos funcionales para la variable PM_{10}* 39
Figura 3.6 *Scores de las dos primeras componentes principales de la variable tMedia en función de la estación en la que ingresa* 41
Figura 3.7 *Dos primeras autofunciones de la variable tMedia* 41
Figura 3.8 *Scores de las dos primeras componentes principales de la variable prec en función de la estación en la que ingresa* 42
Figura 3.9 *Dos primeras autofunciones de la variable prec* 42
Figura 3.10 *Scores PC1 vs PC2 de la variable prec separando por individuos imputados y no imputados* 43

Figura 3.11 *Scores de las dos primeras componentes principales de la variable O_3 en función de la esatción en la que ingresa* 44
Figura 3.12 *Tres primeras autofunciones de la variable O_3* 44
Figura 3.13 *Scores de las dos primeras componentes principales de la variable PM_{10}* en función de la estación en la que ingresan 45
Figura 3.14 *Cuatro primeras autofunciones de la variable PM_{10}* 45
Figura 3.15 *Outlier PM_{10}* 46
Figura 3.16 *Curvas suavizadas de PM_{10}: outlier vs resto* 46
Figura 3.17 *Scores PC1 vs PC2 de la variable PM_{10} sin los outliers* 47
Figura 3.18 *Primera componente principal funcional* 48
Figura 3.19 *Segunda componente principal funcional* 48
Figura 3.20 *Tercera componente principal funcional* 49
Figura 3.21 *Scores PC1 vs PC3 por estación* 50

Índice de tablas

Tabla 3.1 *Definición de las variables* 30
Tabla 3.2 *Resumen de ausencias* 32

Capítulo 1

Introducción

1.1. Contextualización

Históricamente, las condiciones ambientales han sido consideradas un factor clave en el estudio de diversas enfermedades, incluidas las psiquiátricas. La relación entre el medio ambiente y la salud humana ha sido documentada en múltiples estudios, especialmente en los últimos años. (Hu, Xu, Wang, Su, y Guo, 2023) (Rizavas et al., 2023). Actualmente, se reconoce que la salud mental y el bienestar están influenciados por una interacción compleja de factores genéticos, psicológicos, sociales, de estilo de vida y ambientales (Ordóñez-Iriarte, 2020) (Son y Shin, 2021).

El presente Trabajo Fin de Grado (TFG) se enmarca en una línea de investigación desarrollada por la Unidad de Psiquiatría del Hospital de León. El objetivo es analizar la posible asociación entre factores ambientales (meteorológicos y de contaminación) y la salud mental, mediante el estudio del número de ingresos urgentes en psiquiatría, prestando especial atención a la identificación de patrones temporales relacionados con condiciones climáticas y ambientales específicas.

La línea de trabajo se inició en 2023, tras la implementación de una nueva metodología docente en el servicio de Psiquiatría. Se propusieron distintos temas de investigación a los profesionales en formación (MIR/EIR), con el fin de promover la actividad investigadora. A una residente se le asignó el análisis del posible efecto de variables ambientales sobre los ingresos psiquiátricos. La elección pudo estar motivada por la creencia extendida de un aumento de descompensaciones durante los cambios estacionales.

En septiembre de 2023 se creó una base de datos que incluía diagnósticos clínicos de pacientes e información meteorológica procedente de la Agencia Estatal de Meteorología (AEMET). La recogida de datos fue autorizada por el Comité de Ética de Investigación con medicamentos (CEIm). El análisis preliminar comenzó en enero de 2024.

Ante la falta de resultados concluyentes, se revisó la literatura científica. Se identificó un Trabajo de Fin de Máster (Gómez González, 2022) que utilizaba análisis de retardos o *lags*, técnica que considera la posible influencia de condiciones ambientales de días anteriores al ingreso. Esta metodología fue incorporada en marzo de 2024, junto con la inclusión de nuevas variables, tanto meteorológicas como de contaminación ambiental, con distintos niveles de retardo. Esta segunda fase se desarrolló hasta abril de 2024, cuando se inició el análisis estadístico.

Dicho estudio dejó abierta la posibilidad de continuar con su desarrollo y análisis. En este trabajo, se retoma y amplía la investigación, con el fin de profundizar en esta relación, pues no sólo tiene un interés científico, sino también un claro valor práctico para la planificación sanitaria. Comprender esta relación podría facilitar una mejor anticipaciónde de picos asistenciales, optimizando así la asignación de recursos humanos y materiales. Además, este conocimiento permitiría mejorar la toma de decisiones clínicas, ajustar la temporalización de ciertas terapias o intervenciones y diseñar estrategias preventivas en momentos de mayor riesgo.

En el estudio se recogen dos tipos de datos. Por un lado, los datos clínicos de los pacientes ingresados de urgencia en la Unidad de Psiquiatría del Hospital de León, entre los que se incluyen edad, el sexo, la fecha de ingreso y el diagnóstico. Por otro lado, se registraron para cada ingreso variables ambientales, como los niveles de contaminación atmosférica (e.g., monóxido de carbono (CO), óxido nítrico (NO), dióxido de nitrógeno (NO_2) y ozono (O_3), entre otros) e indicadores meteorológicos tales como la temperatura máxima y mínima, junto con otras variables relevantes relacionadas con las condiciones climáticas, tanto el día del ingreso como en los 20 días previos. A este tipo de registros, susceptibles de ser considerados más como curvas que como una respuesta escalar o un vector, se les conoce como datos funcionales.

Desde una perspectiva estadística, la explotación de este conjunto de datos con el objetivo de analizar el número de ingresos urgentes en la Unidad de Psiquiatría plantea un desafío metodológico significativo debido a que se combinan datos escalares y categóricos con datos funcionales, lo que requiere enfoques específicos para su integración en el análisis.

El análisis de datos funcionales (FDA, por sus siglas en inglés *Functional Data Analysis*) es una rama de la estadística que se centra en el estudio de datos que pueden representarse como curvas o funciones continuas. Aunque los datos subyacentes son funciones (de dimensionalidad infinita), los datos observados suelen ser una representación discretizada de estas. De esta manera, cada curva queda representada por un vector de dimensión finita. Este enfoque resulta especialmente útil cuando las observaciones no son valores aislados, sino que se extienden a lo largo de un dominio continuo, como ocurre en curvas temporales (por ejemplo, los registros de temperatura y contaminación en los 20 días anteriores al ingreso del paciente) o, en extensiones

más complejas, en superficies definidas sobre dominios espaciales, como distribuciones geográficas o imágenes médicas. El análisis estadístico y la modelización de este tipo de datos difiere sustancialmente del análisis convencional, requiriendo la adaptación de metodologías clásicas, como la reducción de dimensionalidad o el agrupamiento, a las características específicas de los datos funcionales. De no ser así, es fácil ver que los procedimientos clásicos proporcionan resultados imprecisos, poco informativos o erróneos.

El FDA es un enfoque relativamente reciente que comenzó a consolidarse en la década de 1980, con los trabajos pioneros de James O. Ramsay (Ramsay y Silverman, 2002), quien acuñó el término y estableció las bases metodológicas del campo. Su relevancia ha crecido paralelamente al aumento de la disponibilidad de datos de alta resolución temporal, como los generados en medicina, economía, climatología o ingeniería. En estos contextos, FDA permite modelizar la evolución dinámica de un fenómeno, proporcionando una representación más fiel y completa de la información. La metodología ha sido ampliada con técnicas como regresión funcional, clasificación y *clustering*, entre otras. En la actualidad, FDA se aplica en áreas como neuroimagen, análisis del movimiento humano, estudios medioambientales y predicción en salud pública, consolidándose como una herramienta clave para abordar datos de alta dimensión con estructura temporal o espacial compleja (Chiou y Li, 2007; Greven, Crainiceanu, Caffo, y Reich, 2011; Wang et al., 2021).

El FDA presenta diversos desafíos metodológicos, siendo uno de los más relevantes la alta dimensionalidad inherente a este tipo de datos. En el FDA, cada observación se representa como una función continua (o curva) evaluada en un número elevado de puntos del dominio, lo que genera estructuras de datos altamente correlacionadas y de gran dimensión, dificultando su tratamiento estadístico directo. Para abordar este problema, es habitual recurrir a técnicas de reducción de dimensión, que permiten representar las funciones de forma más compacta sin perder la información esencial. Una de las técnicas más empleadas persiguiendo este objetivo es el Análisis de Componentes Principales Funcionales (FPCA, del inglés *Functional Principal Component Analysis*), que identifica las direcciones principales de variabilidad en el espacio funcional y permite expresar las curvas mediante un número reducido de componentes ortogonales. También se emplean técnicas como la expansión en bases funcionales, como las bases de Fourier o *B-splines*, ya que permiten aproximar las funciones observadas mediante combinaciones lineales de funciones base predefinidas, facilitando el análisis y el modelado.

Adicionalmente, los datos funcionales suelen estar afectados por fuentes de ruido y variabilidad no estructurada. Por ello, es fundamental aplicar técnicas de suavizado para recuperar la señal subyacente, así como abordar la presencia de valores ausentes mediante métodos apropiados

de imputación funcional, que preserven la continuidad y estructura de dependencia de las curvas.

Otro aspecto crucial es la resolución temporal o espacial con la que se registran las funciones. Contar con un número suficiente de puntos de observación, especialmente en las regiones de alta curvatura o cambios rápidos, resulta esencial para estimar correctamente la forma funcional y preservar las características relevantes del fenómeno estudiado.

Además, en muchos contextos, las observaciones funcionales no sólo varían en forma sino también en fase o localización temporal, por lo que resulta conveniente aplicar técnicas de alineamiento o registro temporal que separen la variabilidad estructural de la variabilidad asociada al desfase entre funciones.

Finalmente, cuando se analizan funciones multivariantes (es decir, múltiples funciones por sujeto), es necesario aplicar herramientas como el Análisis de Componentes Principales Funcionales Multivariante (MFPCA, del inglés *Multivariate Functional Principal Component Analysis*), que permite captar patrones simultáneos de variación entre varias variables funcionales observadas en dominios comunes.

Este TFG aborda estos desafíos mediante un enfoque metodológico y académico presentando algunas de las técnicas de FDA mencionadas, con el objetivo de proporcionar una aproximación a un marco analítico riguroso para el estudio de la influencia de factores climatológicos y de contaminación en los ingresos psquiátricos. Además de su relevancia metodológica, este trabajo puede aportar información valiosa para la planificación de recursos en salud mental, especialmente en el contexto del cambio climático y su posible impacto en la incidencia de crisis psiquiátricas.

1.2. Objetivo

El presente TFG tiene como objetivo general explorar la estructura y la variabilidad de las curvas temporales funcionales de las variables meteorológicas y de contaminación, con el fin de descubrir patrones de comportamiento y posibles relaciones entre las variables en los días previos al ingreso del sujeto.

Para lograrlo, se plantean los siguientes objetivos específicos:

1. Explorar la evolución funcional de cada variable meteorológica y de contaminación mediante estadísticas descriptivas funcionales, obteniendo curvas representativas que permitan comprender su dinámica a lo largo del tiempo.

2. Identificar los modos de variabilidad más influyentes de cada variable meteorológica y de contaminación a través de análisis de componentes principales funcional.

3. Analizar las posibles interacciones entre las variables mediante un análisis de componentes principales funcional multivariante, revelando combinaciones de fluctuaciones que conectan la meteorología con los niveles de contaminación.

4. Evaluar la eficacia y las limitaciones de las metodologías utilizadas en el contexto de datos funcionales, destacando sus aportaciones para el estudio de fenómenos meteorológicos y contaminantes atmosféricos en el ingreso urgente psiquiátrico.

1.3. Asignaturas relacionadas

Es importante señalar que, aunque no se ha cursado ninguna asignatura específica sobre datos funcionales en el grado, diversas asignaturas previas han sido fundamentales como base para poder llevar a cabo este análisis. Por ejemplo, el manejo profundo de R para el análisis de datos ha sido potenciado mediante los conocimientos adquiridos en *Computación Estadística* o *Métodos Estadísticos de Computación Intensiva*, lo que ha permitido implementar y automatizar procedimientos estadísticos avanzados. Asimismo, asignaturas como *Análisis de Datos* y *Análisis Multivariante* han facilitado la adaptación de métodos clásicos a técnicas propias del análisis de datos funcionales, posibilitando la aplicación de métodos como el análisis de componentes principales en series de datos continuos. Por otro lado, la asignatura *Métodos de Suavizado en Estadística* ha sido clave para adquirir los conocimientos necesarios sobre las técnicas de suavizado aplicables a este contexto, permitiéndome abordar problemas como la reducción de ruido en funciones observadas y el suavizado de curvas en series temporales. Finalmente, la formación en *Series Temporales* ha proporcionado las herramientas complementarias para analizar la dinámica temporal de los datos.

Capítulo 2

Metodología estadística para el FDA

2.1. Notación y conceptos básicos

En el análisis estadístico, cuando el fenómeno de interés se desarrolla de forma continua a lo largo del tiempo, resulta necesario adoptar un enfoque que capture dicha continuidad. En este contexto, el FDA representa un cambio de paradigma, ya que desplaza el foco desde observaciones discretas hacia funciones definidas sobre un dominio continuo.

Concretamente, en la práctica, para cada individuo ($i = 1, \ldots, n$) se observan los pares (t_{ij}, y_{ij}), con $i = 1, \ldots, n$, $j = 1, \ldots, J_i$, que representan las mediciones de una variable de interés (y) en cada individuo (i) en J_i instantes de tiempo (t). Este conjunto de datos proporciona una aproximación al comportamiento de una función subyacente continua y suave $x_i(\cdot)$, que aproxima de manera coherente las observaciones puntuales del sujeto i-ésimo. En adelante, por simplicidad, supondremos que los instantes de tiempo para los que se evalúa la variable coinciden para todos los individuos, y que el número de observaciones registradas es el mismo para todos los individuos, es decir, $t_{ij} = t_j$ para $j = 1, \ldots, J$, independiente del individuo i. Tanto la notación como las definiciones formales que se presentan a continuación han sido tomadas de (Ramsay y Silverman, 2002; Horváth y Kokoszka, 2012; Levitin, Nuzzo, Vines, y Ramsay, 2007).

Las curvas mostradas en la Figura 2.1 han sido obtenidas a partir de datos discretos y generadas a partir del conjunto de datos *growth* incluido en la librería *fda* de R. Para más detalles sobre este conjunto de datos, veáse la sección 1 del Anexo.

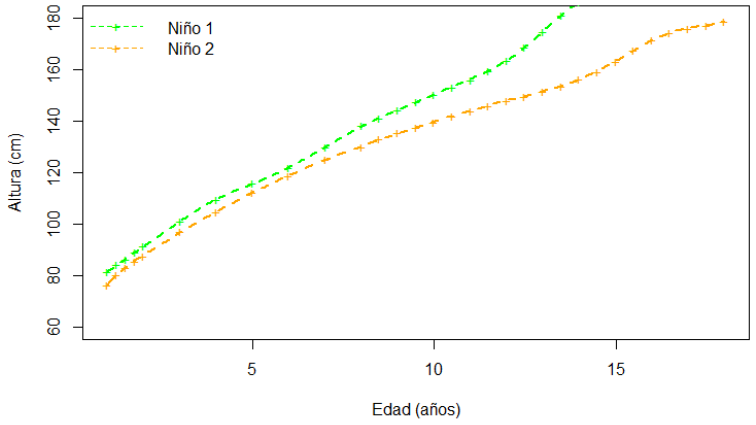


Figura 2.1. Curvas de crecimiento de dos niños, junto con los datos reales de altura marcados con el signo +. Creado a partir del conjunto de datos growth de la librería fda en R.

Definición 2.1.1. ***Dato funcional***

Un dato funcional es una función real definida sobre un intervalo compacto del conjunto de los reales. Es decir, si $T \subset \mathbb{R}$ representa el dominio de interés (por ejemplo, el tiempo), entonces un dato funcional es una aplicación

$$x : T \to \mathbb{R},$$

donde $x(t)$ describe el estado o la evolución del fenónemo en cada punto $t \in T$. Cada unidad observacional de la muestra no se modeliza como un vector, sino como una función entera que expresa de manera continua su comportamiento en el dominio considerado.

Definición 2.1.2. ***Conjunto de de datos funcionales***

Un conjunto de datos funcionales, en este marco, es en una colección de funciones $\{x_i(t)\}_{i=1}^n$, cada una de las cuales representa un individuo, una unidad experimental o una observación, y todas ellas están definidas sobre el mismo dominio continuo T.

Como se ha mencionado, las funciones $x(t)$ no suelen estar dadas de forma explícita, sino que se aproximan a partir de observaciones discretas y potencialmente ruidosas. Asumiendo que los datos discretos registrados para un sujeto i son un conjunto de J pares $(t_1, y_1), \ldots, (t_J, y_J)$, la filosofía del FDA se basa en considerar que cada observación i pasa a ser una función x_i que puede tomar valores en cualquier punto del dominio, de manera que cada punto t le corresponde un valor $x_i(t)$. Por ello, uno de los principales asceptos del FDA es la representación funcional de los datos: es decir, el proceso de construir una función suave $x(t)$ que aproxime las observaciones puntuales de un sujeto.

2.2. Suavizado de una función

El suavizado es una técnica que permite reducir el impacto del ruido o los errores de medición en los datos de una función. Su objetivo principal es resaltar las características esenciales de la función subyacente, eliminando las fluctuaciones aleatorias que no aportan información útil (Bowman y Azzalini, 1997).

En un escenario ideal, si los datos disponibles fueran completamente precisos, es decir, si cada valor observado se correspondiera exactamente con el valor real de la función, expresado como $y_{ij} = x_i(t_j),\ \forall i = 1, \ldots, n$ y $j = 1, \ldots, J$, entonces sería apropiado aplicar técnicas de interpolación para reconstruir la función orginal. Sin embargo, en la práctica esto rara vez ocurre. Es mucho más común que los datos estén contaminados por algún tipo de ruido o error. Esto puede modelizarse como:

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

Aquí, ε_{ij} representa el error o ruido, que normalmente se modeliza como una variable aleatoria con distribución normal, media cero y varianza constante σ^2 ($\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$) (Picón Llamas, 2019). A continuación, se presenta una defínción formal para una función suave en el contexto del FDA.

Definición 2.2.1. ***Función suave***

Se dice que $x(t)$ es una función suave si pertenece a una clase de funciones $C^m(T)$, es decir, si es m veces continuamente diferenciable en su dominio $t \in T \subset \mathbb{R}$.

Formalmente, la curvatura de una función se puede medir a través de su segunda derivada, la cual refleja la rapidez con la que cambia su pendiente. En zonas de alta curvatura (es decir, con variaciones abruptas), se necesita una mayor densidad de puntos para estimar con precisión la forma de la función. Si el muestreo es escaso, los métodos de suavizado pueden no detectar correctamente picos o inflexiones, generando estimaciones imprecisas. En cambio, en regiones suaves (de baja curvatura), se puede usar una menor densidad de puntos sin comprometer la precisión. Por ello, para optimizar los resultados del suavizado se precisa que la densidad de muestreo se adapte a la variabilidad local de la función (Simonoff, 2012).

A la hora de aplicar técnicas de suavizado, existen distintos enfoques para representar la función subyacente de forma que se ajuste a la naturaleza de los datos. Entre los métodos más utilizados destacan aquellos centrados en bases de Fourier y *splines*. En las siguientes subsecciones se describen ambos enfoques, haciendo énfasis en su fundamento teórico, su aplicabilidad y los criterios prácticos para la elección del número de componentes o funciones base, aspecto fundamental que influye en la capacidad del modelo para capturar la estructura real de los datos

sin sobreajustarse al ruido (Manso, 2019).

2.2.1. Bases de Fourier en FDA

Las bases de Fourier son especialmente adecuadas cuando se quiere representar funciones que muestran un comportamiento periódico. Se construyen a partir de funciones trigonométricas, como seno y coseno, que son por naturaleza periódicas (Ramsay y Silverman, 2002).

Definición 2.2.2. ***Sistema de bases de Fourier***

Una función $x(t)$ puede aproximarse mediante una serie de Fourier como:

$$x(t) \approx c_0 + \sum_{r=1}^{R} [c_{2r-1} \sin(r\omega t) + c_{2r} \cos(r\omega t)]$$

donde:

- $c_0, c_1, \ldots, c_{2R}$ *son los coeficientes de Fourier, definidos como:*

$$c_k = \frac{2}{\mathcal{T}} \int_0^{\mathcal{T}} x(t)\phi_k(t)\, dt$$

  siendo $\phi_k(t)$ la base ortonormal correspondiente (por ejemplo, $\sin(r\omega t)$ o $\cos(r\omega t)$).

- $\mathcal{T}$ *determina el período y se verifica la relación mediante $\omega = \frac{2\pi}{\mathcal{T}}$.*

El sistema de funciones base está dado por:

$$\left\{ \begin{array}{c} \phi_0(t) = 1 \\ \phi_{2r-1}(t) = \sin(r\omega t) \\ \phi_{2r}(t) = \cos(r\omega t) \end{array} \right\}_{r=1}^{R}$$

Estas funciones forman una base ortogonal en el intervalo $[0, \mathcal{T}]$.

Ejemplo

La Figura 2.2 ilustra un ejemplo de las tres primeras componentes de la base de Fourier utilizadas para aproximar una función con periodo $\mathcal{T} = [0, 1]$.

