

# Tarea 1

Juan Rubio Cobeta

January 17, 2026

## Contents

<b>1</b>	<b>Definición de Dato Funcional y Objetivos del Análisis</b>	<b>2</b>
<b>2</b>	<b>Revisión Bibliográfica de Investigaciones Recientes</b>	<b>3</b>
2.1	Modelización predictiva en agricultura de precisión . . . . .	3
2.2	Análisis de datos espaciales en oncología a nivel de célula única . . . . .	3
2.3	Dinámica espaciotemporal de la pandemia COVID-19 . . . . .	4
2.4	Clasificación de perfiles de consumo energético industrial . . . . .	4
2.5	Detección de anomalías funcionales en procesos industriales y de imagen . . . . .	5
2.6	Análisis cinemático de corredores recreativos: un enfoque multivariante . . . . .	5

# 1 Definición de Dato Funcional y Objetivos del Análisis

El **Análisis de Datos Funcionales (FDA)** surge como una extensión natural de la estadística multivariante clásica para abordar el estudio de fenómenos cuya naturaleza intrínseca es continua. Mientras que la estadística tradicional opera con vectores de dimensión finita, el FDA considera que la unidad observacional básica es una **función**  $x(t)$  definida sobre un dominio continuo  $T$  (usualmente un intervalo de tiempo, longitud de onda o espacio).

Desde una perspectiva formal, tal y como se detalla en los apuntes de la asignatura y en la literatura especializada (Horváth y Kokoszka, 2012), una variable funcional se define como una variable aleatoria que toma valores en un espacio de dimensión infinita. Típicamente, se asume un **espacio de Hilbert**  $\mathcal{L}_2(T)$  de funciones de cuadrado integrable, dotado del producto escalar estándar:

$$\langle f, g \rangle = \int_T f(t)g(t)dt \quad (1)$$

Una distinción crucial en la teoría es la diferencia entre el dato observado y la entidad teórica subyacente. Aunque en la práctica los datos se registran en una malla discreta de puntos  $\{t_1, \dots, t_m\}$ , a menudo contaminados por ruido de medida, el paradigma del FDA asume el siguiente modelo generador:

$$y_j = x(t_j) + \epsilon_j \quad (2)$$

donde  $x(t)$  es una función suave y  $\epsilon_j$  representa el error aleatorio. Esta **suavidad** es la propiedad fundamental que permite el uso de información sobre derivadas y la estructuración de la covarianza entre puntos vecinos.

Los **objetivos principales** que persigue el Análisis de Datos Funcionales pueden sintetizarse en:

- **Representación y Reconstrucción:** Transformar las observaciones discretas en objetos funcionales continuos mediante técnicas de suavizado o expansión en bases (B-splines, Fourier), eliminando el ruido observacional para recuperar la señal subyacente  $x(t)$ .
- **Análisis Exploratorio Funcional:** Extender los estadísticos descriptivos al contexto funcional, caracterizando la tendencia central mediante la función media y la estructura de variabilidad a través de operadores de covarianza y superficies de correlación.
- **Estudio de la Dinámica:** Aprovechar la naturaleza continua para estimar derivadas de las curvas (velocidad, aceleración), revelando características del fenómeno invisibles en los datos brutos.
- **Inferencia y Modelización:** Desarrollar modelos estadísticos complejos (ANOVA funcional, regresión funcional, Análisis de Componentes Principales Funcionales - FPCA) que permitan la reducción de dimensión, la clasificación y la predicción, superando las limitaciones de los métodos multivariantes ante la alta dimensionalidad y la colinealidad.

## 2 Revisión Bibliográfica de Investigaciones Recientes

A continuación se presentan cinco estudios publicados en el periodo 2024-2025 que ilustran la versatilidad del Análisis de Datos Funcionales (FDA) en la resolución de problemas reales, abarcando disciplinas como la agronomía, la biomedicina y la economía.

### 2.1 Modelización predictiva en agricultura de precisión

**Referencia:** Matsui, H., & Mochida, K. (2024). Functional data analysis-based yield modeling in year-round crop cultivation. *Horticulture Research*, 7 páginas.

<https://academic.oup.com/hr/article/11/7/uhae144/7680591?login=false>

#### Resumen de la investigación:

Este trabajo propone un marco metodológico basado en regresión funcional para modelar y predecir el rendimiento de cultivos de ciclo continuo (específicamente tomate y fresa) en entornos controlados. El **tipo de datos analizados** consiste en curvas longitudinales de variables ambientales monitorizadas diariamente, tales como la temperatura acumulada, la radiación solar y la concentración de CO<sub>2</sub>. A diferencia de los enfoques tradicionales que resumen estas series temporales en escalares estáticos (medias o totales mensuales), este estudio trata las variables ambientales como predictores funcionales  $X(t)$ , preservando la estructura temporal completa y la dinámica de las fluctuaciones climáticas a lo largo del ciclo de crecimiento.