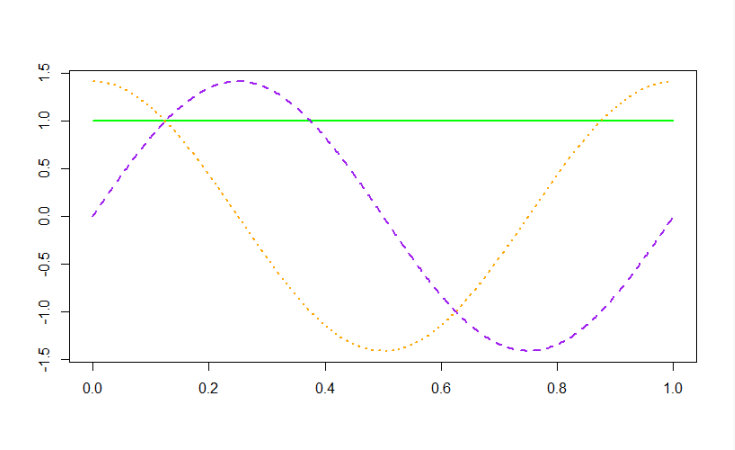


Figura 2.2. Tres primeras componentes de la base de Fourier definidas en el intervalo [0,1]. Creado con la funcion create.fourier.basis() de la librería fda.

Las bases de Fourier presentan varias ventajas cuando se está trabajando con funciones periódicas. En primer lugar, permiten representar estas funciones utilizando combinaciones de senos y cosenos que capturan ese patrón repetitivo con pocos términos. Además, existen algoritmos computacionales altamente eficientes, como la Transformada Rápida de Fourier (FFT, por sus siglas en inglés), que permiten calcular los coeficientes de la serie de forma rápida incluso para grandes volúmenes de datos (Ramsay y Silverman, 2002). La FFT reduce significativamente la complejidad computacional del cálculo y cuenta con implementaciones optimizadas que funcionan con diferentes tamaños de muestra y condiciones. Otra ventaja importante es que las funciones de Fourier se pueden derivar fácilmente. Como las derivadas de senos y cosenos siguen siendo funciones trigonométricas, resulta muy sencillo estudiar cómo cambia la función con el tiempo, lo cual es muy útil para analizar fenómenos dinámicos, como la velocidad o la aceleración de un proceso.

Sin embargo, el uso de estas bases tiene ciertas limitaciones. No son adecuadas cuando los datos presentan comportamientos irregulares, asimétricos o no periódicos, ya que este tipo de funciones no capturan bien las variaciones locales ni las singularidades. Por eso, en otros contextos es preferible emplear bases alternativas como son los *splines*.

2.2.2. Bases de *spline* en FDA

Las bases *spline* en FDA ofrecen una mayor sensibilidad a la estructura local de los datos, permitiendo capturar posibles irregularidades. Aunque existen versiones periódicas, su princi-

pal ventaja radica en la flexibilidad para modelizar comportamientos locales no necesariamente periódicos (Ramsay y Silverman, 2002).

Definición 2.2.3. ***Spline***

Sea una partición del intervalo $[a, b]$ *dada por los nodos:*

$$a = t_0 < t_1 < \cdots < t_n = b$$

Un spline de grado d *(o de orden* $d + 1$*) es una función* $S : [a, b] \to \mathbb{R}$ *tal que:*

- S *es* $d - 1$ *veces continuamente derivable en todo el intervalo.*
- *En cada subintervalo* $[t_i, t_{i+1}]$*, la función* S *coincide con un polinomio* $P_i(t)$ *de grado a lo sumo* d*:*

$$\forall i \in \{0, \ldots, n-1\}, \quad S(t) = P_i(t), \quad \textit{con } \deg(P_i) \leq d, \quad \forall t \in [t_i, t_{i+1}]$$

- *Las derivadas hasta orden* $d - 1$ *coinciden en los nodos* $t_1, \ldots, t_{n-1}$*, garantizando que no haya saltos ni quiebres en la transición entre los distintos polinomios.*

Los nodos t_i definen las posiciones en las que se conectan los distintos polinomios que componen el *spline*, y son esenciales porque controlan su flexibilidad local: al aumentar el número de nodos, se permite una mayor capacidad de adaptación a los datos. Por otro lado, el grado d del *spline* determina la suavidad y flexibilidad del *spline*: si $d = 0$, se obtiene una función por tramos constante; si $d = 1$, se construye un *spline* lineal formado por segmentos rectos; y si $d = 3$, se obtiene un *spline* cúbico, el más utilizado en la práctica por su equilibrio entre suavidad y flexibilidad.

La flexibilidad del *spline* depende del número de nodos (cuantos más haya, mejor se adaptarán a los datos) y del orden del polinomio que se use en cada tramo. Si no se colocan nodos interiores, el *spline* se convierte en un polinomio global, perdiendo su capacidad de adaptación local. Por eso, seleccionar correctamente la cantidad y la posición de los nodos es clave para obtener un buen ajuste.

Una vez definidos los nodos y el orden, se puede representar cualquier *spline* mediante una combinación lineal de funciones base. Las más utilizadas son las *B-splines*.

Los *B-splines* (*Basis splines*) son funciones base especialmente diseñadas para construir *splines* por tramos, de forma estable y eficiente. Son *splines* de grado d asociado a un conjunto de nodos.

Definición 2.2.4. ***Funciones base B-splines***

Una función $x(t)$ puede aproximarse mediante una base de B-splines:

$$x(t) \approx \sum_{m=1}^{M} c_m B_m(t)$$

donde:

- $B_m(t)$ *son funciones base B-spline de grado d, definidas sobre un conjunto de nodos.*
- c_m *son los coeficientes del desarrollo funcional.*

Una cuestión fundamental es determinar el número M de *splines*. Como criterio orientativo, puede emplearse la relación $K = m + L - 2$, que vincula el orden m del *spline* con el número total de nodos L. Aunque puede parecer que cuantos más *B-splines* se usen, mejor será el ajuste, esto no siempre es cierto. Si el número de funciones base aumenta, pero también cambia el reparto de los nodos, puede ocurrir que el nuevo sistema se ajuste peor a los datos. Por eso, es muy importante tener cuidado al seleccionar cuántos *B-splines* usar y cómo distribuir los nodos.

Ejemplo

A modo de ejemplo, se considera un *spline* cúbico (de orden 4), definido en el intervalo $[0, 1]$ con 5 nodos interiores (7 en total contando los nodos de los extremos) y 9 funciones *B-spline*. A continuación en la Figura 2.3 se muestra la forma de estas nueve bases:

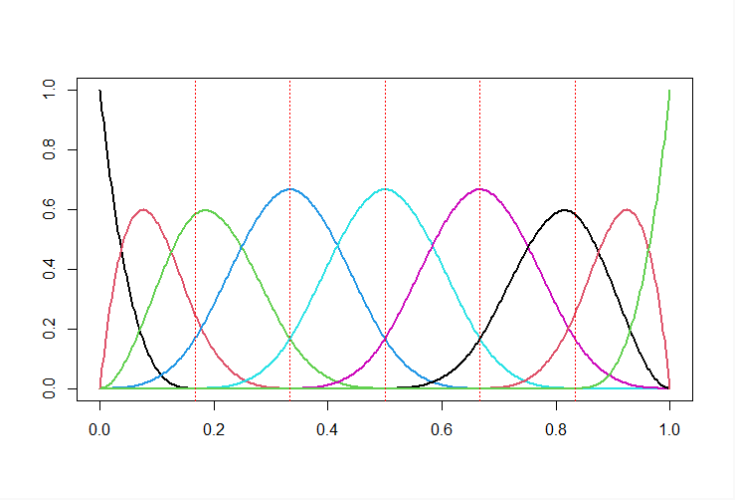


Figura 2.3. Sistema de nueve bases B-spline en el intervalo [0,1].

Las nueve bases *B-splines* cúbicas con cinco nodos interiores equiespaciados generan un espacio de dimensión nueve cuyos elementos (cada función base $B_i(x)$) valen cero fuera del intervalo $[0, 1]$. Las funciones base asociadas a los dos primeros y dos últimos nodos tienen soporte asimétrico frente al dominio (el tramo cúbico comienza o termina en los límites) mientras que las siete bases centrales tienen perfiles simétricos centrados en el punto medio de su soporte. Todas las bases son C^2 (dos veces continuamente diferenciable en todo el dominio) y satisfacen ($\sum_{i=1}^{9} B_i(x) = 1 \quad \forall\, x \in [0, 1]$), lo que garantiza un control preciso al ajustar las curvas (Picón Llamas, 2019).

Además de las variantes ya mencionadas, resulta relevante en ciertos contextos considerar los *splines* periódicos.

Los *splines* periódicos son una variante de los *splines* clásicos en los que, además de las condiciones de suavidad y definición por tramos en los nodos interiores, se imponen condiciones de periodicidad en los extremos del intervalo.

Definición 2.2.5. ***Splines periódicos***

Sean $a = t_0 < t_1 < \ldots < t_n = b$ los nodos que definen una partición del intervalo. Un spline periódico de grado d es una función $S : [a, b] \to \mathbb{R}$ que cumple:

$$S^{(k)}(a) = S^{(k)}(b), \quad \text{para } k = 0, 1, \ldots, d-1$$

Esta condición garantiza que la función spline resultante sea suave y continua a lo largo de todo

el dominio cerrado $[a, b]$.

En consecuencia, los *splines* periódicos son adecuados para modelizar datos funcionales de naturaleza cíclica. Desde el punto de vista de la construcción de bases, también es posible definir funciones base *B-spline* periódicas, adaptando la secuencia de nodos e imponiendo las condiciones de periodicidad. Una referencia clásica para la construcción y análisis de estos *splines* es (De Boor y De Boor, 1978).

Ejemplo

A continuación, la Figura 2.4 muestra un sistema de funciones base *B-spline* periódicas.

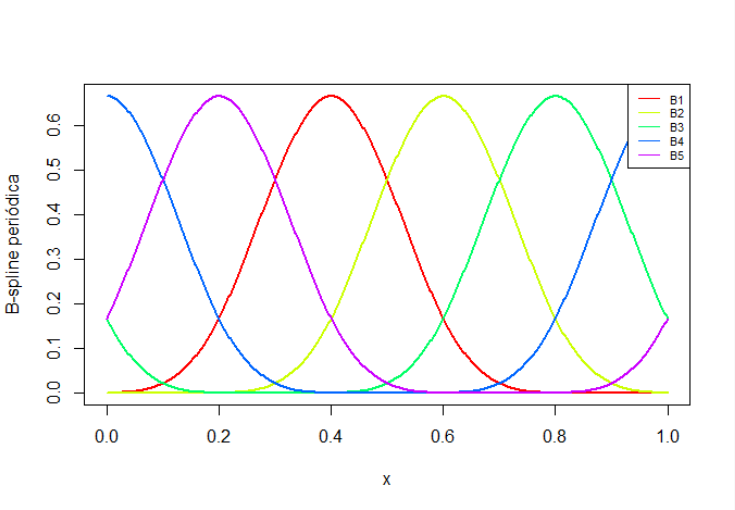


Figura 2.4. Funciones base B-spline periódicas de grado 3 en el intervalo [0,1].

2.2.3. Imputación

La ausencia de datos completos puede suponer un obstáculo para el correcto desarrollo del FDA. La información faltante, ya sean debidos a errores de medición, pérdidas de señal o simplemente a la imposibilidad de observar ciertas regiones del dominio, dificulta la aplicación directa de técnicas estadísticas pudiendo llegar a invalidar los resultados obtenidos. Para resolver este problema se recurre a la imputación, es decir, el proceso mediante el cual se estiman los valores faltantes a partir de la información observada, preservando la estructura y continuidad de las funciones.

Una estrategia común para abordar la imputación en este contexto es el suavizado funcional. Este enfoque parte del supuesto de que las funciones subyacentes que representan los fenómenos observados son regulares y continuas, lo que permite aproximarlas mediante combinaciones lineales de funciones base, como las bases de Fourier, *B-splines* o *splines* periódicos, definidas previamente. Al ajustar estas bases a los datos disponibles, el suavizado permite obtener una estimación global coherente de la función, incluyendo las regiones no observadas, guiándose por

la estructura general de la curva y la regularidad impuesta por las bases seleccionadas.

Desde este punto de vista, el suavizadado actúa simultáneamente como una herramienta de reconstrucción y de imputación, al proporcionar una estimación suave de la señal subyacente completa a pesar de que parte de ella no haya sido observada directamente. Esta estrategia es especialmente útil cuando el patrón de valores ausentes no es aleatorio y puede ser compensado mediante las propiedades estructurales del proceso funcional.

Este tipo de enfoque está muy extendido en la práctica (Ramsay y Silverman, 2002; Yao, Müller, y Wang, 2005).

2.3. Estadística descriptiva funcional

En esta sección se presentan los conceptos clave de la estadística descriptiva aplicados a datos funcionales, como son la media, varianza, covarianza y correlación, herramientas esenciales para analizar tendencias, variabilidad y relaciones a lo largo de un dominio continuo (Kokoszka y Reimherr, 2017).

Definición 2.3.1. ***Media funcional***

Dado un conjunto de n funciones suaves $x_1(t), x_2(t), \ldots, x_n(t)$, todas definidas sobre un dominio común T, la media funcional se define punto a punto como:

$$\overline{x}(t) = \frac{1}{n} \sum_{i=1}^{n} x_i(t), \quad \forall t \in T$$

La media funcional representa para cada instante t, el comportamiento promedio de las funciones individuales $x_i(t)$ en dicho instante.

A continuación la Figura 2.5 ilustra la evolución de la altura promedio de un grupo de niñas en función de la edad, a partir de un conjunto de observaciones reales (para más detalles sobre el dataset *growth*, ver la sección 1 del Anexo).

Ejemplo

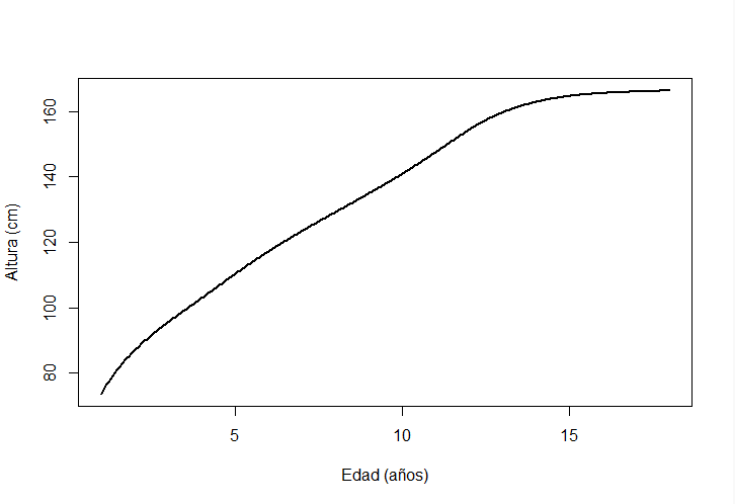


Figura 2.5. Representación gráfica de la altura media funcional de un grupo de niñas. Creado a partir del conjunto de datos growth de la librería fda en R.

A la vista del gráfico se puede decir que se trata de una curva suave y monótona creciente, lo que indica un crecimiento continuo y constante desde la infancia hasta la adolescencia. A su vez, podemos observar que el crecimiento es ligeramente más acelerado entre los 8 y 13 años, lo que probablemente refleja el estirón puberal típico en esa franja de edad. Después de los 13-14 años, la curva se estabiliza, alcanzando una altura cercana a la definitiva alrededor de los 16-17 años.

Definición 2.3.2. ***Varianza funcional***

La varianza funcional representa la dispersión de los valores funcionales respecto a la media, y se define como:

$$\text{var}(t) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i(t) - \overline{x}(t)\right)^2, \quad \forall t \in T$$

La Figura 2.6 representa gráficamente la varianza funcional del crecimiento, lo que permite analizar la dispersión de las curvas individuales respecto a la media funcional.

Ejemplo

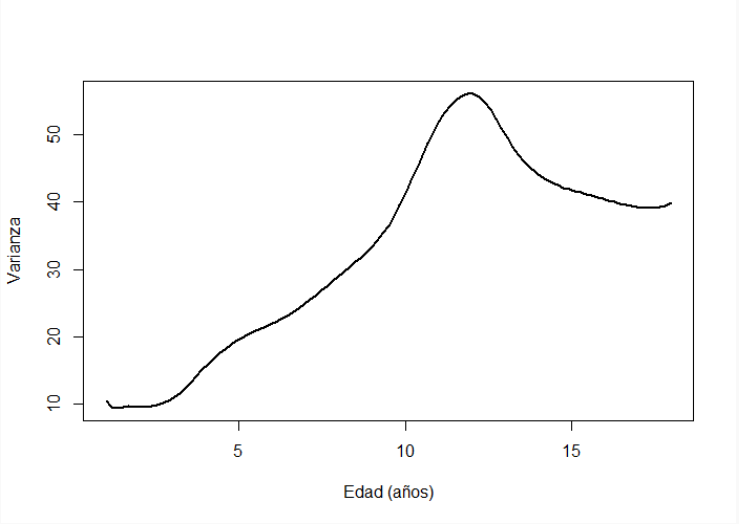


Figura 2.6. Gráfica de la varianza funcional del crecimiento. Creado a partir del conjunto de datos growth de la librería fda en R.

La variabilidad de la estatura en las niñas sigue un patrón definido a lo largo de la infancia y de la adolescencia. Durante los primeros años (de 2 a 5 años aproximadamente), la varianza es baja, lo que indica que las alturas son bastante homogéneas. A medida que las niñas se acercan a la pubertad, esta variabilidad crece, alcanzando su punto máximo a los 11-12 años. Este aumento refleja que el estirón en la pubertad no ocurre al mismo tiempo para todas: unas incian antes y otras después, es decir, existen diferencias más marcadas en la estatura. Tras ese pico, la varianza se reduce progresivamente, una vez completado el crecimiento acelerado, las alturas vuelven a ser más parecidas y el crecimiento se estabiliza.

Definición 2.3.3. ***Desviación típica funcional***

La desviación típica funcional se define como:

$$\mathrm{SD}(t) \;=\; \sqrt{\mathrm{var}(t)} \;=\; \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(x_i(t)-\bar{x}(t)\big)^2}, \quad \forall t \in T$$

Esta desviación típica funcional informa sobre la variabilidad de las curvas en los distintos instantes de tiempo, pero no nos proporciona información sobre cómo se relacionan los valores de las curvas entre dos instantes de tiempo cualesquiera, algo primordial en FDA.

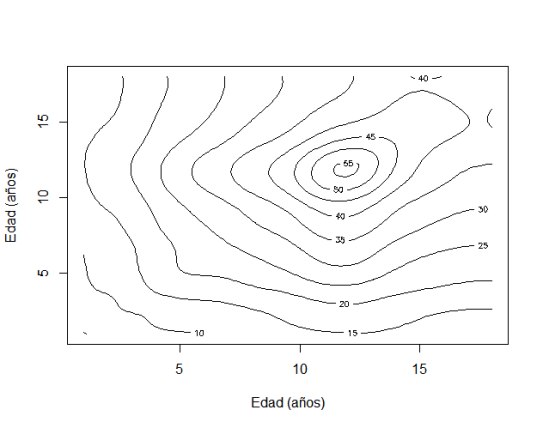
Definición 2.3.4. ***Covarianza funcional***

La covarianza funcional entre dos instantes $t, s \in T$ se define como la media de los productos de las desviaciones respecto a la media funcional en cada punto:

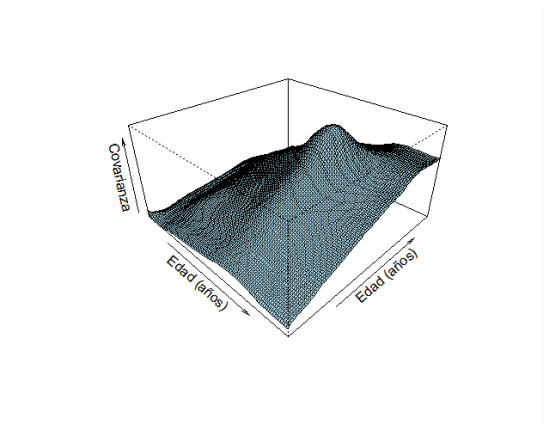
$$\mathrm{cov}(t,s) \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(x_i(t)-\bar{x}(t)\big)\,\big(x_i(s)-\bar{x}(s)\big), \quad \forall t,s \in T$$

A modo de ejemplo, evaluaremos la $\text{cov}(t, s)$ sobre las curvas de crecimiento utilizadas anteriormente para el grupo de niñas. En la Figura 2.7 se representa dicha función de dos formas complementarias:

Ejemplo



(a) Gráfico de contorno (2D) de la covarianza funcional en torno al valor máximo de la covarianza



(b) Representación 3D de la covarianza funcional

Figura 2.7. Representaciones de la covarianza funcional de la altura en niñas

La función de covarianza funcional muestra cómo varía conjuntamente la estatura de las niñas en distintos momentos de su desarrollo. En la superficie estimada, se observa un máximo claro de covarianza en torno a los 11–12 años, lo que indica que en esa etapa del crecimiento las alturas de las niñas presentan una mayor coherencia y variabilidad compartida.

Este máximo sugiere que las diferencias individuales en altura son especialmente marcadas y consistentes durante el llamado "estirón puberal", momento en el que el crecimiento se acelera. La alta covarianza en esta región implica que las niñas que presentan una mayor (o menor) estatura en torno a los 11–12 años tienden a mantener esa diferencia con respecto al resto durante todo ese intervalo.

Fuera de esa franja, las covarianzas disminuyen, lo que refleja que en etapas anteriores o posteriores las diferencias entre estaturas son más estables o menos asociadas entre edades distintas. Esto es coherente con la idea de que el momento exacto en el que se produce el estirón puede variar entre individuos, generando un patrón de variabilidad más localizado.

Para interpretar la intensidad de la asociación entre los valores del proceso en distintos ins-

tantes, independientemente de la escala de variación, es habitual considerar una versión estandarizada (entre -1 y 1) de la covarianza funcional denominada correlación funcional.

Definición 2.3.5. ***Correlación funcional***

La correlación funcional entre dos instantes de tiempos s y t se define como:

$$corr(t,s) = \frac{cov(t,s)}{\sqrt{var(t)}\sqrt{var(s)}}$$

2.4. Análisis de Componentes Principales Funcional

El análisis de componentes principales funcionales (FPCA, del inglés *Functional Principal Component Analysis*) es una extensión del análisis de componentes principales (PCA) al contexto de datos funcionales. Se trata de una herramienta estadística fundamental para estudiar la variabilidad estructural en procesos que pueden representarse como funciones continuas en el tiempo, el espacio o cualquier otro dominio funcional.

El objetivo principal del FPCA es descomponer la variabilidad total del conjunto de funciones observadas en una serie de componentes ortogonales ordenadas según la cantidad de variabilidad que explican. Esta representación permite reducir la dimensionalidad de los datos funcionales de forma eficiente, capturando los patrones dominantes de variación en pocos términos.

Una de las principales ventajas del FPCA es que proporciona una interpretación más manejable y visual de la estructura de covarianza del proceso funcional, cuya forma integral puede resultar difícil de analizar directamente. De forma análoga al PCA en el análisis multivariante clásico, el FPCA permite representar cada función observada como una combinación lineal de funciones propias (componentes funcionales principales), multiplicadas por sus respectivas puntuaciones (*scores*), facilitando así tanto el análisis exploratorio como la modelización o clasificación posterior. A continuación, se presenta una breve revisión del PCA clásico que aunque puede resultar conocido para el autor sirve como base para la notación y descripción del FPCA de la Subsección 2.4.2.

2.4.1. PCA: revisión del enfoque clásico

El PCA es una técnica estadística multivariante que permite reducir la dimensionalidad de un conjunto de datos transformando las variables originales en un nuevo conjunto de variables no correlacionadas, denominadas componentes principales. Estas componentes capturan la mayor parte de la variabilidad presente en los datos, y se ordenan de forma decreciente según la proporción de varianza explicada. En esta sección se produce un abuso de notación de x para mantener el paralelismo con la versión funcional del PCA (Hellton y Thoresen, 2017).

El objetivo del PCA es encontrar combinaciones lineales de las variables originales que capturen la mayor parte posible de la variabilidad de los datos. Estas combinaciones están definidas por vectores de pesos, conocidos como *loadings*, que determinan las direcciones de proyección. Para ello, dado un conjunto de datos con n observaciones de p variables con $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ denotando el vector de valores de la observación i-ésima, el procedimiento a seguir para llevar a cabo un PCA se resume como sigue:

1. Estandarización (si es necesario): si las variables están en escalas distintas, se estandarizan para que todas tengan media cero y varianza uno.

2. Cálculo de la matriz de covarianza (o de correlación):

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$$

3. Descomposición espectral de la matriz de covarianza: se obtienen los autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ y los autovectores correspondientes $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_p$ tal que:

$$\boldsymbol{\Sigma}\boldsymbol{\xi}_k = \lambda_k \boldsymbol{\xi}_k$$

4. Construcción de las componentes principales: la k-ésima componente principal es:

$$f_k = \boldsymbol{\xi}_k^\top \boldsymbol{x}$$

   donde $\boldsymbol{\xi}_k$ es el vector de pesos (*loading*) de la k-ésima componente, y $\boldsymbol{x}$ es el vector de variables originales.

5. Ordenación y selección de componentes: las componentes principales se ordenan según la varianza explicada:

   - La primera componente f_1 captura la mayor varianza posible.
   - La segunda f_2 es ortogonal a f_1 y captura la mayor varianza restante.
   - Así sucesivamente, hasta p componentes.

En la interpretación de PCA los *loadings* $\boldsymbol{\xi}_k$ indican la contribución de cada variable original a la componente k, es decir, son los vectores de dirección sobre los que se proyectan los datos. Mientras que los *scores* f_k se corresponden con las coordenadas de cada individuo en la nueva base generada por las componentes principales. De forma que el PCA proyecta los datos en las direcciones donde la varianza es máxima, lo que facilita la visualización, exploración e interpretación de estructuras complejas en los datos.

2.4.2. FPCA: Extensión del PCA al contexto funcional

La solución a este problema, en el marco funcional, ha sido ampliamente desarrollado por (Ramsay y Silverman, 2002). Desde este enfoque, el problema se formula como una búsqueda de funciones propias (autofunciones) y valores propios (autovalores) de un operador integral descrito a continuación, análogo a la búsqueda de autovectores y autovalores en el contexto multivariante clásico (Ferraty, 2006).

El FPCA extiende el PCA clásico al contexto en el que las observaciones no son vectores finitos, sino funciones continuas definidas sobre un dominio continuo. Esta metodología permite descomponer la variabilidad de un conjunto de curvas observadas en modos principales de variación, de forma análoga al caso multivariante, pero utilizando funciones como base en lugar de vectores.

Al igual que en el PCA clásico, el objetivo del FPCA es identificar direcciones principales de variación en los datos. En el contexto funcional, esto implica encontrar funciones de peso $\xi_m(t)$ que permitan representar las observaciones $x_i(t)$ mediante combinaciones ortogonales que capturen, en orden, la mayor parte de la variabilidad entre curvas. Para ello, dado un conjunto de funciones observadas $x_1(t), \ldots, x_n(t)$, definidas sobre un intervalo $T \subset \mathbb{R}$, el procedimiento del FPCA consta de los siguientes pasos:

1. Cálculo de la función media:
$$\bar{x}(t) = \frac{1}{n} \sum_{i=1}^{n} x_i(t)$$

2. Definición del operador (funcional) de covarianza:

   Se construye el operador integral de covarianza C a partir de la función de covarianza empírica $\text{cov}(t, s)$:
$$\text{cov}(t, s) = \frac{1}{n} \sum_{i=1}^{n} (x_i(t) - \bar{x}(t))(x_i(s) - \bar{x}(s))$$
   Este operador (funcional) actúa sobre una función $u(s)$ como:
$$C(u)(t) = \int_T \text{cov}(t, s) \cdot u(s)\, ds$$

3. Búsqueda de autofunciones (funciones de peso):

   Se desea encontrar una función $\xi_1(t)$ tal que los *scores* funcionales definidos por:
$$f_{i1} = \langle \xi_1, x_i \rangle = \int_T \xi_1(t) \cdot x_i(t)\, dt$$
   tengan varianza máxima entre las curvas, bajo la restricción de normalización:
$$\|\xi_1\|^2 = \int_T \xi_1^2(t)\, dt = 1$$

4. Obtención de componentes sucesivas:

   Para cada $m > 1$, se obtiene una nueva función $\xi_m(t)$ que maximiza la varianza de los nuevos *scores*:

   $$f_{im} = \langle \xi_m, x_i \rangle$$

   sujeto a:

   - Normalización: $\|\xi_m\|^2 = 1$
   - Ortogonalidad:

   $$\int_T \xi_m(t) \cdot \xi_k(t)\, dt = 0 \quad \text{para todo } k < m$$

Cada función de peso $\xi_m(t)$ representa un modo principal de variación en el conjunto de curvas observadas. Los autovalores asociados a cada ξ_m miden la importancia relativa de cada componente, y el conjunto de *scores* permite representar y analizar las funciones originales en un espacio reducido y más interpretable. Por otro lado, los valores f_{im} cuantifican en qué medida la función $x_i(t)$ refleja ese patrón de variación.

A continuación, para ilustrar de forma práctica todo lo expuesto hasta ahora, se va a aplicar el FPCA al conjunto de datos utilzado anteriormente (*growth* de la libreria *fda* en R). En este ejemplo se trabajará con las curvas de altura de las niñas, suavizadas mediante una base *B-splines*. Se utilizará la representación de las autofunciones para identificar el tipo de variación de cada componente.

Ejemplo

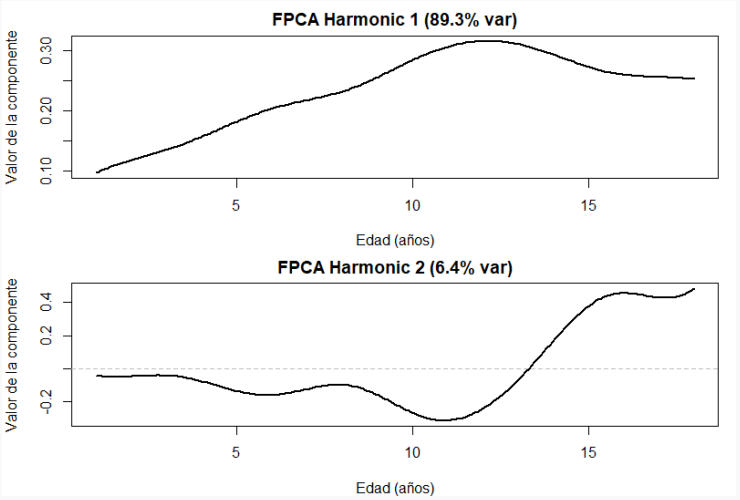


Figura 2.8. FPCA para las curvas de altura de las niñas. Creado a partir del conjunto de datos growth de la librería fda en R

La Figura 2.8 muestra los dos primeros modos de variación extraídos por FPCA sobre las curvas de altura de las niñas. Estas dos componentes nos permiten resumir más del 95 % de la variabilidad de las curvas de altura: FPC1 describe la intensidad global del crecimiento y FPC2 el desfase temporal respecto al patrón medio. Esto facilita tanto la visualización como la clasificación de los distintos perfiles de crecimiento en un espacio reducido e interpretable.

2.4.3. MFPCA: Análisis de Componentes Principales Funcional Multivariante

Tras haber explorado el FPCA, resulta natural extender este enfoque al caso multivariante. Mucha veces, las observaciones funcionales provienen de varias dimensiones interrelacionadas y al analizarlas por separado se podría perder información relevante. En este contexto, el análisis de componentes principales funcional multivariante (MFPCA por sus siglas en inglés *Multivariate Functional Principal Component Analysis*) constituye una herramienta idónea para capturar estructuras comunes y dependencias entre múltiples funciones, enriqueciendo así la comprensión del fenómeno de estudio.

Esta sección se ha elaborado siguiendo principalmente la metodología expuesta en (Berrendero, Justel, y Svarc, 2011).

El MFPCA extiende el FPCA al caso en el que cada unidad de observación está asociada con un vector de funciones, en lugar de una sola función. Esta metodología permite capturar la variabilidad conjunta y la estructura de correlación entre varias dimensiones funcionales registradas

simultáneamente.

Sea

$$\mathbf{x}_i(t) = \begin{pmatrix} x_{i1}(t) \\ x_{i2}(t) \\ \vdots \\ x_{iK}(t) \end{pmatrix}, \quad t \in T, \quad i = 1, \ldots, n,$$

donde K es el número de dimensiones funcionales. Cada observación consiste en un vector de K funciones reales.

El objetivo del MFPCA es encontrar funciones de peso multivariantes

$$\boldsymbol{\xi}_m(t) = \begin{pmatrix} \xi_{m1}(t) \\ \xi_{m2}(t) \\ \vdots \\ \xi_{mK}(t) \end{pmatrix}, \quad m = 1, 2, \ldots,$$

que permitan representar las observaciones mediante combinaciones ortogonales de los datos, a través de los *scores* funcionales multivariantes:

$$f_{im} = \int_T \boldsymbol{\xi}_m(t)^\top \mathbf{x}_i(t)\, dt = \sum_{k=1}^{K} \int_T \xi_{mk}(t) \cdot x_{ik}(t)\, dt$$

Para definir formalmente el problema, se introduce el operador (funcional) de covarianza multivariado C. Este operador actúa sobre una función $u(t)$ como:

$$C(u)(s) = \int_T C(s,t) \cdot u(t)\, dt$$

donde $C(s,t)$ es la matriz de funciones de covarianza cruzada:

$$C(s,t) = \operatorname{cov}(\mathbf{x}(s), \mathbf{x}(t)) \in \mathbb{R}^{K \times K},$$

y contiene en la posición (k, ℓ) la covarianza entre $x_{ik}(s)$ y $x_{i\ell}(t)$.

Las funciones de peso multivariadas $\boldsymbol{\xi}_m(t)$ se obtienen resolviendo el siguiente problema de valores propios:

$$\int_T \mathbf{C}(s,t)\, \boldsymbol{\xi}_m(t)\, dt = \lambda_m\, \boldsymbol{\xi}_m(s),$$

sujeto a la condición de normalización:

$$\|\boldsymbol{\xi}_m\|^2 = \int_T \boldsymbol{\xi}_m(t)^\top \boldsymbol{\xi}_m(t)\, dt = 1,$$

y a la ortogonalidad entre funciones de peso:

$$\int_T \boldsymbol{\xi}_m(t)^\top \boldsymbol{\xi}_\ell(t)\, dt = 0, \quad \text{para todo } \ell < m.$$

Cada función de peso $\boldsymbol{\xi}_m(t)$ representa un modo principal de variación conjunto entre las K dimensiones funcionales. Los *scores* f_{im} cuantifican la contribución de cada observación a dicho modo. El MFPCA permite, por tanto, identificar patrones coordinados de evolución funcional en contextos multivariantes, siendo especialmente útil en análisis de series temporales multicanal, biomarcadores longitudinales, o registros medioambientales multivariantes (Dugan, Barbuto, Masterson, y Shilt, 2024).

Capítulo 3

Análisis de datos y resultados

3.1. Descripción de los datos

En este TFG se analiza un conjunto de datos que combina dos fuentes de información complementarias. Por un lado, se dispone de datos clínicos de los pacientes atendidos en la Unidad de Psiquiatría del Hospital de León entre el 3 de noviembre de 2011 y el 27 de diciembre de 2022. Por otro lado, se incorporan registros ambientales y meteorológicos asociados tanto al día del ingreso como en los veinte días previos.

Los indicadores meteorológicos se obtuvieron directamente de la web de la Agencia Estatal de Mateorología (AEMET) (`https://opendata.aemet.es/centrodedescargas/inicio`), concretamente las medidas representativas de la provincia de León se recogieron de la estación *2661-León, Virgen del Camino*, mientas que los datos de contaminación proceden de la plataforma de datos abiertos de la Junta de Castilla y León (`http://servicios.jcyl.es/esco/index.action`).

Cabe destacar, que se desconocen los valores de cuatro indicadores meteorológicos entre el 2 de abril de 2014 y el 31 de julio de 2017 en la estación *2661-León, Virgen del Camino*, en concreto de la velocidad media, horas de sol y la presión atmosférica máxima y mínima. Según la AEMET, la causa de esta ausencia pudo ser debida a que dicha estación sufriera alguna incidencia en ese periodo de tiempo. Ante la ausencia de datos, la AEMET comunica que se pueden utilizar los datos procedentes de la estación *2661B-León, Aeropuerto*, muy próxima a la estación de la Virgen del Camino. Por lo tanto, entre abril de 2014 y julio de 2017 los datos de las variables velocidad media, horas de sol y la presión atmosférica máxima y mínima procederán de la estación *2661B-León, Aeropuerto*.

En la Tabla 3.1 se proporciona la descripción detallada de cada una de las variables que se encuentran en la base de datos.

Tabla 3.1

Definición de las variables

Variable	**Niveles**	**Descripción**
	Variables clínicas	
Edad	Cuantitativa discreta	Edad del paciente en años
Sexo	Mujer Hombre	Sexo del paciente
DX	Conducta autolítica Delirium Depresión Trastorno adaptativo Trastorno delirante Trastorno psicótico y del espectro de la esquizofrenia Trastorno relacionado con sucesos traumáticos y estresores Trastornos bipolares y relacionados Trastornos conductuales y del control de impulsos Trastornos de ansiedad Trastornos de la conducta alimentaria Trastornos de la personalidad Trastornos del neurodesarrollo Trastornos disociativos Trastornos obsesivos compulsivos Trastornos relacionados con el consumo de sustancias adictivas Trastornos somáticos Otros	Diagnostico psiquiatrico

Variable	**Niveles**	**Descripción**
Fecha Ingreso	Fecha (DD-MM-AAAA)	Día en el que el paciente fue ingresado en la Unidad de Psiquiatria del Hospital de León
Variables meteorológicas		
tMed	ºC	Temperatura media
prec	mm	Precipitación acumulada diaria. Un valor de *Ip* indica precipitación inferior a $0,1$mm (lluvia insignificante)
tMin	ºC	Temperatura mínima
tMax	ºC	Temperatura maxíma
velMedia	m/s	Velocodad media del vieno
Sol	Horas	Insolación
PresMax	hPa	Valor más alto de la presión barométrica registraso en el día
PresMin	hPa	Valor más bajo de la presión barométrica registraso en el día
Variables contaminación		
CO	mg/m^3	Monóxido de carbono
NO	$\mu g/m^3$	Óxido de nitrógeno
NO_2	$\mu g/m^3$	Dióxido de nitrógeno
O_3	$\mu g/m^3$	Ozono
PM_{10}	$\mu g/m^3$	Particulas en suspensión< $10\mu m$
$PM_{2,5}$	$\mu g/m^3$	Particulas en suspensión< $2,5\mu m$
SO	$\mu g/m^3$	Dióxido de azufre

Para el análisis de los datos y presentación de los resultados de este TFG se han manejado fundamentalmente tres librerías de R que han aportado rigor y solidez metodológica a este TFG. La libreeía *AER* (`https://cran.r-project.org/web/packages/AER/index.html`) para

abordar el problema de los datos censurados. La librería *fda* (`https://cran.r-project.org/web/packages/fda/index.html`) para llevar a cabo la descripción funcional de las series y la extracción de sus componentes principales funcionales. Por último, *MFPCA* (`https://cran.r-project.org/web/packages/MFPCA/index.html`) para enriquecer el análisis al considerar la estructura conjunta de múltiples variables. Estos resultados son reproducibles a partir del código que se adjunta en la sección 5 del Anexo.

3.2. Procesado e imputación de datos

Antes de aplicar las técnicas de análisis funcional, se llevó a cabo un análisis exploratorio de la base de datos con el fin de detectar posibles anomalías. No se identificaron valores atípicos ni patrones sistemáticos que pudieran comprometer la validez del análisis. Sin embargo, se observaron dos tipos de observaciones faltantes que podrían tener implicaciones analíticas: (i) valores ausentes propiamente dichos, en los que no se dispone de ningún registro para alguna de las variables en días puntuales y de forma aislada, y (ii) valores censurados, correspondientes a mediciones registradas por debajo de un umbral mínimo predefinido. A continuación, se describe el tratamiento específico adoptado para cada caso.

3.2.1. Tratamiento para valores ausentes

En esta sección, se describe el proceso de imputación aplicado a nuestras variables de interés, explicando los criterios de selección del método de imputación, así como la decisión de eliminar justificadamente alguna de las variables si fuera el caso.

Para valorar el alcance de las ausencias y qué método de imputación utilizar, se calcula tanto el porcentaje total de valores faltantes como el máximo de días consecutivos sin datos para cada variable. Estos resultados se muestran a continuación en Tabla 3.2.

Tabla 3.2

Resumen de ausencias

Variable	**% de NA's**	**Máx. días consecutivos con NA**
$tMed$	0,00073	2
$prec$	0,0348	5
$tMin$	0,00073	2
$tMax$	0,00073	2
$velMedia$	0,0029	2

Variable	% de NA's	Máx. días consecutivos con NA
sol	0,0022	4
$PresMax$	0,0015	2
$PresMin$	0,0017	2
CO	0,027	23
NO	0,0041	12
NO_2	0,0075	12
O_3	0,013	11
PM_{10}	0,0092	4
$PM_{2,5}$	0,47	1855
SO_2	0,01	12

En las variables funcionales con un porcentaje de datos ausentes inferior a 0,1 % se opta por imputar aplicando interpolación con *splines*. Este estrategia como se ha comentado en la Sección 2.2 está extendida en la literatura cuando hay un porcentaje bajo de valores ausentes y no están distribuidos de manera consecutiva. La implementación se realizó en R usando la función *smooth.spline()* a través de una función personalizada que recibe un vector númerico y un parámetro de suavizado, ver sección 5 del Anexo.

Por otro lado, la variable $PM_{2,5}$ presentó dos limitaciones que impidieron su imputación aplicando interpolación mediante *splines*:

1. Lagunas excesivamente largas. Faltan datos completos en seis anualidades, lo que origina huecos demasiado extensos para garantizar una interpolación continua y fiable.

2. Variabilidad no homogénea entre años (ver Figura 3.1).

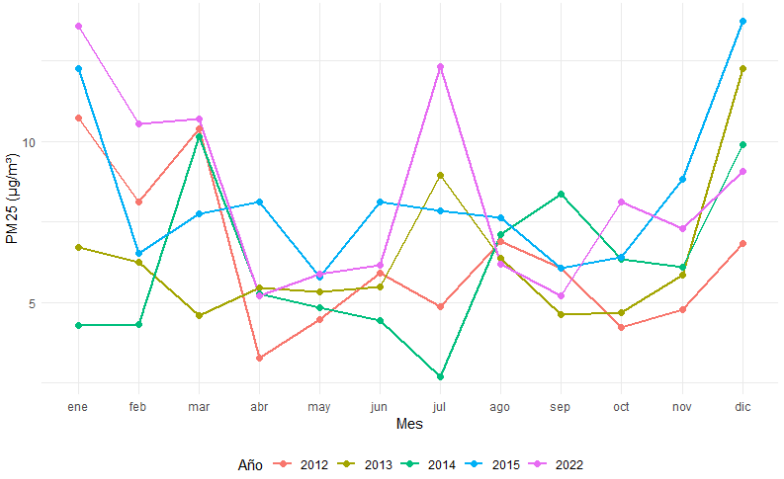


Figura 3.1. Medias mensuales de $PM_{2,5}$ para 2012, 2013, 2014, 2015 y 2022.

Al comparar las curvas completas de 2012, 2013, 2014, 2015 y 2022 en la Figura 3.1, se observa que el 2022 presenta un patrón estacional similar al resto (mínimos en primavera, máximos en invierno y verano), pero con picos significativamente más altos en enero, febrero y julio pudiendose interpretar que el promedio anual de $PM_{2,5}$ en 2022 fuera superior al de el resto de años. En consecuencia, ante la falta de información en los años intermedios y no resultado factible al evaluar la evolución o posible incremento de la variable $PM_{2,5}$ en los años intermedios, se optó por descartar esta variable del análisis, ya que cualquier intento de imputación mediante *spline* o media anual arriesgaría a introducir sesgos.