En cuanto a los **resultados principales**, la investigación demuestra que los modelos de regresión lineal funcional superan significativamente en capacidad predictiva a los modelos convencionales (como la regresión de mínimos cuadrados parciales basada en valores agregados). Los autores identifican, mediante el análisis de los coeficientes funcionales estimados  $\beta(t)$ , ventanas temporales críticas donde las condiciones ambientales ejercen un impacto mayor sobre el rendimiento final. Esto permite no solo una predicción más robusta, sino también una interpretación biológica de cómo la variabilidad ambiental en etapas específicas del desarrollo fenológico afecta a la productividad del cultivo.

### 2.2 Análisis de datos espaciales en oncología a nivel de célula única

**Referencia:** Wrobel, J., Soupir, A., & Fridley, B. L. (2024). mxdfa: a comprehensive toolkit for functional data analysis of single-cell spatial data. *Bioinformatics Advances*, 5 páginas.

<https://academic.oup.com/bioinformaticsadvances/article/4/1/vbae155/7899878>

#### Resumen de la investigación:

Este estudio se sitúa en la frontera entre la biología computacional y la estadística, abordando el análisis del microambiente tumoral mediante tecnologías de imagen multiplexada. El **tipo de datos analizados** no son curvas temporales clásicas, sino funciones espaciales derivadas de la ubicación de células individuales. Específicamente, los autores transforman los patrones de puntos espaciales (posiciones  $(x, y)$  de células inmunes y tumorales) en datos funcionales mediante "funciones de resumen espacial" (como la función K de Ripley o densidades radiales). De esta forma, se modela la densidad de células inmunes como una función continua de la distancia al tumor o a otras células de interés, permitiendo tratar la arquitectura del tejido como un objeto funcional  $X(r)$ .

Entre los **resultados principales**, el trabajo presenta la validación del paquete de software 'mxdfa' desarrollado en R, aplicándolo a un conjunto de datos reales de cáncer de ovario. El análisis funcional reveló que la distribución espacial de las células inmunes (perfil de infiltración) tiene una asociación significativa con la supervivencia de los pacientes y el estadio del tumor, información que se pierde si solo se utilizan conteos simples o porcentajes de células. El enfoque FDA permitió identificar firmas espaciales específicas (picos de densidad inmunológica a distancias concretas del tumor) que actúan como biomarcadores pronósticos.

## 2.3 Dinámica espaciotemporal de la pandemia COVID-19

**Referencia:** Ribeiro, M., Azevedo, L., Santos, A. P., Pinto Leite, P., & Pereira, M. J. (2024). Understanding spatiotemporal patterns of COVID-19 incidence in Portugal: A functional data analysis from August 2020 to March 2022. *PLOS ONE*, 23 páginas.

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0297772>

### Resumen de la investigación:

Esta investigación aplica el Análisis de Datos Funcionales para desentrañar la complejidad de la propagación epidemiológica del SARS-CoV-2 en Portugal, superando las limitaciones de los análisis clásicos de series temporales que a menudo ignoran la estructura de correlación intrínseca de la curva epidémica. El **tipo de datos analizados** consiste en las tasas de incidencia diaria por municipio, las cuales son suavizadas y transformadas en objetos funcionales continuos  $X_i(t)$ , donde  $i$  representa cada municipio y  $t$  el tiempo en el intervalo estudiado (agosto 2020 - marzo 2022). Este enfoque permite capturar no solo la magnitud de los contagios, sino la velocidad y la aceleración de las olas pandémicas (a través de las derivadas  $X'_i(t)$  y  $X''_i(t)$ ), tratando la evolución sanitaria como un proceso estocástico suave en lugar de una colección de puntos discretos ruidosos.

En lo referente a los **resultados principales**, el estudio implementa técnicas de agrupamiento funcional (*functional clustering*) para clasificar los municipios según la "forma" de sus curvas de contagio, independientemente de sus niveles absolutos. Los autores identificaron patrones espaciotemporales distintivos que revelan una clara dicotomía entre las áreas metropolitanas densamente pobladas y las regiones rurales. El análisis funcional permitió detectar desfases temporales (*time-lags*) en el inicio y pico de las olas entre diferentes clústeres, demostrando que la difusión del virus siguió dinámicas funcionales heterogéneas que las medidas de resumen estáticas no logran detectar. Estas conclusiones sugieren que las políticas de salud pública deberían adaptarse a estas "firmas funcionales" regionales en lugar de aplicarse uniformemente.