3.2.2. Tratamiento para datos censurados

Por otro lado, hay que tener en cuanta que la variable *prec* presenta datos censurados. Esta variable funcional, incluye además de valores *Ip* (valor inferior a 0,1mm) un número considerable de ceros, aproximadamente más del 60 % en cada uno de los días previos al ingreso. Esto seguramente se deba a que se haya utilizado algún tipo de corrección en el aparato de medida que haga que arroje de forma automática este valor, aunque no se haya llegado al mínimo.

Como conclusión, se tiene una variable funcional (*prec*) donde cada uno de los días puede tener hasta un 70 % de observaciones censuradas por la izquierda (los ceros mencionados anteriormente se tratarán como valores *Ip*), donde su valor estará entre $[0, Ip]$. Para abordar el problema de los datos censurados, se propone el uso del modelo Tobit (Hernández y Garcés, 2002), una extensión del modelo de regresión lineal diseñada específicamente para el análisis de variables censuradas. Este modelo permite estimar la relación entre una variable dependiente censurada y un conjunto de covariables, bajo el supuesto de que existe una variable latente continua no observable en toda su extensión, cuyo valor real se observa únicamente cuando supera (o está por debajo de) un determinado umbral de censura. La estimación de los coeficientes del modelo se realiza mediante máxima verosimilitud, considerando de forma explícita tanto las observaciones completamente observadas como aquellas censuradas. Esta estrategia resulta preferible frente a métodos más simples, como la imputación directa por el umbral de detección, que pueden inducir sesgos sistemáticos y provocar una pérdida sustancial de información al no modelizar adecuadamente el mecanismo de censura. En el caso concreto de la variable *prec*, se construye un modelo Tobit con variable respuesta la precipitación registrada en el día de la hospitalización y en los veinte días previos considerados. Mientras que como variables explicativas se consideraron la irradiancia solar (*sol*), la estación del año y la temperatura media diaria (*tMedia*). Tanto la implementación en R mediante la funcion *tobit* de la librería *AER* como los detalles metodológicos se incluyen en el Anexo en las secciones 2 y 5.

Debido a que el objetivo final consiste en llevar a cabo un FPCA sobre las curvas de precipitación, se opta por trabajar con el valor absoluto de las predicciones generadas por el modelo. En este contexto, interesa fundamentalmente capturar la variabilidad y la forma de las curvas de precipitación, más que la magnitud exacta de cada valor individual.

Una vez construido el modelo Tobit para cada fecha, se procede a la imputación de los valores de precipitación ausentes mediante *splines*, siguiendo la misma metodología aplicada anteriormente al resto de variables funcionales.

3.3. Suavizado de los datos mediante bases *B-splines*

Antes de proceder con el análisis descriptivo funcional y la posterior aplicación de técnicas de reducción de dimensionalidad (FPCA y MFPCA), se realiza un proceso de suavizado de los datos funcionales. Este paso es esencial para transformar los datos discretos observados (valores diarios recogidos durante los 20 días previos al ingreso hospitalario de cada paciente) en funciones suaves.

En este trabajo se ha optado por utilizar como método de suavizado las bases *B-splines*. Esta elección se debe a que estas bases permiten una representación flexible de funciones en dominios no periódicos, como es el caso de las variables analizadas, cuya evolución en los días previos al ingreso no siguen un patrón cíclico.

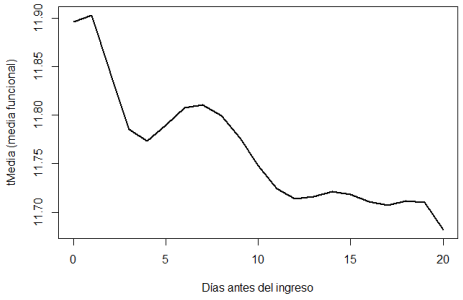
En todos los casos se ha utilizado una base *B-splines* con diez funciones base ($nbasis = 10$) definida en el intervalo $[0, 20]$, correspondiente al dominio temporal de interés (desde el día del ingreso hasta los 20 días anteriores). Se han decidido fijar diez funciones base por ser un valor que proporciona un equilibrio adecuado entre flexibilidad y suavidad en el ajuste funcional. El proceso de suavizado se ha implementado en R con la función *smooth.basis()* del paquete *fda*. El código empleado para la creación de los objetos funcionales y el ajuste por *B-splines* se incluye con detalle en la sección 5 del Anexo.

3.4. Análisis descriptivo funcional de las variables funcionales observadas

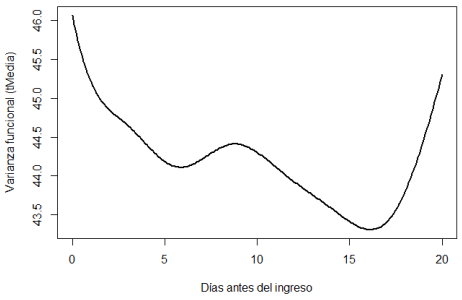
En esta sección, se presenta un análisis descriptivo para cada variable funcional. A continuación, se mostrarán los resultados únicamente para un subconjunto de las variables funcionales, lo que facilita la legibilidad y mantiene la coherencia con las variables de interés que serán analizadas en detalle en las secciones posteriores. Esto permitirá identificar tendencias temporales, zonas de mayor o menor dispersión y patrones de dependencia entre los días previos a la hospitalización.

Los resúmenes descriptivos para el resto de variables funcionales se encuentran recogidos en la sección 3 del Anexo.

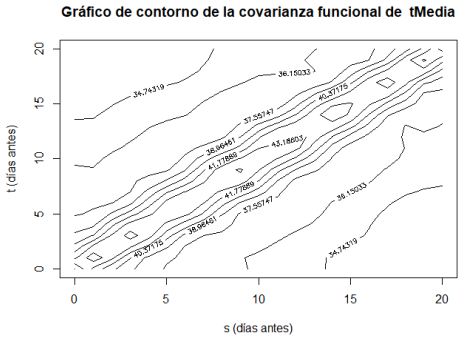
3.4.1. Análisis descriptivo para *tMedia*



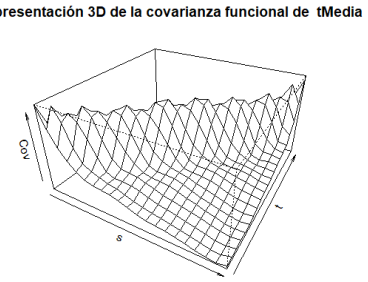
(a) Media funcional de tMedia



(b) Varianza funcional de tMedia



(c) Covarianza funcional de tMedia



(d) Covarianza funcional de tMedia

Figura 3.2. Representación de los estadísticos descriptivos funcionales para la variable tMedia

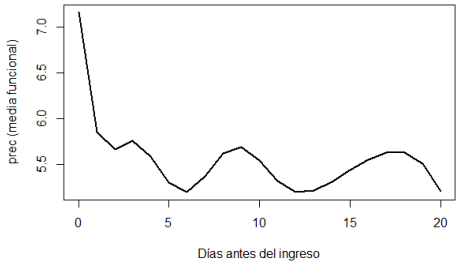
La media funcional de la temperatura media (Figura 3.2a) muestra una tendencia descendente en los 20 días previos al ingreso, especialmente marcada entre el día de la hospitalización y los 10 días antes.

En la Figura 3.2b se observa un aumento de la varianza cerca del ingreso, esto podría indicar que existe un grupo de pacientes expuestos a condiciones térmicas más extremas. Entre los días 5-15 antes del ingreso, la varianza disminuye, pero los valores de temperatura media muestran mucha variabilidad entre sujetos en los días más lejanos al ingreso, hecho que entre otros factores podría deberse a diferencias estacionales.

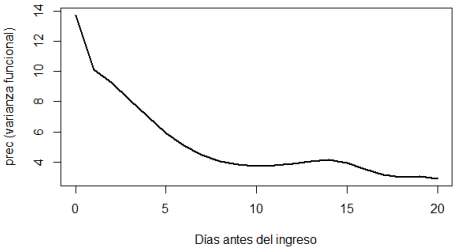
En la covarianza funcional representada en Figura 3.2c se aprecia, alta autocorrelación en la diagonal, como es habitual (temperaturas cercanas en el tiempo están fuertemente correlacionadas). Las líneas de nivel alargadas en diagonal indican dependencia temporal prolongada. Hay una banda de alta covarianza que se extiende desde el día 5 al 15, lo que sugiere que esos

días tienen patrones compartidos de evolución entre individuos. La representación tridimensional (Figura 3.2d) de la covarianza funcional muestra la misma información que el gráfico anterior pero en 3D, identificándose una zona de máxima covarianza en torno al centro del dominio temporal. A medida que aumenta la separación entre días, la covarianza decrece, como es esperable en procesos temporales suaves.

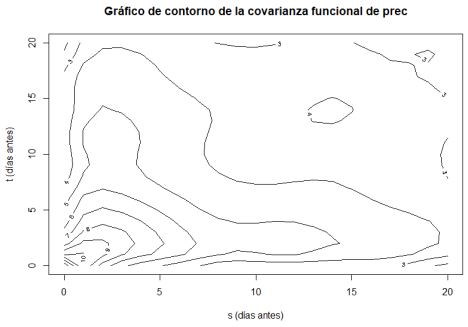
3.4.2. Análisis descriptivo para *prec*



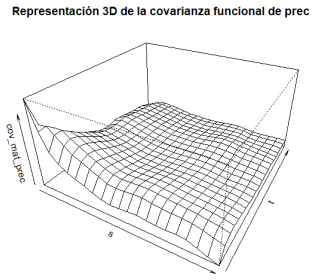
(a) Media funcional de prec



(b) Varianza funcional de prec



(c) Covarianza funcional de prec



(d) Covarianza funcional de prec

Figura 3.3. Representación de los estadísticos descriptivos funcionales para la variable prec

La media funcional de la precipitación (Figura 3.3a) muestra un descenso rápido desde el ingreso hacia los primeros días previos, seguido de dos fases de recuperación: un repunte a media distancia temporal y otro algo más lejano, antes de estabilizarse hacia el final del período.

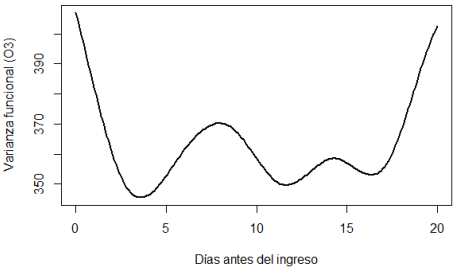
La representación obtenida para la varianza funcional de esta temperatura (Figura 3.3b) es acorde con lo observado para la media, con un patrón temporal decreciente, con tendencia más marcada y mayor variabilidad en los días previos al ingreso.

En cuanto a la correlación funcional, en la Figura 3.3c se observa un núcleo concentrado en la esquina inferior izquierda (0-5 días antes) que denota que esos días comparten patrones muy similares entre individuos. A medida que aumenta la separación entre dos días, la covarianza disminuye rápidamente. La representación tridimensional (Figura 3.3d) refuerza lo anterior: se

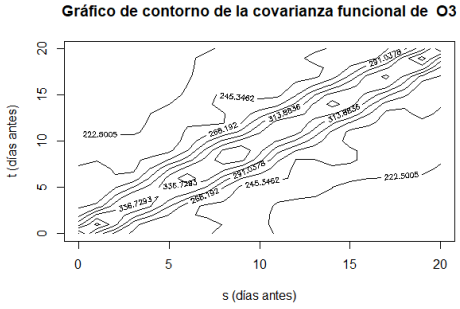
identifica la zona de máxima covarianza en la esquina inferior y se va atenuando a medida que los días estén más separados.

3.4.3. Análisis descriptivo para O_3

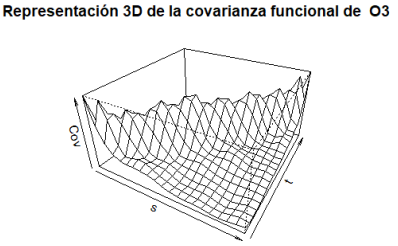
(a) Media funcional de O_3



(b) Varianza funcional de O_3



(c) Covarianza funcional de O_3



(d) Covarianza funcional de O_3

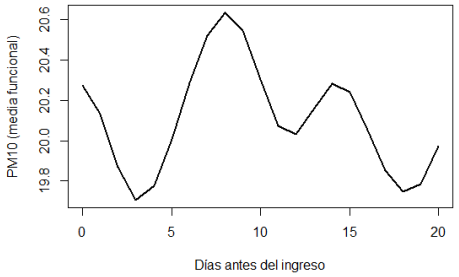
Figura 3.4. Representación de los estadísticos descriptivos funcionales para la variable O_3

La media funcional del nivel de ozono (Figura 3.4a) alcanza su máximo en torno a tres días antes del ingreso. La tendencia que se observa indica un aumento de los niveles de ozono desde 14 días antes del ingreso.

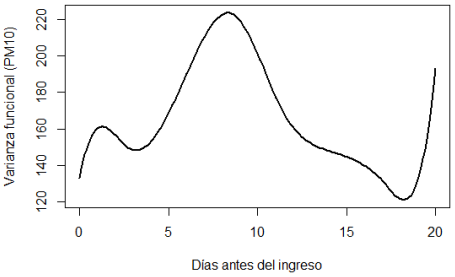
La varianza funcional (Figura 3.4b) no es homogénea a lo largo del tiempo. Se observa una variabilidad alta en los días intermedios al ingreso, con una dispersión mayor en los días previos y más lejanos al ingreso, siguiendo un patrón análogo al de la temperatura (ver Figura 3.2b).

Se observa una fuerte autocorrelación a lo largo de la diagnoal principal (ver Figura 3.4c), lo que indica que los niveles de ozono en días cercanos están altamente relacionados, de forma similar a lo observado en la temperatura (Figura 3.2c). La superficie tridimensional (Figura 3.4d) confirma este patrón, mostrando una banda de máxima covarianza en la diagonal $t = s$.

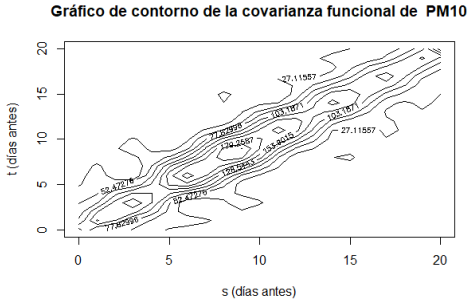
3.4.4. Análisis descriptivo para PM_{10}



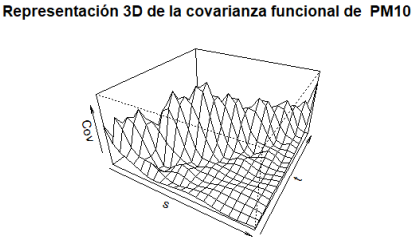
(a) Media funcional de PM_{10}



(b) Varianza funcional de PM_{10}



(c) Covarianza funcional de PM_{10}



(d) Covarianza funcional de PM_{10}

Figura 3.5. Representación de los estadísticos descriptivos funcionales para la variable PM_{10}

La media funcional de la variable PM_{10} (Figura 3.5a) presenta un aspecto bimodal: comienza con un ligero descenso, luego crece progresivamente hasta alcanzar un máximo en la región media del período, tras lo cual desciende de nuevo y experimenta un repunte intermedio antes de estabilizarse hacia el final. En conjunto, se trata de una tendencia con dos oscilaciones principales, destacando la más pronunciada alrededor de los días centrales.

La varianza funcional representada en la Figura 3.5b es menor en los extremos del dominio temporal y alcanza su nivel más alto en el tramo medio donde se producen incrementos y descensos.

En la Figura 3.5c se observa que conforme crece la separación entre dos días, la covarianza va disminuyendo gradualmente. La superficie tridimensional representada en la Figura 3.5d refuerza este esquema.

3.5. Implementación del FPCA y MFPCA

EL FPCA sintetiza la variabilidad de un conjunto de curvas en componentes que capturan sus patrones de variación más significativos. En lugar de limitarse a describir la tendencia y

dispersión globales, el FPCA identifica patrones específicos de variación a lo largo del dominio temporal. Además, al tratarse de una técnica de reducción de la dimensionalidad, resulta idonea para análisis posteriores, como *clustering* o regresión.

3.5.1. FPCA en variables climatológicas y de contaminación

En este apartado, se va a realizar un FPCA de manera independiente para cada una de las variables funcionales seleccionadas en la sección anterior (Sección 3.4), empleando los registros de los 4104 pacientes a lo largo de los 20 días previos al ingreso. Los resultados para el resto de variables pueden consultarse en la sección 4 del Anexo. La aplicación de FPCA de forma independiente a cada variable resulta muy didáctica porque facilita la ilustración de la metodología descrita en Subsección 2.4.2, aportando una interpretación clara a las autofunciones obtenidas y permitiendo generar, para cada individuo, unos *scores* que resumen la evolución temporal de cada variable. No obstante, conviene tener en cuenta que esta metodología no es capaz de describir las posibles correlaciones entre las distintas variables funcionales.

A lo largo de la sección seguiremos el mismo esquema, para cada variable se obtendrá un *plot de scores* de las dos primeras componentes principales ilustrando la estación en la que el individuo ingresa. Este gráfico permite visualizar cómo se distribuyen las observaciones respecto a los ejes de máxima variabilidad, representando la estructura global de la muestra, mostrando agrupamientos subyacentes entre individuos con curvas similares, así como, la relación entre las dos direcciones de mayor varianza y la presencia de valores atípicos. Por otro lado, también se van a representar las autofunciones hasta recoger un 90 % de la variabilidad total. Este gráfico es una herramienta clave para interpretar cada componente principal, mostrando de forma visual qué tipo de variación captura cada eje.

La implementación en R utilizada para llevar a cabo estos análisis se apoya en el paquete *fda* de R y queda recogida en la sección 5 del Anexo.

3.5.1.1. FPCA para *tMedia*

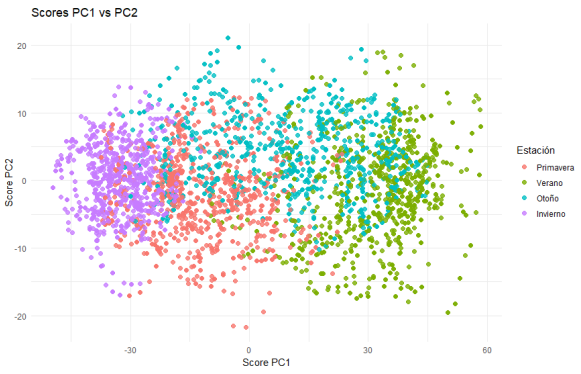


Figura 3.6. Scores de las dos primeras componentes principales de la variable tMedia en función de la estación en la que ingresa

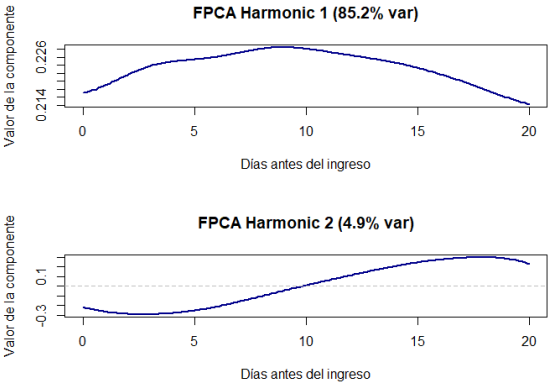


Figura 3.7. Dos primeras autofunciones de la variable tMedia

Las dos primeras componentes principales funcionales de la temperatura media (Figura 3.7) permiten caracterizar patrones diferenciados en los 20 días previos al ingreso psiquiátrico. La primera componente (PC1) recoge el 85,2 % de variabilidad total, presentando un máximo en torno a los días 10–12 antes del ingreso y mínimos tanto al inicio del periodo como en el día del ingreso. Este patrón sugiere una tendencia de aumento intermedio y descenso posterior de la temperatura, de modo que los sujetos con puntuaciones positivas (*scores*) en esta componente habrían experimentado temperaturas más altas hacia la mitad del periodo y más bajas en los extremos, mientras que puntuaciones negativas reflejarían un perfil más plano o invertido, sin esa elevación térmica intermedia. Este patrón podría reflejar cambios estacionales transitorios, como olas de frío o calor. Por su parte, la segunda componente (PC2) recoge un 4,9 % de la variabilidad explicada y muestra una forma similar a una "S" suave, con valores máximos hacia los días más alejados del ingreso y mínimos en los días cercanos al ingreso, describiendo así un descenso progresivo de la temperatura media a lo largo del periodo. Los sujetos con puntuaciones positivas en PC2 habrían estado expuestos a temperaturas más elevadas en los días lejanos al ingreso, seguidas de una bajada pronunciada hacia los días finales. En cambio, puntuaciones negativas indicarían un patrón opuesto, con temperaturas más bajas en los primeros días y más altas cerca del ingreso. Esta componente podría estar captando tendencias térmicas descendentes prolongadas, que podrían influir en descompensaciones previas al ingreso.

Por otro lado, en el *plot de scores* representado por estación (Figura 3.6) se aprecia cómo PC1, que resume el nivel medio y la tendencia global de la temperatura, discrimina claramente los perfiles estacionales de los pacientes. Los ingresos en verano (verde) quedan mayoritariamente

en la zona de *scores* altos en PC1, coherente con temperaturas más elevadas en los días previos, mientras que los de invierno (morado) se agrupan a la izquierda (*scores* bajos), reflejando exposiciones térmicas sustancialmente menores. Las estaciones intermedias (primavera y otoño) se ubican en posiciones centrales y muestran solapamiento, lo que sugiere que sus perfiles de temperatura media y tendencia son parecidas y que el FPCA no los separa tan nítidamente en este eje. Por su parte, PC2 apenas distingue por estación, confirmando que las diferencias estacionales en la temperatura media se deben principalmente a la tendencia global (PC1) y no tanto a la aceleración del cambio térmico capturada por la segunda componente.

3.5.1.2. FPCA para *prec*

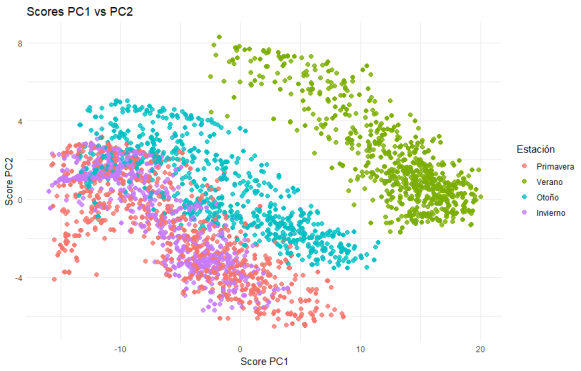


Figura 3.8. Scores de las dos primeras componentes principales de la variable prec en función de la estación en la que ingresa

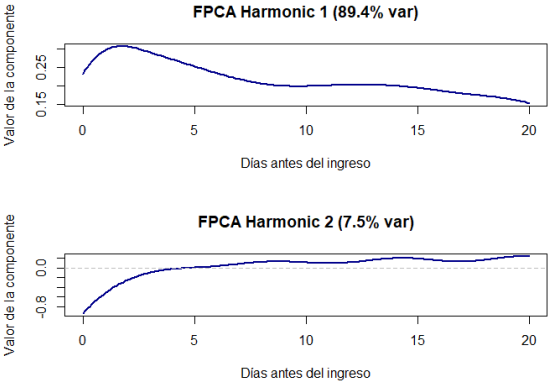


Figura 3.9. Dos primeras autofunciones de la variable prec

En el caso de la variable precipitación, el gráfico de *scores* de las dos primeras componentes principales funcionales (Figura 3.8) permite distinguir con claridad los grupos estacionales, observándose un patrón especialmente diferenciado para el verano. Además, se aprecia una tendencia decreciente conjunta entre los *scores* de PC1 y PC2, lo que sugiere una correlación negativa entre ambas componentes.

Concretamente, los individuos cuyo ingreso se produjo en verano (puntos verdes) se agrupan en la parte superior derecha del plano (valores positivos de PC1 y de PC2), lo que indica niveles más altos de precipitación total, típicamente asociados a episodios intensos y concentrados, como las tormentas estivales. El grupo de otoño (azul) se sitúa también en valores positivos de PC1, aunque algo más próximos al centro, lo que refleja una precipitación de intensidad media-alta distribuida de forma más regular. En contraste, las estaciones de invierno (morado) y primavera (rojo) se concentran en torno a valores más bajos o negativos de PC1, lo que sugiere niveles de

precipitación más moderados en esos periodos del año.

Cabe destacar que la primera componente principal funcional (PC1) explica casi el 90 % de la variabilidad total, por lo que concentra la mayor parte de la información relativa al volumen global de precipitación a lo largo del periodo. Por su parte, la segunda componente (PC2), que explica un 7,5 % de la varianza (Figura 3.9), presenta una evolución temporal característica: comienza en valores negativos en el entorno del ingreso, asciende rápidamente y se estabiliza en torno a valores positivos a partir del día 5. Esta forma sugiere que PC2 puede interpretarse como un gradiente temporal de precipitación. En particular, *scores* negativos en PC2 indicarían que la mayor parte de la lluvia se concentró en los días más cercanos al ingreso (lluvia temprana). En cambio, *scores* positivos reflejarían un patrón en el que la precipitación se desplazó hacia los días más lejanos al ingreso (lluvia tardía o más distribuida).

Este análisis sugiere que el perfil funcional de precipitación en las semanas previas al ingreso varía sistemáticamente según la estación del año, no solo en cuanto a la cantidad total, sino también en cuanto al momento de acumulación, lo que podría tener implicaciones relevantes para el estudio del contexto ambiental previo al ingreso psiquiátrico.

Adicionalmente para la variable *prec*, se procedió a comprobar que los individuos con datos imputados no formen un subgrupo separado, sino que se integren de manera proporcional con los no imputados.

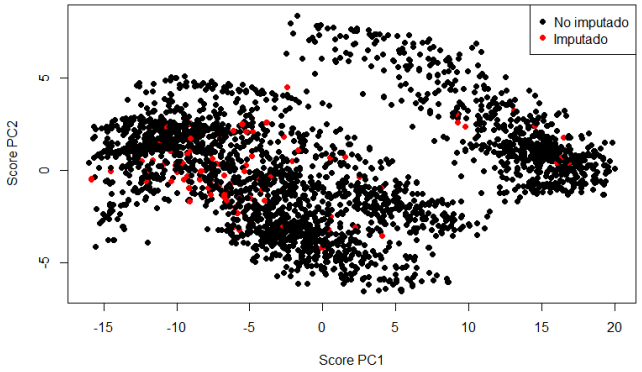


Figura 3.10. Scores PC1 vs PC2 de la variable prec separando por individuos imputados y no imputados

Como se muestra en la Figura 3.10, los individuos representados en rojos (imputados) están distribuidos homogéneamente dentro de la nube de puntos negros (no imputados). Esto confirma que la imputación se ha realizado correctamente y no se ha generado un sesgo estructural en los patrones principales de variabilidad de la precipitación.

3.5.1.3. FPCA para O_3

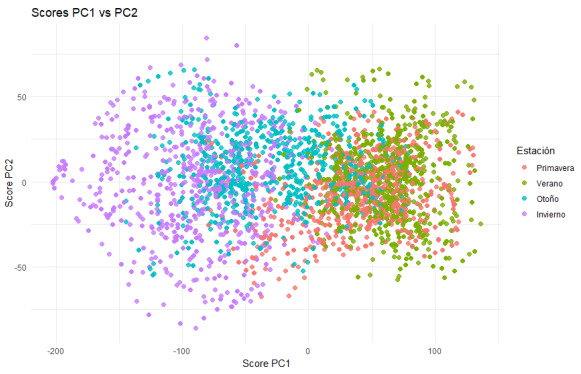


Figura 3.11. Scores de las dos primeras componentes principales de la variable O_3 en función de la esatción en la que ingresa

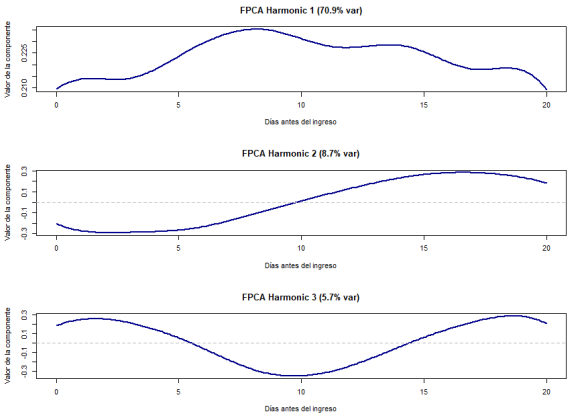


Figura 3.12. Tres primeras autofunciones de la variable O_3

En el caso de la variable ozono (O_3), la Figura 3.11 presenta los *scores* correspondientes a las dos primeras componentes principales. Es posible distinguir las estaciones de primavera y verano hacia la parte derecha del gráfico, mientras que invierno se agrupa hacia la izquierda, y el otoño ocupa una posición intermedia. No obstante, en este caso, la diferenciación entre estaciones no resulta tan clara y no se aprecian tendencias definidas.

Las tres primeras componentes principales exlican aproximadamente el 85 % de la varianza total. Dado que la cuarta autofunción no aporta información relevante, se optó por extraer únicamente las tres primeras componentes.

La primera componente principal funcional (PC1) del ozono explica el 70,9 % de la variabilidad y podría actuar como índice de nivel medio de ozono: su autofunción (representada en Figura 3.12) adopta una meseta suave entre los 5 y 15 días antes del ingreso, de modo que un *score* positivo eleva uniformemente la curva en esa región y uno negativo la atenúa.

La segunda componente (PC2, 8,7 %) presenta una forma que cruza el cero en el medio del periodo, pudiendo actuar como un gradiente temporal de la exposición del ozono: un *score* negativo en PC2 podría enfatizar los picos de ozono en los diez días inmediatamente anteriores al ingreso, mientras que un *score* positivo podría interpretarse como que atenúa esa acumulación tardía y realza la exposición más alejada en el tiempo (10-20 días antes). Así, más allá de cuantificar el nivel medio con PC1, la segunda componente permitiría distinguir a los pacientes según cuándo se concentró su exposición al ozono en el periodo prehospitalario.

Finalmente, la tercera componente (PC3, 5,7 %) presenta oscilaciones locales con un valle central y picos en los extremos, pudiendo interpretarse como episodios de cambio abrupto en

momentos puntuales.

3.5.1.4. FPCA para PM_{10}

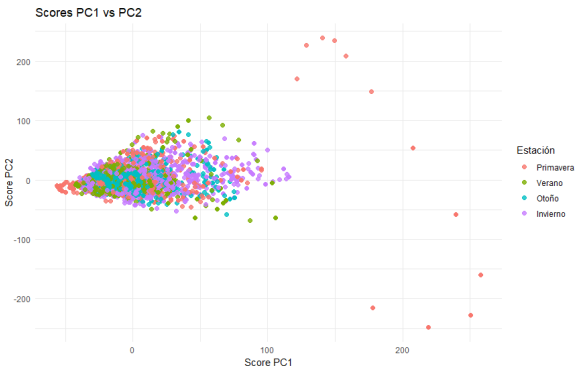


Figura 3.13. Scores de las dos primeras componentes principales de la variable PM_{10} en función de la estación en la que ingresan

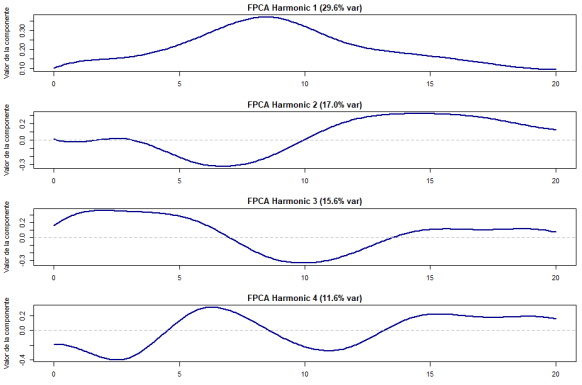


Figura 3.14. Cuatro primeras autofunciones de la variable PM_{10}

En la Figura 3.13 se muestran los *scores* correspondientes a las dos primeras componentes principales funcionales de la variable PM_{10}. En este caso, no se aprecia una diferenciación clara entre estaciones, lo que sugiere una distribución más homogénea a lo largo del año. No obstante, se identifican algunos individuos atípicos, que podrían estar distorsionando la visualización global del gráfico. Su posible impacto e interpretación serán abordados más adelante.

Para el FPCA de la variable PM_{10} se ha decidido representar las primeras cuatro componentes principales a pesar de no llegar a recoger el 90 % de variabilidad total porque en la quinta no se han observado patrones relevantes.

Las cuatro componentes representadas en la Figura 3.14 explican cerca del 74 % de la variabilidad funcional total. La PC1 de la variable PM_{10} recoge cerca del 30 % y podría resumir la intensidad media o global de exposición al contaminante. La Figura 3.13 muestra como a lo largo de la PC1, los *scores* se extienden horizontalmente, por lo que podría ser razonable pensar que: (i) Individuos con *score* alto en PC1 tienen curvas con más exposición a PM_{10} en los días previos. (ii) *Scores* bajos en PC1 podrían corresponder a baja o estable exposición; la PC2 recoge un 17 % y podría corresponder a una tendencia lineal que baja hasta el sexto día anterior al ingreso y luego sube hasta el final, distinguiendo acumulaciones más intensas en la segunda mitad del periodo; la PC3 recoge un 15,6 % y refleja picos tempranos y un valle profundo en torno al día 10, pudiendo capturar episodios de caída brusca tras un inicio elevado; y finalmente, la PC4 reocoge alrededor del 12 % y describe oscilaciones de alta frecuencia, señalando fluctuaciones puntuales que las otras tres componentes principales no recogen.

Como se ha mencionado, en el *plot de scores* de las dos primeras componentes funcionales de la variable PM_{10} (Figura 3.13), se observan valores atípicos en la parte derecha de dicho gráfico, correspondientes todos a valores de individuos cuyo ingreso ocurre en primavera.

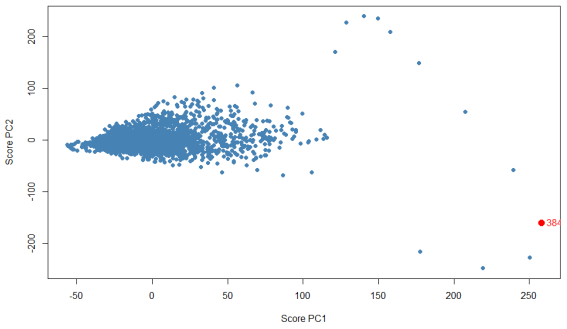


Figura 3.15. Outlier PM_{10}

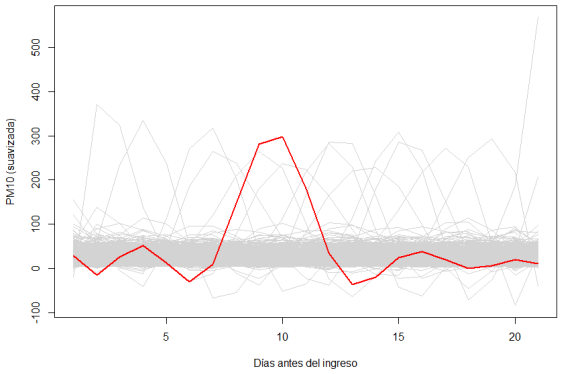


Figura 3.16. Curvas suavizadas de PM_{10}: outlier vs resto

Entre estos *outliers*, se selecciona uno, representado en la Figura 3.16 junto con el resto de indiviuos. Como se puede observar (Figura 3.15), este sujeto presenta $PC1 > 250$ (su exposición media diaria supera en más de $250\mu g/m^3$ la curva promedio) y $PC2 \approx -150$ (refleja un pico de más de $300\mu g/m^3$ alrededor de los 10 días previos al ingreso y un valle de $-80\mu g/m^3$ 12 días antes).

Al observar el gráfico de los *scores* del FPCA aplicado a la variable PM_{10} (Figura 3.13), surge la sospecha de que los *outliers* visibles en la parte derecha del eje correspondiente a la primera componente (PC1>125), podrían corresponder a ingresos hospitalarios concentrados en un mismo periodo. Tras comprobarlo, se confirma que todos estos individuos ingresaron durante la primavera de 2022. Este hallazgo plantea la posibilidad de que se trate de un error de medición o, alternativamente, de un comportamiento ambiental anómalo no detectado previamente mediante análisis exploratorios. En cualquier caso, se opta por representar en la Figura 3.17 nuevamente el gráfico de *scores* excluyendo estos individuos, lo que ofrece una visualización más clara y estructurada. No obstante, a diferencia de lo observado en otras variables como la precipitación, en el caso de PM_{10} no se aprecia una influencia marcada de la estación del año en los patrones de variabilidad, lo que sugiere una dinámica más estable o menos estacional de este contaminante.

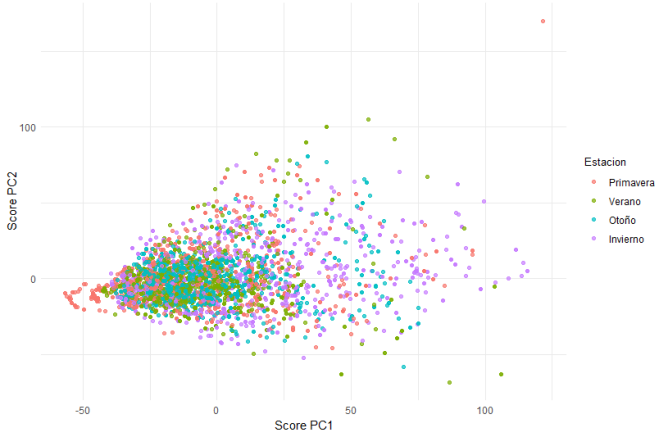


Figura 3.17. Scores PC1 vs PC2 de la variable PM_{10} sin los outliers

3.5.2. MFPCA aplicado a las variables seleccionadas

Con el fin de obtener una descripción más específica de los patrones de variación que afectan a nuestro conjunto de variables, se propone seleccionar para el MFPCA la temperatura media ($tMedia$), la precipitación ($prec$), el ozono (O_3) y las partículas en suspensión PM_{10}. Se ha tomado esta decisión porque, en el FPCA preliminar, estas cuatro series mostraron patrones interesantes en los *plot de scores* y autofunciones; además, representan fenónemos centrales en la dinámica climática y la calidad del aire.

Al aplicar el MFPCA a este conjunto de variables, nuestro objetivo es capturar mediante las componentes principales el patrón conjunto de variación de las cuatro curvas ambientales en los 20 días previos al ingreso. La forma que adopta cada curva en el gráfico, nos indicará cómo se manifiesta esa dimensión de variación en cada variable a lo largo del tiempo.

Dado que las variables consideradas presentan diferentes unidades, se ha considerado necesario escalar dichas funciones antres de aplicar MFPCA.

A continuación, en la Figura 3.18 se muestra el primer modo de variación conjunta.

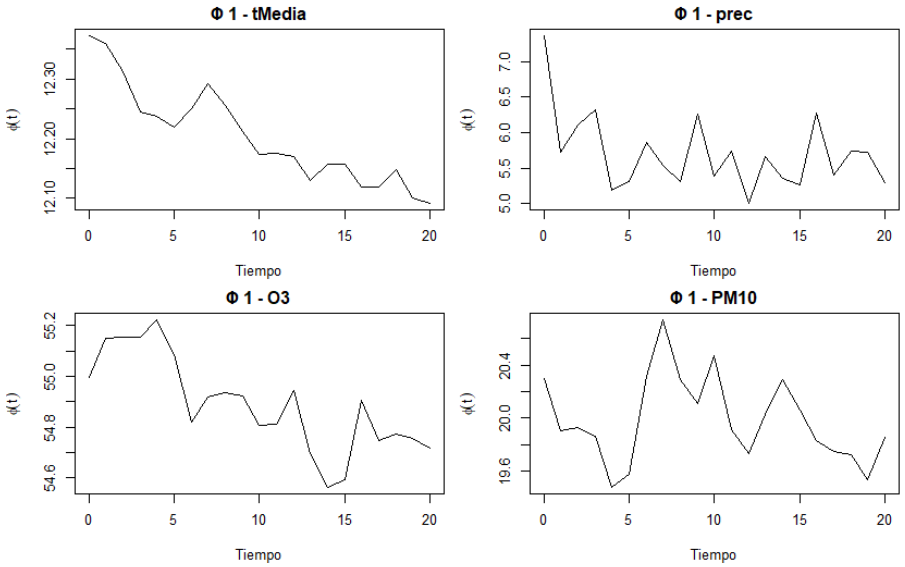


Figura 3.18. Primera componente principal funcional

Esta primera componente recoge el 67,76 % de la variabilidad total. Al comparar los valores medios de la primera mitad del dominio (desde el día del ingreso hasta los 10 días antes) con la segunda mitad (10-20), se observa un patrón conjunto donde la temperatura media, precipitación, ozono y partículas de PM_{10} presentan su máxima intensidad justo en la primera mitad del dominio y disminuyen de manera progresiva al retroceder en el tiempo, conservando además un repunte de ozono muy cercano al ingreso y un máximo de PM_{10} alrededor de los 8 días previos; así, una puntuación positiva en PC1 reproduce este perfil de carga y una negativa invierte su dirección.

A continuación, en la Figura 3.19 se representa el segundo modo de variación conjunta.

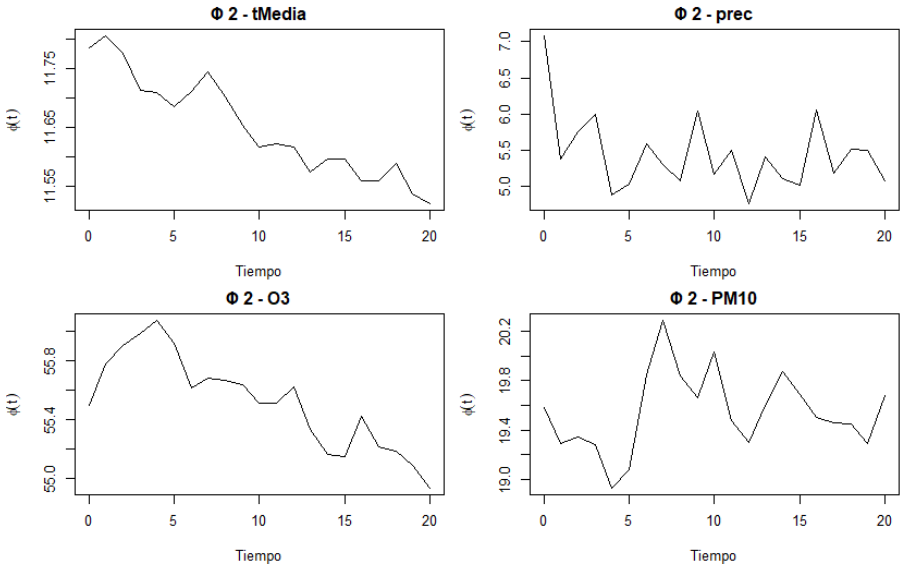


Figura 3.19. Segunda componente principal funcional

Esta segunda componente recoge el 13,53 % de la variabilidad total y muestra un patrón bastante similar a PC1 en las cuatro variables (temperaturas y precipitaciones que alcanzan su

punto más alto el día del ingreso y van decreciendo a medida que nos alejamos en el tiempo, junto a los mismos repuntes de ozono y de PM_{10}), aunque con un leve desplazamiento a niveles más bajos en la temperatura media. Tal similaridad entre PC1 y PC2, podría sugerir que ambas capturan esencialmente el mismo patrón de variación conjunta, con una ligera diferencia en el nivel de temperatura: PC1 agrupa a los individuos con temperaturas globalmente más altas y PC2 a los de temperaturas más bajas, manteniendo en ambos el mismo perfil de valores máximos en el ingreso y tendencia decreciente hacia atrás. Por ello, resulta necesario la representación de PC3 (Figura 3.20) para poder observar otro modo de variación que no se limite a esta distinción por nivel térmico.