## 2.4 Clasificación de perfiles de consumo energético industrial

**Referencia:** Rivera-García, D., & Mateu, J. (2025). Exploring energy consumption patterns in Colombian companies: a functional data clustering approach. *Communications for Statistical Applications and Methods*, 32 páginas. <http://www.csam.or.kr/journal/view.html?uid=2198&vmd=Full>

### Resumen de la investigación:

Este trabajo aborda un problema crítico en la economía energética moderna: la caracterización eficiente de la demanda eléctrica en el sector empresarial. El **tipo de datos analizados** corresponde a curvas de carga horaria, es decir, el consumo de electricidad registrado continuamente a lo largo del tiempo para un conjunto heterogéneo de empresas. En el marco del FDA, cada empresa  $i$  se representa mediante una función  $X_i(t)$  que describe su perfil de potencia diario o semanal. El tratamiento funcional de estos datos permite capturar características dinámicas complejas, como la rampa de subida al inicio de la jornada laboral, la estabilidad del consumo nocturno o la variabilidad durante las horas pico, elementos que se diluyen al utilizar promedios mensuales.

Los **resultados principales** se derivan de la aplicación de técnicas de análisis de clúster para datos funcionales (*functional clustering*). Los autores logran segmentar la cartera de clientes en grupos homogéneos con comportamientos de consumo distintivos (por ejemplo, perfiles con picos diurnos marcados frente a perfiles planos de consumo continuo). Esta segmentación es superior a la clasificación tradicional basada en códigos de actividad económica (CNAE), ya que se basa en la forma real de la curva de demanda. El estudio concluye que este enfoque permite a las compañías eléctricas diseñar tarifas personalizadas más ajustadas y estrategias de gestión de la demanda más eficientes, optimizando la red eléctrica basándose en patrones reales de uso.

## 2.5 Detección de anomalías funcionales en procesos industriales y de imagen

**Referencia:** Alcacer, A., & Epifanio, I. (2024). Outlier detection of clustered functional data with image and signal processing applications by archetype analysis. *PLOS ONE*, 23 páginas.

**Disponible en:** <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0311418>

### Resumen de la investigación:

Este artículo presenta una metodología innovadora para el control de calidad y el procesamiento de señales mediante el análisis de datos atípicos (*outliers*) en conjuntos de datos funcionales que presentan una estructura de grupos latente. El **tipo de datos analizados** es variado, abarcando desde espectrometrías en la industria alimentaria hasta perfiles de imágenes médicas. En el contexto FDA, estas observaciones se tratan como funciones  $X(t)$  o superficies  $X(u, v)$  donde la continuidad es esencial para definir qué constituye una desviación "normal" frente a una anomalía patológica. A diferencia de los métodos univariantes que buscan valores extremos puntuales, aquí se buscan funciones con formas anómalas (*outliers* de forma) o magnitudes desviadas respecto a su clúster de pertenencia.

Los **resultados principales** destacan la eficacia del uso del Análisis de Arquetipos (*Archetype Analysis*) adaptado al contexto funcional. Los autores proponen un algoritmo que no asume una única distribución homogénea para todos los datos, sino que permite que existan múltiples patrones de comportamiento estándar. Al aplicar esta técnica a datos reales de espectrometría de carne y dígitos manuscritos, el método demostró una sensibilidad superior para detectar anomalías sutiles que métodos tradicionales (como la profundidad funcional estándar) pasaban por alto, reduciendo significativamente la tasa de falsos positivos en entornos de producción industrial.

## 2.6 Análisis cinemático de corredores recreativos: un enfoque multivariante

**Referencia:** Gunning, E., Golovkine, S., Simpkin, A. J., et al. (2025). Analysing kinematic data from recreational runners using functional data analysis. *Computational Statistics*, 44 páginas.

**Disponible en:** <https://arxiv.org/abs/2408.08200>

### Resumen de la investigación:

Este trabajo presenta una aplicación avanzada del FDA en el ámbito de las ciencias del deporte, abordando la complejidad del movimiento humano mediante modelos de efectos mixtos. El **tipo de datos analizados** consiste en curvas cinemáticas multivariantes, específicamente los ángulos de flexión de la cadera y la rodilla en el plano sagital, registrados durante la carrera en cinta de un amplio grupo de corredores recreativos. En lugar de reducir estas series temporales a valores discretos (como el ángulo máximo o mínimo), los autores modelan la trayectoria completa del ciclo de la zancada como una función bivalente  $X(t) = (X_{cadera}(t), X_{rodilla}(t))$ , preservando la estructura de correlación continua entre ambas articulaciones.

Entre los **resultados principales**, destaca la propuesta de un "Modelo Lineal Funcional de Efectos Mixtos Multivariante" que permite separar la variabilidad debida a factores fijos (como la velocidad de carrera) de la variabilidad aleatoria específica de cada sujeto. El estudio confirma que la velocidad de carrera altera significativamente la forma funcional de los ángulos articulares, pero lo hace de manera no uniforme a lo largo del ciclo de la marcha. Además, el análisis de las componentes principales funcionales reveló que cada corredor posee una "firma cinemática" única y altamente repetible (fuerte correlación intra-sujeto), lo que sugiere que las estrategias de prevención de lesiones deben personalizarse basándose en el perfil funcional completo del deportista y no en métricas estándar generales.