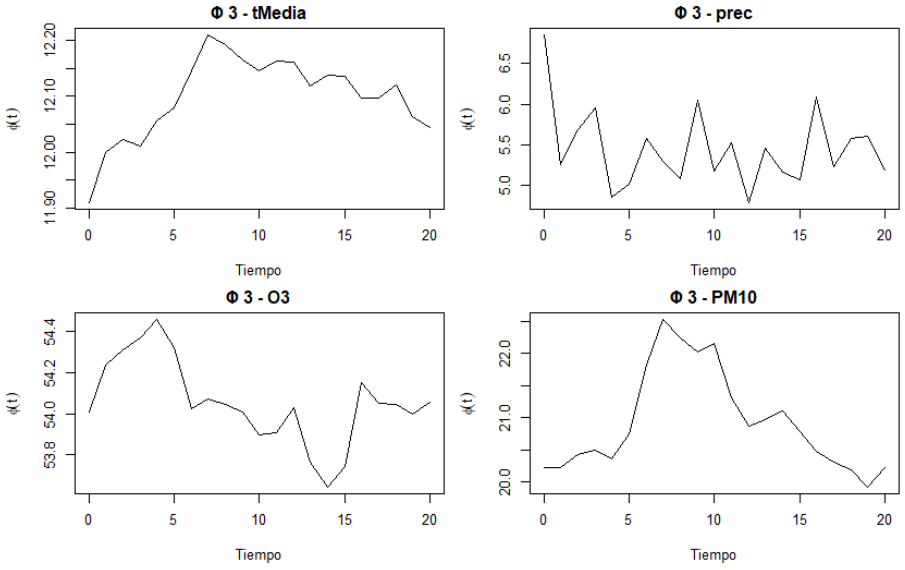


Figura 3.20. Tercera componente principal funcional

Esta tercera componente, que recoge aproximadamente el 8 % de la variabilidad total, pone de manifiesto un modo de variación conjunto disinto al perfil que se observaba en las dos primeras componentes: alrededor de los 6-10 días antes del ingreso, todas las curvas exhiben un incremento conjunto. En concreto, la temperatura media asciende hasta un máximo cercano a 7-8 días antes, para luego estabilizarse en niveles algo más bajos; simultáneamente, las partículas de PM_{10} presentan un máximo aún más pronunciado en ese mismo intervalo, seguido de un descenso progresivo hacia los extremos del dominio temporal. El ozono también alcanza su punto más alto en los 4-6 días previos, desciende a continuación y ofrece un repunte leve al final, mientras que la precipitación, pese a su mayor irregularidad, tiende a exhibir mínimos inmediatos al ingreso, un repunte en el tramo medio y luego oscilaciones menores.

Este patrón podría indicar que existe un tercer modo de variación ambiental que podría concentrar un incremento simultáneo de temperatura, contaminantes y lluvia en torno a la segunda semana antes del ingreso, aportando una dimensión temporal que las componentes 1 y 2 no habían capturado.

A continuación, en la Figura 3.21 se presenta un *plot de scores* en el espacio PC1-PC3. Se ha decidido representar en estas dos componentes porque PC2 resultaba redundante con PC1 al compartir prácticamente el mismo patrón de variación, mientras que PC3 aporta un eje nuevo al capturar el refuerzo intermedio de todas las variables en la segunda semana previa al ingreso. Tras comprobar que ni el sexo ni el diagnóstico generan ningún agrupamiento ni influyen en la distribución de los *scores*, seguimos el mismo criterio aplicado en el FPCA y representamos los *scores* según las estaciones del año.

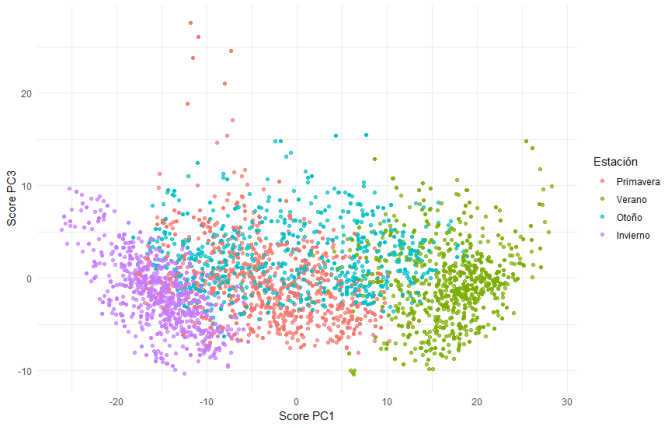


Figura 3.21. Scores PC1 vs PC3 por estación

Los individuos ubicados en la derecha del gráfico podrían reproducir el perfil de valores elevados en esas cuatro variables en el día del ingreso y descenso hacia atrás, mientras que los situados a la izquierda invierten esa dinámica, mostrando intensidades crecientes al retroceder en el tiempo. Por otra parte, quienes aparecen en la parte superior del gráfico posiblemente presenten el incremento simultáneo de todas las variables en la segunda semana previa al ingreso; en cambio, los puntos bajos carecen de ese pulso intermedio. Además, se aprecia un patrón estacional: PC1 tiende a separar los ingresos de invierno (puntuaciones negativas, hacia la izquierda) de los de verano (puntuaciones positivas, hacia la derecha), mientras que otoño y primavera ocupan posiciones intermedias.

Capítulo 4

Conclusiones. Trabajo futuro.

Este trabajo ha partido de una comprensión profunda de los datos funcionales de temperatura media, precipitación y contaminación registrados en los 20 días previos al ingreso psiquiátrico urgente. Este tipo de datos exige un tratamiento cuidadoso (suavizado, elección de bases, tratamiento de valores ausentes y datos censurados, etc) y plantea desafíos de alta dimensionalidad que motivan el uso de métodos de reducción de dimensión (FPCA y MFPCA) que faciliten la interpretación y los análisis posteriores, y que en esta memoria se ha abordado desde el punto de vista teórico, práctico y computacional.

En este contexto, el estudio llevado a cabo en este TFG consiste en una aproximación y exploración inicial al problema, tanto desde el punto de vista metodológico como clínico, en la que se han identificado variables y patrones de variación que podrían sintetizar la dinámica ambiental en los días previos al ingreso, pudiendo servir este trabajo como base para análisis posteriores más profundos.

Se seleccionaron un conjunto de datos meteorológicos (temperatura media y precipitación) y contaminantes atmosféricos (O_3 y PM_{10}) medidos durante los 20 días previos al ingreso hospitalario.

De este análisis se extraen varias conclusiones:

1. A partir del FPCA realizado sobre la temperatura, precipitación, ozono y PM_{10} se identificaron perfiles funcionales que difieren en la magnitud media de los valores registrados, en el momento temporal en el que se producen los máximos picos de exposición, así como en la presencia de fluctuaciones irregulares a lo largo del periodo analizado. Estas diferencias podrían estar moduladas por la estación del año en la que el paciente ingresa, lo que sugiere que las condiciones ambientales podrían influir en la evolución clínica previa al ingreso y convendría incorporarlas en futuros modelos predictivos que capturen mejor la dinámica temporal de estos factores.

2. Al aplicar el MFPCA sobre las cuatro series ambientales, se identificaron tres componentes principales funcionales que describen, sucesivamente, perfiles de intensidad alta en el ingreso (PC1), la existencia de dos subgrupos diferenciados por temperaturas globales altas o bajas (PC2), así como una intensidad centrada (PC3), con el mayor nivel de exposición alrededor de los 10 días previos al ingreso. En conjunto, estos ejes explican casi el 90 % de la variabilidad y generan *scores* para cada individuo facilitando, por ejemplo, la estratificación de pacientes utilizando herramientas como el *clustering* funcional.

3. Al contrastar los *scores* con variables demográficas y clínicas (sexo y diagnóstico psiquiátrico principal) en nuestro análisis exploratorio no se observó ningún agrupamiento claro en función de estas variables, pero podría haber otras terceras variables no recogidas en el estudio que estuvieran influyendo.

A partir de estos resultados preliminares, se proponen dos líneas principales para el desarrollo futuro del estudio:

- **Estratificación estacional del análisis:**

  La agrupación natural observada en los *scores* funcionales por estación del año sugiere que el efecto de las variables ambientales podría estar modulado por la estación. En consecuencia, el análisis combinado de datos procedentes de diferentes estaciones podría estar introduciendo cierta confusión en la interpretación de patrones latentes. Una vía prometedora consiste en repetir el análisis de componentes principales funcionales de forma estratificada por estación, lo que permitiría identificar patrones específicos estacionales, mejorar la robustez de los resultados, y detectar posibles mecanismos diferenciales de la temporalidad estacional.

- **Transición hacia modelos predictivos:**

  Dado que el presente trabajo ha permitido esbozar unas primeras representaciones funcionales robustas de los entornos ambientales previos al ingreso, el siguiente paso natural es avanzar hacia modelos con enfoque predictivo. En este sentido, se plantean dos estrategias complementarias:

  1. La aplicación de *clustering* funcional para identificar subgrupos clínicamente relevantes en función de los perfiles ambientales.
  2. La implementación de modelos de regresión funcional o de aprendizaje automático (por ejemplo, redes funcionales o modelos aditivos funcionales) que incorporen los *scores* de las componentes principales como predictores, con el objetivo de predecir

riesgo de ingreso, categoría diagnóstica u otros eventos clínicos como el números de ingresos por año.

Estas líneas permitirán consolidar el valor del análisis funcional como herramienta en el estudio del impacto de factores ambientales sobre la salud mental, facilitando el desarrollo de modelos explicativos más precisos y, eventualmente, estrategias de intervención preventiva informadas por factores ambientales.

Referencias

Berrendero, J. R., Justel, A., y Svarc, M. (2011). Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, *55*(9), 2619–2634.

Bowman, A. W., y Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with s-plus illustrations* (Vol. 18). OUP Oxford.

Chiou, J.-M., y Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *69*(4), 679–699.

De Boor, C., y De Boor, C. (1978). *A practical guide to splines* (Vol. 27). springer New York.

Dugan, E. L., Barbuto, A. E., Masterson, C. M., y Shilt, J. (2024). Multivariate functional principal component analysis and k-means clustering to identify kinematic foot types during gait in children with cerebral palsy. *Gait & Posture*, *113*, 40–45.

Ferraty, F. (2006). *Nonparametric functional data analysis*. Springer.

Gómez González, L. (2022). Impacto a corto plazo de factores ambientales sobre los ingresos por trastorno mental en la comunidad de madrid.

Greven, S., Crainiceanu, C., Caffo, B., y Reich, D. (2011). Longitudinal functional principal component analysis. En *Recent advances in functional data analysis and related topics* (pp. 149–154).

Hellton, K. H., y Thoresen, M. (2017). When and why are principal component scores a good tool for visualizing high-dimensional data? *Scandinavian Journal of Statistics*, *44*(3), 581–597.

Hernández, M. B., y Garcés, A. T. (2002). Aplicación de los modelos de regresión tobit en la modelización de variables epidemiológicas censuradas. *Gaceta Sanitaria*, *16*(2), 188–195.

Horváth, L., y Kokoszka, P. (2012). *Inference for functional data with applications* (Vol. 200). Springer Science & Business Media.

Hu, T., Xu, Z.-Y., Wang, J., Su, Y., y Guo, B.-B. (2023). Meteorological factors, ambient air pollution, and daily hospital admissions for depressive disorder in harbin: A time-series study. *World Journal of Psychiatry*, *13*(12), 10–61.

Kokoszka, P., y Reimherr, M. (2017). *Introduction to functional data analysis*. Chapman and Hall/CRC.

Levitin, D. J., Nuzzo, R. L., Vines, B. W., y Ramsay, J. (2007). Introduction to functional data analysis. *Canadian Psychology/Psychologie canadienne*, *48*(3), 135.

Manso, P. P. (2019). Análisis de datos funcionales: representación en bases y regresión funcional en scikit-fda.

Ordóñez-Iriarte, J. M. (2020). Salud mental y salud ambiental. una visión prospectiva. informe

sespas 2020. *Gaceta sanitaria*, *34*, 68–75.

Picón Llamas, D. M. (2019). *Análisis de datos funcionales aplicado a datos de temperatura en españa* (Trabajo fin de grado). Universidad de Valladolid.

Ramsay, J. O., y Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*. Springer.

Rizavas, I., Gournellis, R., Pantazis, N., Chatzinikolaou, F., Douzenis, P., Efstathiou, V., ... Douzenis, A. (2023). The impact of meteorological factors on involuntary admission in attica, greece. *Psychiatriki*, *34*(4), 289–300.

Simonoff, J. S. (2012). *Smoothing methods in statistics*. Springer Science & Business Media.

Son, J., y Shin, J. (2021). Bimodal effects of sunlight on major depressive disorder. *Comprehensive psychiatry*, *108*, 152232.

Wang, D., Li, X., Tian, S., He, L., Xu, Y., y Wang, X. (2021). Quantifying the dynamics between environmental information disclosure and firms' financial performance using functional data analysis. *Sustainable Production and Consumption*, *28*, 192–205.

Yao, F., Müller, H.-G., y Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, *100*(470), 577–590.

Anexo

1. Conjunto de datos *growth*: descripción para ejemplificar la metodología del TFG

El conjunto de datos *growth* forma parte del paquete *fda* en R y recopila las series de estatura de niños y niñas del estudio de crecimiento de Berkeley (*Berkeley Growth Study*). Se compone de:

- *growth$age*: vector de 31 puntos temporales que van desde 1 hasta 18 años.
- *growth$hgtm*: matriz de dimensión 31 x 39 en la que cada columna contiene las 31 medidas de altura de cada uno de los 39 niños, tomadas en las edades indicadas.
- *growth$hgtf*: matriz de dimensión 31 x 54 con las 31 medidas de altura de cada una de las 54 niñas, en los mismos instantes de medición.

Se decide seleccionar este conjunto de datos para ilustrar la metodología empleada porque reúne curvas de crecimiento con un dominio temporal común (31 edades de 1 a 18 años) y un tamaño manejable (39 niños y 54 niñas). Estas características facilitan el ajuste de bases funcionales y la extracción de componentes principales.

2. Metodología modelo Tobit

Los modelos Tobit permiten tratar adecuadamente la censura por la izquierda, estimando la media real sin necesidad de imputar datos por valores fijos (como el umbral o la mitad del umbral), ni descartando datos.

Este modelo asume que existe una variabe latente continua, es decir, una variable que no se puede observar completamente, que sigue una distribución normal, y estima los parámetros mediante máxima verosimilitud considerando que los residuos del modelo son independientes y normalmente distribuidos con media cero y varianza constante. Por ejemplo:

- Existe una precipitación real (y^*) que sigue un modelo lineal con error normal:

$$y^* = \beta_0 + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- Pero solo observamos:

$$y = \begin{cases} y^*, & \text{si } y^* > c \quad (\text{con } c \text{ umbral de detección}), \\ 0{,}1, & \text{si } y^* \leq c \end{cases}$$

La estimación se realiza maximizando una función que se construye a partir de la contribución de cada observación i, diferenciando entre las observaciones censuradas ($y_i = 0$) y las no censuradas ($y_i > 0$). Primero, se define la variable indicadora:

$$D_i = \begin{cases} 0, & \text{si } y_i = 0 \\ 1, & \text{si } y_i > 0 \end{cases}$$

Sabemos que existe una variable latente continua y_i^* tal que $y_i^* = x_i'\beta + \varepsilon \quad \text{con} \quad \varepsilon \sim N(0, \sigma^2)$, y que la variable observada y_i viene dada por:

$$y_i = \begin{cases} y_i^*, & \text{si } y_i^* > c \text{ (con umbral } c) \\ 0, & \text{si } y_i^* \leq c \end{cases}$$

Para simplificar la notación, fijamos $c = 0$. De este modo, $y_i = 0$ equivale a $y_i^* \leq 0$ y $y_i > 0$ equivale a $y_i^* > 0$. Entonces, la densidad de la observación i, condicionada en x_i, se escribe de la siguiente manera:

$$f(y_i \mid x_i) = \begin{cases} \Pr\big(y_i = 0 \mid x_i\big) = \Phi\Big(-\frac{x_i'\beta}{\sigma}\Big) & \text{si } D_i = 0, \\ f\big(y_i \mid x_i,\, y_i > 0\big) = \dfrac{1}{\sigma}\,\varphi\Big(\frac{y_i - x_i'\beta}{\sigma}\Big), & \text{si } D_i = 1 \end{cases}$$

donde:

- $\Phi(\cdot)$ es la función de distribución acumulada de la normal estándar
- $\varphi(\cdot)$ es la densidad de la normal estándar
- $\Pr(y_i = 0 \mid x_i) = \Pr\big(y_i^* \leq 0 \mid x_i\big) = \Phi\big(-x_i'\beta/\sigma\big)$

En forma conjunta, podemos escribirlo como:

$$f(y_i \mid x_i) = \Big[\, \Phi\big(-\tfrac{x_i'\beta}{\sigma}\big)\Big]^{1-D_i} \times \Big[\, \tfrac{1}{\sigma}\,\varphi\big(\tfrac{y_i - x_i'\beta}{\sigma}\big)\Big]^{D_i}$$

Por lo tanto, la función de verosimilitud para la muestra completa de n observaciones es:

$$L(\beta, \sigma) = \prod_{i=1}^{n} \Big[\, \Phi\big(-\tfrac{x_i'\beta}{\sigma}\big)\Big]^{1-D_i} \times \Big[\, \tfrac{1}{\sigma}\,\varphi\big(\tfrac{y_i - x_i'\beta}{\sigma}\big)\Big]^{D_i}$$

Este enfoque permite estimar tanto el vector de parámetros como la varianza, evita el sesgo y utiliza toda la información disponible.

3. Estadísticos descriptivos funcionales

Análisis descriptivo para *tMin*

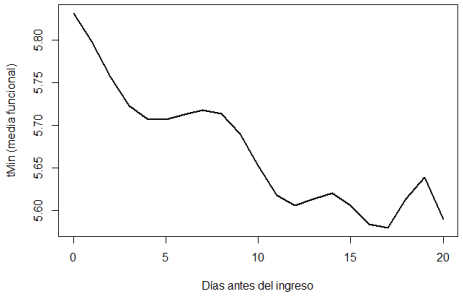


Figura 4.1. Media funcional de tMin

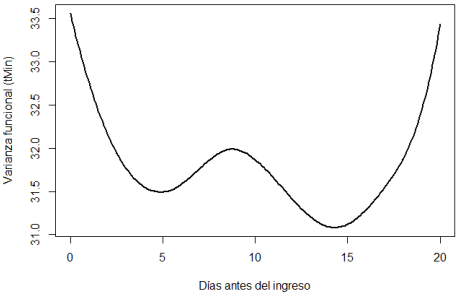


Figura 4.2. Varianza funcional de tMin

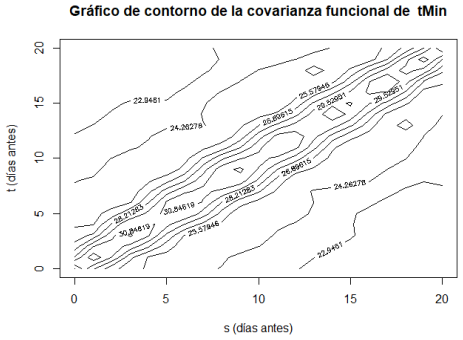


Figura 4.3. Covarianza funcional de tMin

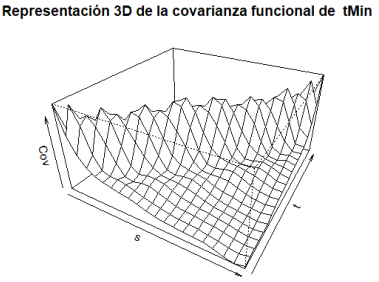


Figura 4.4. Covarianza funcional de tMin

Análisis descriptivo para *tMax*

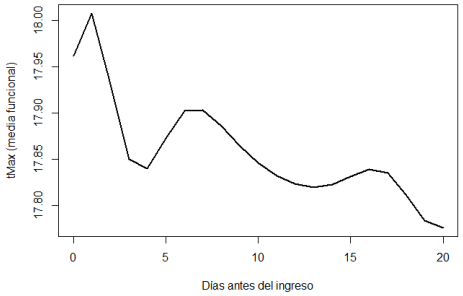


Figura 4.5. Media funcional de tMax

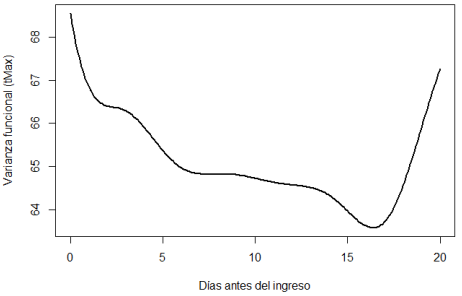


Figura 4.6. Varianza funcional de tMax

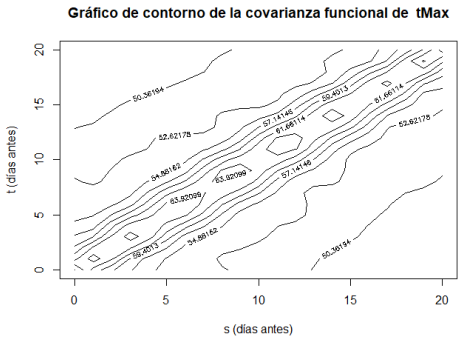


Figura 4.7. Covarianza funcional de tMax

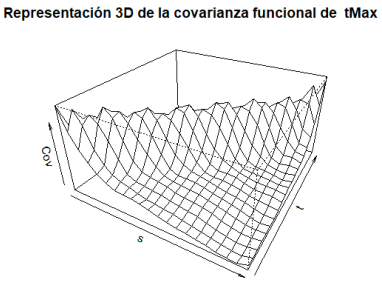


Figura 4.8. Covarianza funcional de tMax

Análisis descriptivo para *velMedia*

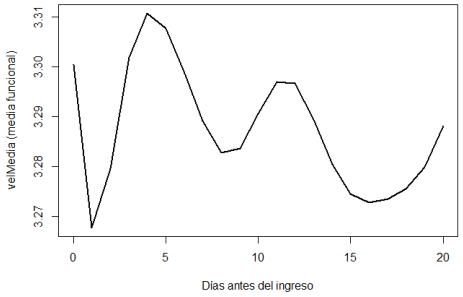


Figura 4.9. Media funcional de velMedia

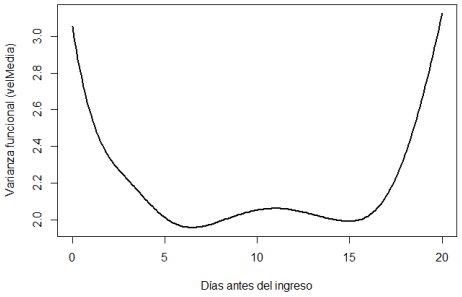


Figura 4.10. Varianza funcional de velMedia

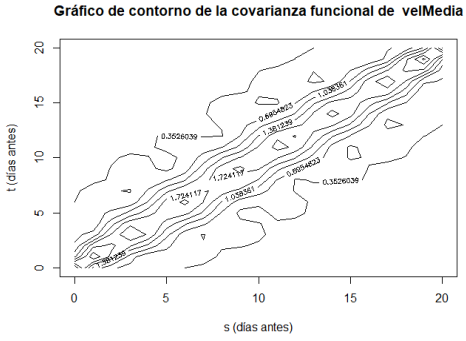


Figura 4.11. Covarianza funcional de velMedia

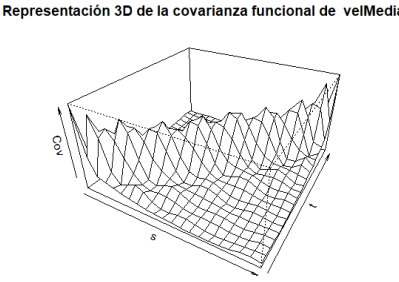


Figura 4.12. Covarianza funcional de velMedia

Análisis descriptivo para *sol*

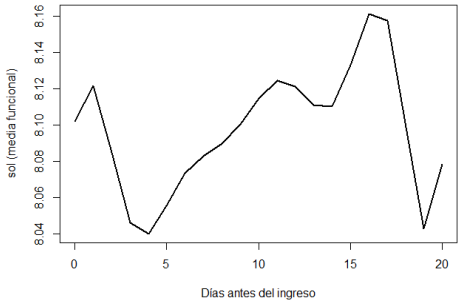


Figura 4.13. Media funcional de sol

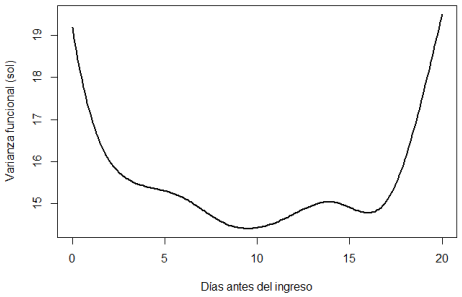


Figura 4.14. Varianza funcional de sol

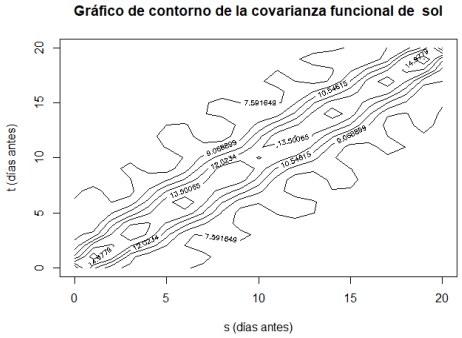


Figura 4.15. Covarianza funcional de sol

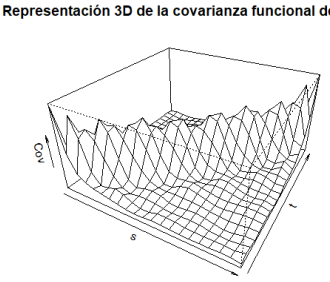


Figura 4.16. Covarianza funcional de sol

Análisis descriptivo para *PresMax*

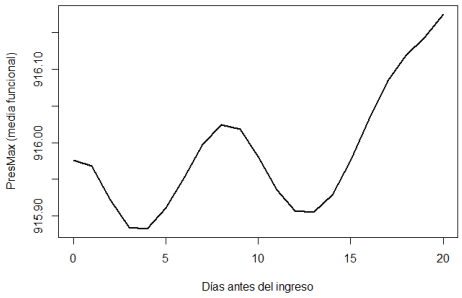


Figura 4.17. Media funcional de PresMax

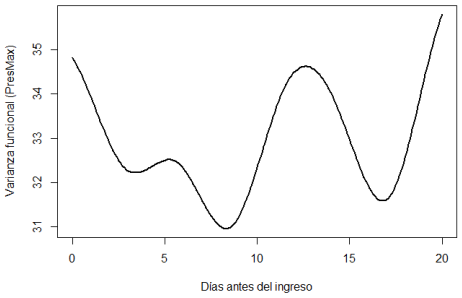


Figura 4.18. Varianza funcional de PresMax

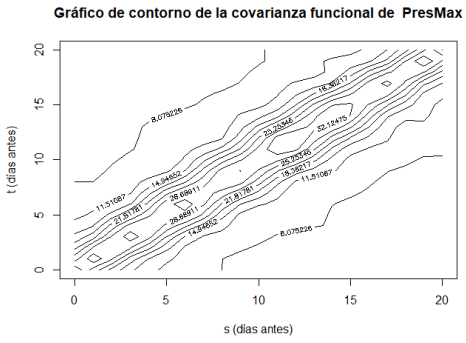


Figura 4.19. Covarianza funcional de PresMax

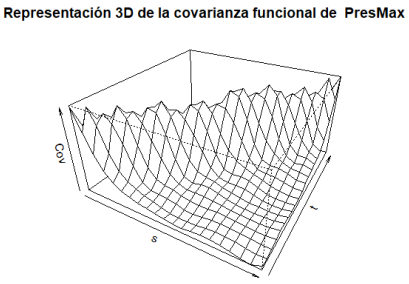


Figura 4.20. Covarianza funcional de PresMax

Análisis descriptivo para *PresMin*

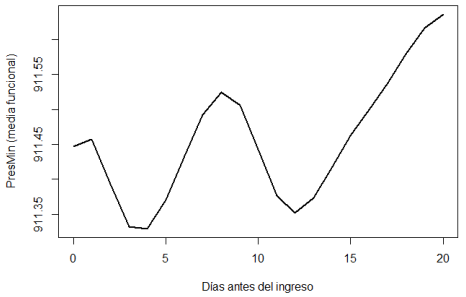


Figura 4.21. Media funcional de PresMin

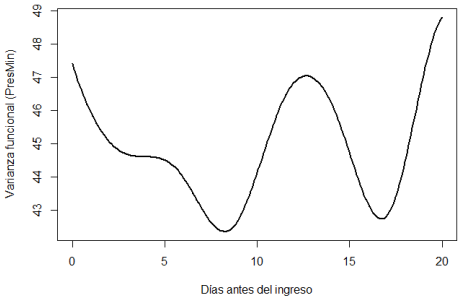


Figura 4.22. Varianza funcional de PresMin

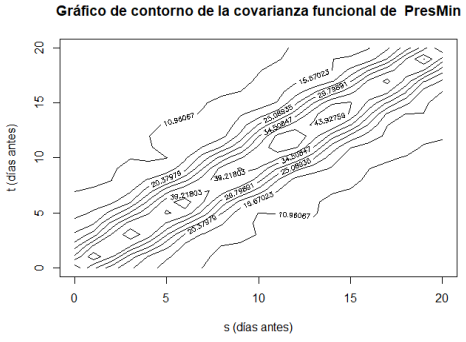


Figura 4.23. Covarianza funcional de PresMin

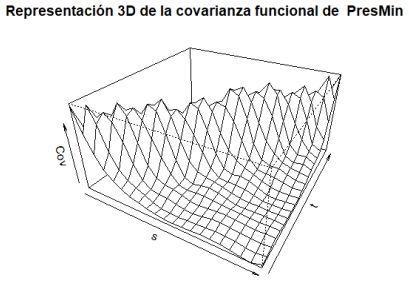


Figura 4.24. Covarianza funcional de PresMin

Análisis descriptivo para *CO*

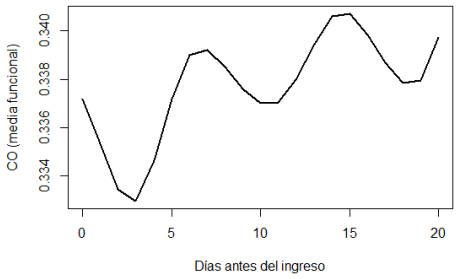


Figura 4.25. Media funcional de CO

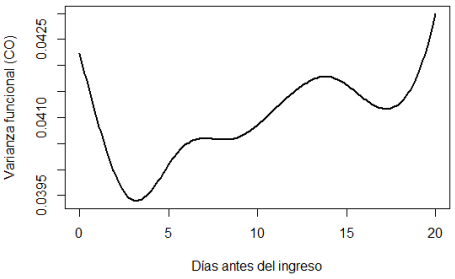


Figura 4.26. Varianza funcional de CO

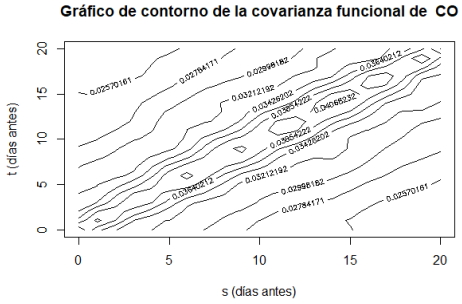


Figura 4.27. Covarianza funcional de CO

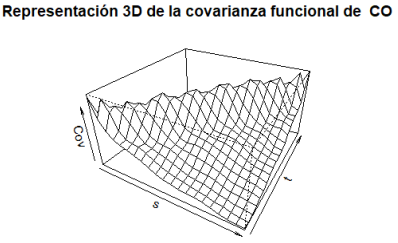


Figura 4.28. Covarianza funcional de CO

Análisis descriptivo para *NO*

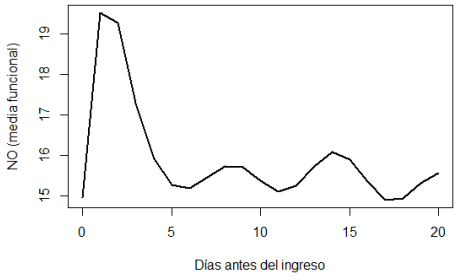


Figura 4.29. Media funcional de NO

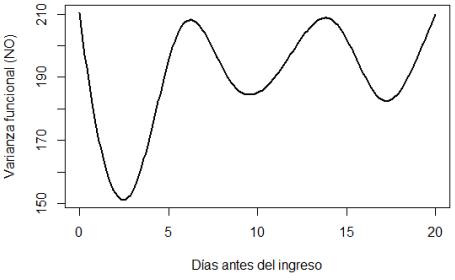


Figura 4.30. Varianza funcional de NO

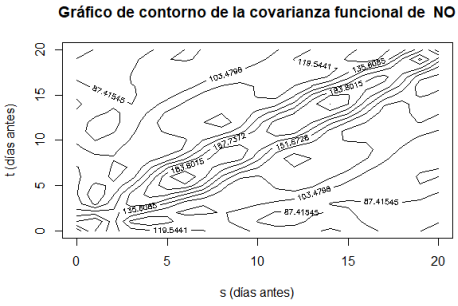


Figura 4.31. Covarianza funcional de NO

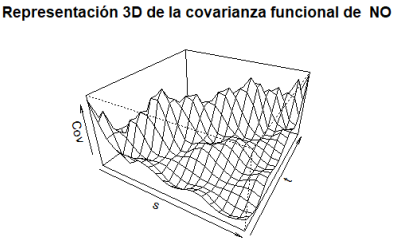


Figura 4.32. Covarianza funcional de NO

Análisis descriptivo para NO_2

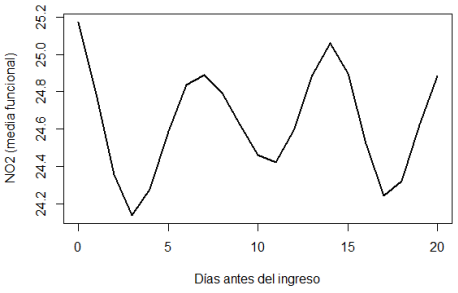


Figura 4.33. Media funcional de NO_2

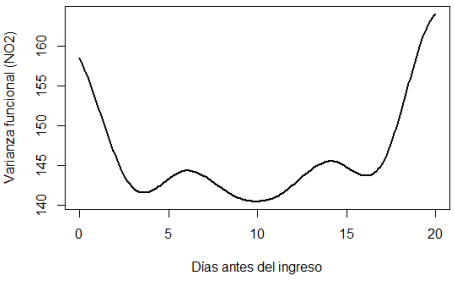


Figura 4.34. Varianza funcional de NO_2

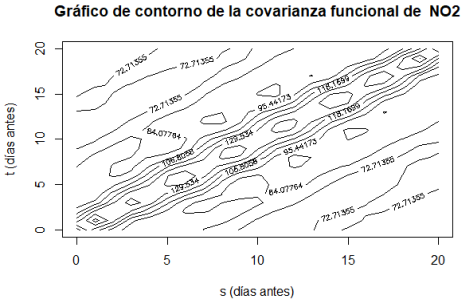


Figura 4.35. Covarianza funcional de NO_2

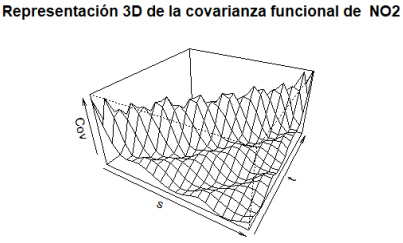


Figura 4.36. Covarianza funcional de NO_2

Análisis descriptivo para SO_2

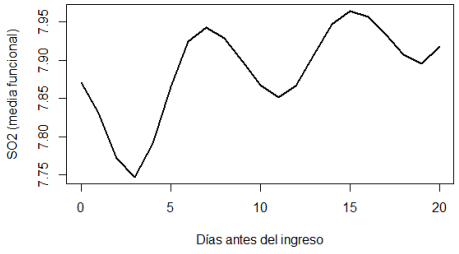


Figura 4.37. Media funcional de SO_2

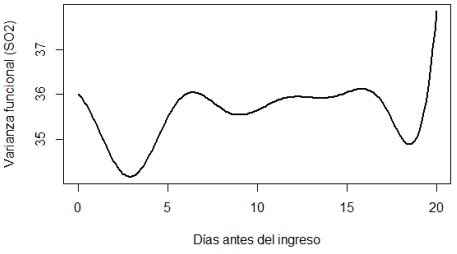


Figura 4.38. Varianza funcional de SO_2

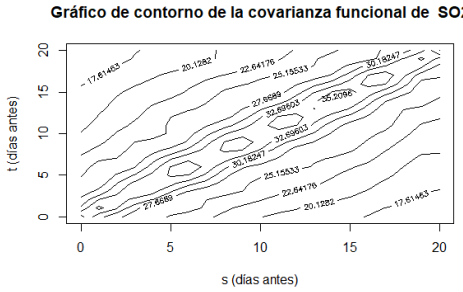


Figura 4.39. Covarianza funcional de SO_2

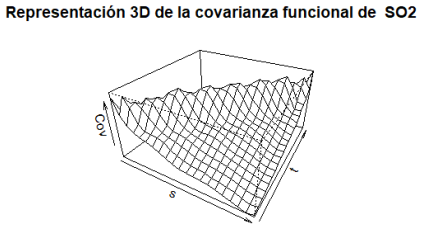


Figura 4.40. Covarianza funcional de SO_2

4. FPCA

FPCA para *tMin*

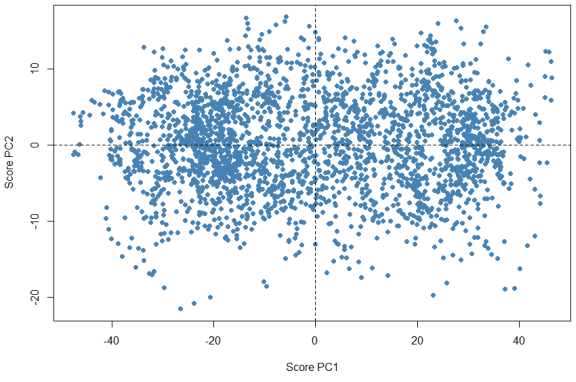


Figura 4.41. Scores de las dos primeras componentes principales de la variable tMin

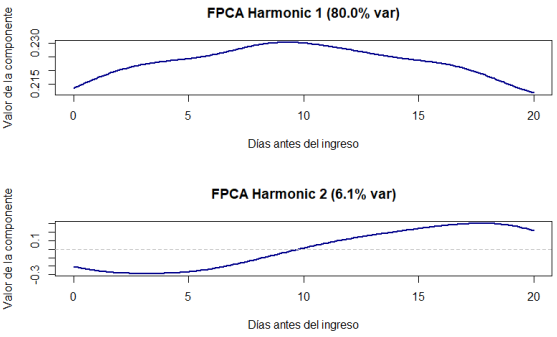


Figura 4.42. Dos primeras autofunciones de la variable tMin

FPCA para *tMax*

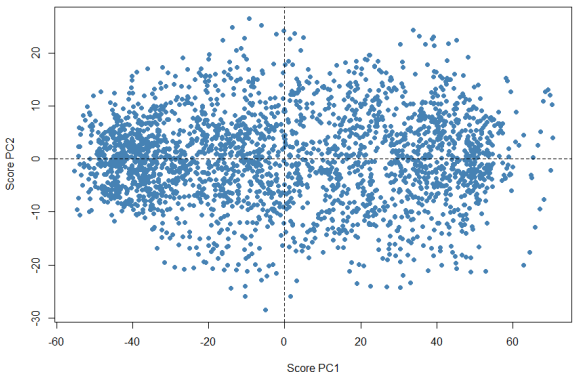


Figura 4.43. Scores de las dos primeras componentes principales de la variable tMax

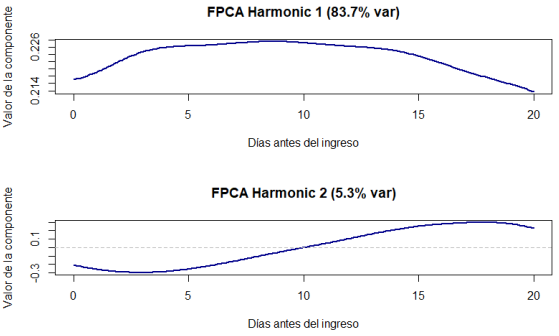


Figura 4.44. Dos primeras autofunciones de la variable tMax

FPCA para *velMedia*

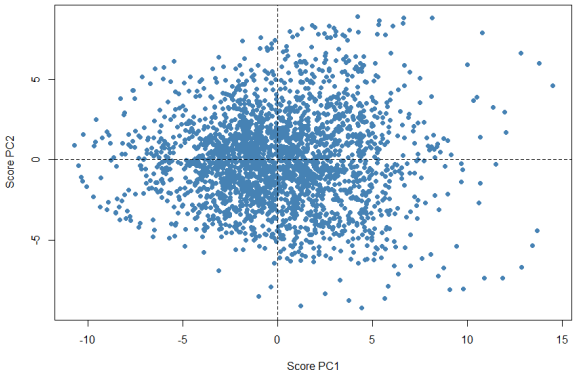


Figura 4.45. Scores de las dos primeras componentes principales de la variable velMedia

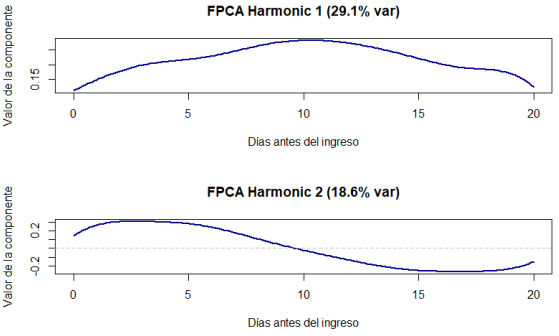


Figura 4.46. Dos primeras autofunciones de la variable velMedia

FPCA para *sol*

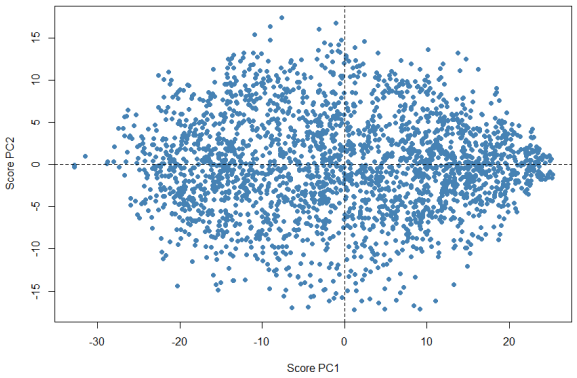


Figura 4.47. Scores de las dos primeras componentes principales de la variable sol

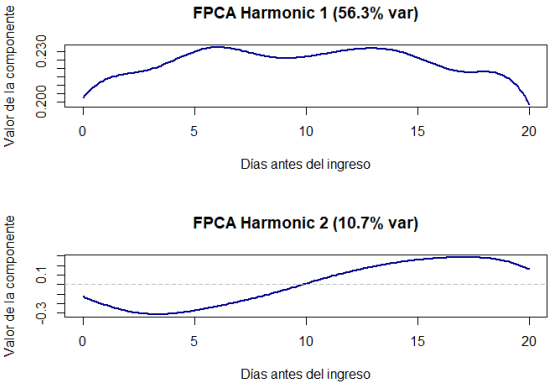


Figura 4.48. Dos primeras autofunciones de la variable sol

FPCA para *PresMax*

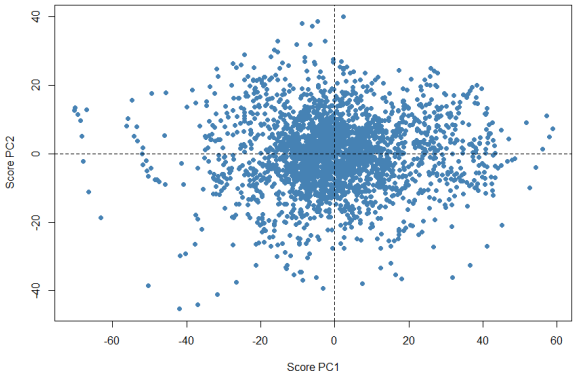


Figura 4.49. Scores de las dos primeras componentes principales de la variable PresMax

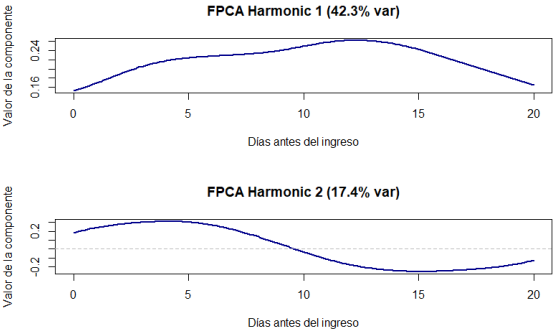


Figura 4.50. Dos primeras autofunciones de la variable PresMax

FPCA para *PresMin*

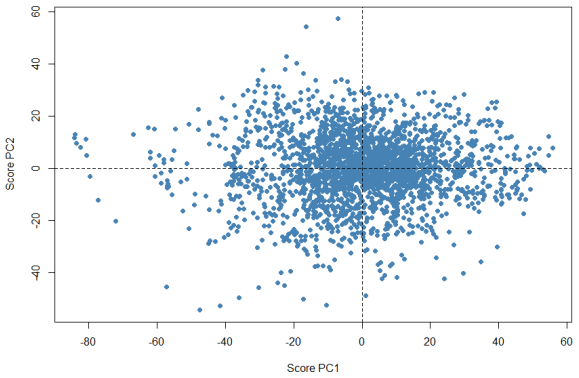


Figura 4.51. Scores de las dos primeras componentes principales de la variable PresMin

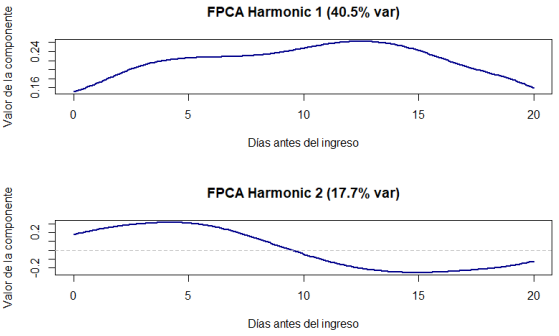


Figura 4.52. Dos primeras autofunciones de la variable PresMin

FPCA para *CO*

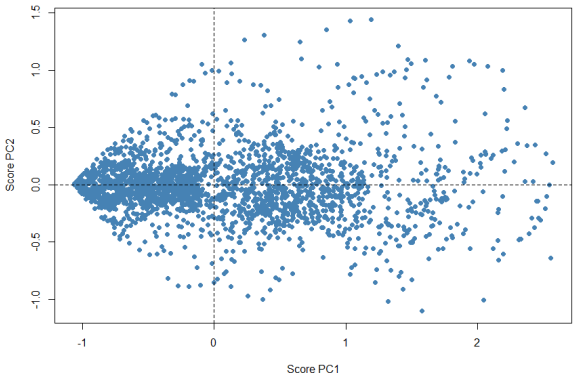


Figura 4.53. Scores de las dos primeras componentes principales de la variable CO

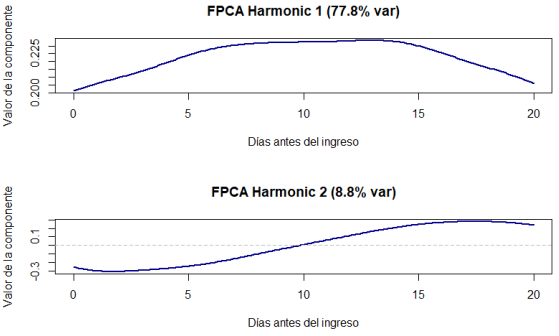


Figura 4.54. Dos primeras autofunciones de la variable CO

FPCA para *NO*

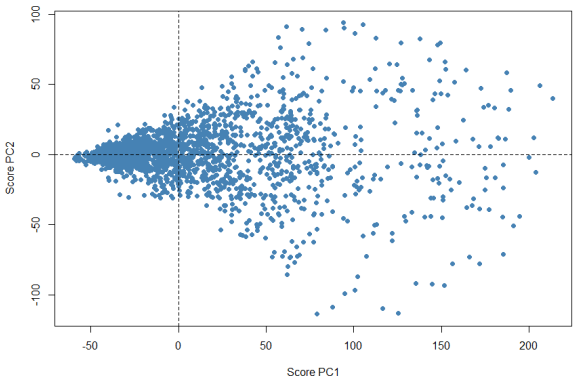


Figura 4.55. Scores de las dos primeras componentes principales de la variable NO

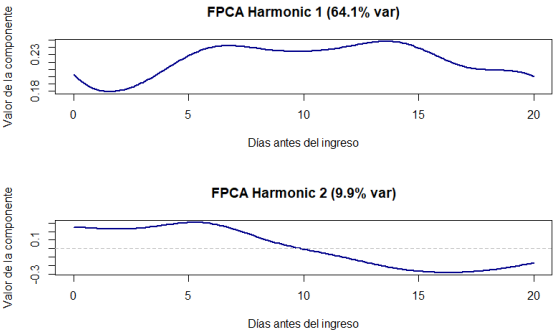


Figura 4.56. Dos primeras autofunciones de la variable NO

FPCA para NO_2

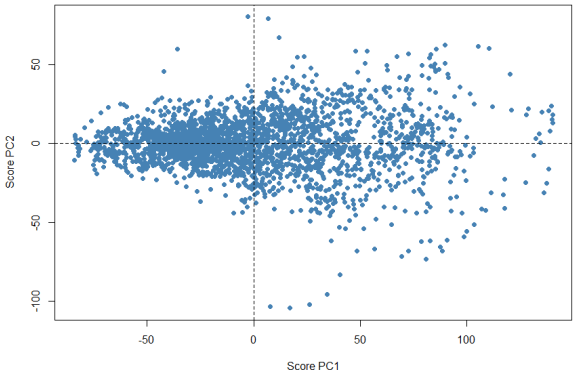


Figura 4.57. Scores de las dos primeras componentes principales de la variable NO_2

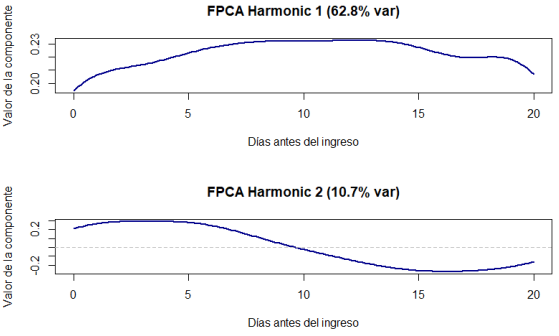


Figura 4.58. Dos primeras autofunciones de la variable NO_2

FPCA para SO_2

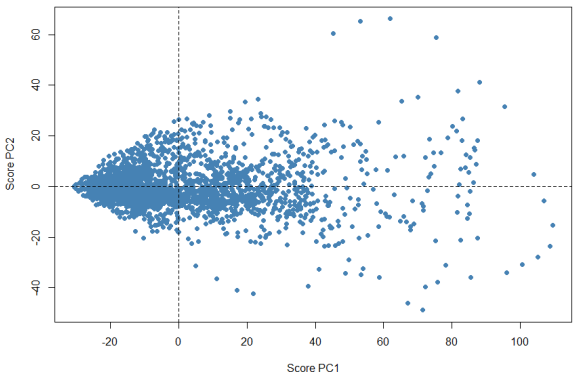


Figura 4.59. Scores de las dos primeras componentes principales de la variable SO_2

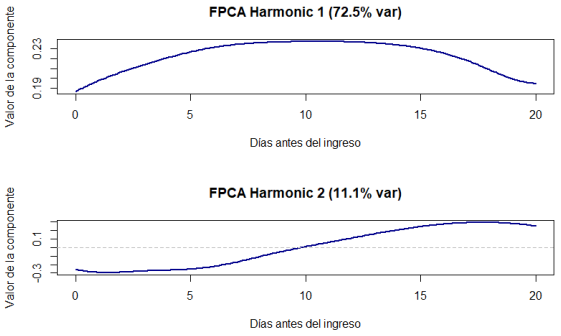


Figura 4.60. Dos primeras autofunciones de la variable SO_2

5. Código R

```
#=============================================================
#================== LIBRERÍAS NECESARIAS =====================
#=============================================================
library(openxlsx)
library(ggplot2)
library(AER)
library(fda)
library(funData)
library(MFPCA)

#=============================================================
#============== LECTURA Y MANIPULACIÓN DE DATOS ==============
#=============================================================
datos <- read.xlsx("D:/TFG/Base.xlsx",detectDates = T,sheet = "GENERICA",rows
    =1:4212)
x<-table(datos$Tipo.Ingreso);x;round(prop.table(x)*100,2)
datos<-datos[-which(datos$Tipo.Ingreso=="Programado"),]
datos_MET <- read.xlsx("D:/TFG/Base.xlsx",detectDates = T,sheet = "Meteo",rows
    =1:4212)
datos_CONT <- read.xlsx("D:/TFG/Base.xlsx",detectDates = T,sheet = "CONTA",rows
    =1:4212)

## Valores ausentes: imputacion mediante splines
imputar_spline <- function(variable_funcional, spar = 0.6) {
  n <- length(variable_funcional)
  tiempo <- seq_len(n)
```

```
  obs <- !is.na(variable_funcional)
  tiempo_obs   <- tiempo[obs]
  valores_obs  <- variable_funcional[obs]
  fit <- smooth.spline(x = tiempo_obs, y = valores_obs, spar = spar)
  y_hat <- predict(fit, x = tiempo)$y
  y_hat[tiempo < min(tiempo_obs)] <- valores_obs[1]
  y_hat[tiempo > max(tiempo_obs)] <- tail(valores_obs, 1)
  imputado <- variable_funcional
  imputado[is.na(imputado)] <- y_hat[is.na(imputado)]
  return(imputado)
}

## Datos censurados: modelo Tobit para la variable prec
datos$FechaIngreso <- as.Date(datos$Fecha.Ingreso, format = "%d/%m/%Y")
# Variables explicativas del modelo Tobit: Estacion, tMedia, sol
datos$Estacion <- factor(
  case_when(
    month(datos$FechaIngreso) %in% c(12, 1, 2)  ~ "Invierno",
    month(datos$FechaIngreso) %in% c(3, 4, 5)   ~ "Primavera",
    month(datos$FechaIngreso) %in% c(6, 7, 8)   ~ "Verano",
    month(datos$FechaIngreso) %in% c(9, 10, 11) ~ "Otoño"
  ),
  levels = c("Primavera", "Verano", "Otoño", "Invierno")
)
datos$tMedia <- as.numeric(datos$tMedia)
datos$sol    <- as.numeric(datos$sol)

prec_vars <- c("prec", paste0("prec", 1:20))

for (var in prec_vars) {
  datos[[paste0(var, "_imputado")]] <- NA  # inicializar columnas de imputacion
}

for (var in prec_vars) {
  datos[[var]][ datos[[var]] == "lp" ] <- 0.1
  datos[[var]][ datos[[var]] == 0 ]      <- 0.1
  datos[[var]] <- as.numeric(datos[[var]])
  idx_completos <- complete.cases(datos[[var]], datos$Estacion, datos$tMedia,
      datos$sol)
  datos_modelo <- datos[idx_completos, ]
  formula_tobit <- as.formula(paste0(var, " ~ Estacion + sol + tMedia"))
  modelo_tobit <- tobit(
```

```
    formula_tobit,
    data = datos_modelo,
    left = 0.1
  )
  # Predecir valores esperados
  pred_val <- predict(modelo_tobit, type = "response")

  # Valor absoluto para garantizar positividad
  pred_val <- abs(pred_val)

  # Asignar las predicciones al dataset original
  datos[[paste0(var, "_imputado")]][idx_completos] <- pred_val
}

#=============================================================
#==================== SUAVIZADO DE DATOS =====================
#=============================================================
time_pts <- 0:20
## Base B-splines
nbasis   <- 10
basis_fd <- create.bspline.basis(rangeval = c(0,20), nbasis = nbasis)
fdPar0   <- fdPar(basis_fd, Lfdobj = 2, lambda = 0)
vars <- c("tMedia",  "tMin", "tMax", "velMedia", "sol", "PresMax", "PresMin", "
    CO", "NO", "NO2", "O3", "PM10", "SO2")
fd_list <- vector("list", length(vars))
names(fd_list) <- vars

for (v in vars) {
  cols <- c(v, paste0(v, 1:20))
  sub  <- datos[, cols]
  mat      <- sapply(sub, function(col) as.numeric(as.character(col)))
  mat_t    <- t(mat)
  smoothR  <- smooth.basis(argvals   = time_pts,
                           y         = mat_t,
                           fdParobj  = fdPar0)
  fd_list[[v]] <- smoothR$fd
}

#=============================================================
#========== ESTADÍSTICOS DESCRIPTIVOS FUNCIONALES ============
#=============================================================
## Media funcional
```

```r
mean_fd_list <- lapply(fd_list, mean.fd)
m_O3 <- eval.fd(time_pts, mean_fd_list[["O3"]])

plot(time_pts, m_O3, type = "l", lwd=2,
     xlab = "Días antes del ingreso",
     ylab = "O3 (media funcional)")

## Varianza funcional
var_vect <- vector("list", length(vars))
names(var_vect) <- vars

var_fd_list    <- vector("list", length(vars))
names(var_fd_list) <- vars

for(v in vars) {
  X       <- eval.fd(time_pts, fd_list[[v]])
  var_vec <- apply(X, 1, stats::var)
  var_vect[[v]] <- var_vec
  smoothVar <- smooth.basis(argvals    = time_pts,
                            y          = matrix(var_vec, ncol=1),
                            fdParobj   = fdPar0)
  var_fd_list[[v]] <- smoothVar$fd
}

plot(time_pts, var_vect[["O3"]], type="b",
     xlab="Días antes del ingreso",
     ylab="Varianza puntual (O3)",
     main="Varianza puntual de O3")

# Gráfico de la varianza suavizada como objeto fd
var_O3_fd <- var_fd_list[["O3"]]
malla_fina <- seq(0,20, length.out=200)
var_smooth_vals <- eval.fd(malla_fina, var_O3_fd)
plot(malla_fina, var_smooth_vals, type="l", lwd=2,
     xlab="Días antes del ingreso",
     ylab="Varianza funcional (O3)")

## Covarianza funcional
cov_mat_list  <- vector("list", length(vars))
names(cov_mat_list) <- vars
cov_bifd_list <- vector("list", length(vars))
names(cov_bifd_list) <- vars
```

```r
for(v in vars) {
  fdobj <- fd_list[[v]]
  X    <- eval.fd(time_pts, fdobj)
  CovM <- cov(t(X), use = "pairwise.complete.obs")
  CovM <- as.matrix(CovM)
  cov_mat_list[[v]]  <- CovM
  cov_bifd_list[[v]] <- var.fd(fdobj)
}

for(v in vars) {
  CovM <- cov_mat_list[[v]]
  zr <- range(CovM, finite = TRUE)
  if(diff(zr) == 0) {
    levels <- zr

  } else {
    levels <- seq(zr[1], zr[2], length.out = 10)
  }

  # Gráfica de contorno
  contour(x      = time_pts,
          y      = time_pts,
          z      = CovM,
          levels = levels,
          xlab   = "s (días antes)",
          ylab   = "t (días antes)",
          main   = paste("Gráfico de contorno de la covarianza funcional de ", v
              ))

  # Representación 3D
  persp(x      = time_pts,
        y      = time_pts,
        z      = CovM,
        theta  = 30,
        phi    = 30,
        expand = 0.5,
        xlab   = "s",
        ylab   = "t",
        zlab   = "Cov",
        main   = paste("Representación 3D de la covarianza funcional de ", v))
}
```

```r
#================================================================
#============================ FPCA ==============================
#================================================================
columnas_tmedia <- grep("^tMedia[0-9]*$", colnames(datos), value = TRUE)
columnas_tmedia <- columnas_tmedia[order(nchar(columnas_tmedia), columnas_tmedia
    )]
matriz_temp <- as.matrix(datos[, columnas_tmedia])

# Crear base de B-splines
nbasis <- 10
basis_fd <- create.bspline.basis(rangeval = c(0,20), nbasis = nbasis)

# Suavizar datos (ajuste funcional)
fd_obj <- Data2fd(argvals = tiempo, y = t(matriz_temp), basisobj = basis_fd)

# FPCA
pca_fd <- pca.fd(fd_obj, nharm = 4)

# Gráficos
plot.pca.fd(pca_fd)

# Porcentaje de varianza explicada
round(pca_fd$varprop * 100, 2)

# Plot FPCA  atofunciones
t_grid <- seq(0, 20, length.out = 200)
n_comp <- 2

par(mfrow = c(n_comp, 1))
for (i in 1:n_comp) {
  harmonic_fd <- pca_fd$harmonics[i]
  y_vals <- eval.fd(t_grid, harmonic_fd)

  plot(t_grid, y_vals, type = "l", lwd = 2, col = "darkblue",
       main = paste("FPCA Harmonic", i,
                    sprintf("(%.1f%% var)", pca_fd$varprop[i]*100)),
       xlab = "Días antes del ingreso", ylab = "Valor de la componente")
  abline(h = 0, col = "gray", lty = 2)
}

# Plot de scores
```

```r
resultados_PCA[["tMedia"]] <- list(fd_suavizada = fd_obj, PCA = pca_fd)
scores <- resultados_PCA[["tMedia"]]$PCA$scores

plot(scores[,1], scores[,2],
     xlab = "Score PC1",
     ylab = "Score PC2",
     pch  = 19, col = "steelblue")
abline(h = 0, v = 0, lty = 2, col = "black")


#=============================================================
#============================ MFPCA ==========================
#=============================================================
var_names   <- c("tMedia", "prec_imputado", "O3", "PM10")
n_vars      <- length(var_names)
t_points    <- seq(0, 20, length.out = 21)
M           <- 5

n_ind       <- nrow(datos)
data_array  <- array(NA, dim = c(n_ind, length(t_points), n_vars))
for (v in seq_along(var_names)) {
  base <- var_names[v]
  if (base == "prec_imputado") {
    cols <- c("prec_imputado", paste0("prec", 1:20, "_imputado"))
  } else {
    cols <- c(base, paste0(base, 1:20))
  }
  data_array[ , , v] <- as.matrix(datos[, cols])
}

# Convertir cada dimensión en un objeto funData
fds_list <- lapply(seq_along(var_names), function(v) {
  funData(argvals = t_points, X = data_array[ , , v])
})
names(fds_list) <- var_names

multi_fd <- multiFunData(fds_list)

uniExp <- lapply(seq_len(n_vars), function(i) {
  list(type = "splines1D", bs = "ps")
})
```

```r
# Escalamos los datos
vars <- sapply(multi_fd, function(fd) mean(apply(fd@X, 2, stats::var)))
weights <- 1 / vars
weights <- weights / sum(weights)

res_mfpca <- MFPCA(
  multi_fd,
  M              = M,
  uniExpansions = uniExp,
  weights        = weights
)

k <- 1 # PC1
nombres_dim <- c("tMedia", "prec",  "O3", "PM10")
basis_list <- predict(res_mfpca, newdata = NULL, scores = diag(1, 5))

par(mfrow = c(2, 2), mar = c(4, 4, 2, 1))

for (j in 1:4) {
  phi_kj <- basis_list[[j]]@X[k, ]  # curva de la dimensión j en la componente k
  t_vals <- basis_list[[j]]@argvals[[1]]

  plot(t_vals, phi_kj, type = "l",
       main = paste("  ",k, "-", nombres_dim[j]),
       xlab = "Tiempo", ylab = expression(phi(t)))
  abline(h = 0, col = "gray", lty = 2)
}

# Plot scores PC1 vs PC3 en funcion de la estación
scores_mat <- res_mfpca$scores
df_scores <- data.frame(
  PC1      = scores_mat[, 1],
  PC3      = scores_mat[, 3],
  Estacion = datos$Estacion
)

ggplot(df_scores, aes(x = PC1, y = PC3, color = Estacion)) +
  geom_point(size = 1.5, alpha = 0.7) +
  labs(
    x     = "Score PC1",
    y     = "Score PC3",
    color = "Estación"
```

```
) +
theme_minimal() +
theme(legend.position = "right")
```