DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Estudio y aplicación de técnicas de Análisis de Datos Funcionales de geoposicionamiento

Sonia María Pérez Plaza

Directores: Fernando Fernández Palacín y Manuel Berrocoso Domínguez

Puerto Real (Cádiz), 2020

AGRADECIMIENTOS

"El agradecimiento es la memoria del corazón" (Lao Tsé)

Gracias a mis abuelos, mis padres y hermana. Sin ellos no sería la persona que hoy soy. Sus consejos, su amor y su ejemplo me enseñaron la importancia de ser honesta, trabajadora y paciente. Gracias por apoyarme siempre. Estuviéseis de acuerdo o no, por muy locas que pareciesen mis ideas, siempre confiasteis y aún confiáis en mí. No podría tener una familia mejor. Papi, Mami, tata os adoro.

Gracias a mis niñas, Carmen, Carla y Sofía. Sois mi apoyo, la ilusión y la alegría de mi vida. Sin vosotras nada tendría sentido y cualquier cosa que haga, y por la que trabaje duro, siempre será por y para vosotras. Os amo.

Gracias a mis amigos y amigas, su cariño y sus ánimos son un pilar muy importante para mí. Soy pobre, pero muy rica en amigos. Mari, Fernando, Maribel, Carmen, Juanito, Roldana, Novo, Carmela, Sara y Elo sois mis tesoros, mi familia del corazón. Os quiero.

Gracias a los compañeros del Departamento de Estadística e Investigación Operativa que siempre han estado y están dispuestos a ayudarme y apoyarme. Auxi, Antonio, Manolo, Gafas, Toñi, Antonio...mil gracias.

Gracias a mis directores de Tesis. Manolo, muchas gracias por tu apoyo y tus ánimos. No lo olvidaré. Y sobre todo GRACIAS FERNANDO.

Sin tí no habría tesis y ni siquiera sé donde estaría ahora mismo, casi seguro que no sería en la Universidad. Eres una persona admirable y la más generosa que conozco. Me has apoyado y animado siempre y doy gracias a Dios por haberte puesto en mi camino. Soy muy afortunada. Con la suerte que tuve con mi padre, mi papi académico es otro regalo del cielo. Siempre tendrás mi apoyo incondicional y mi eterna gratitud.

RESUMEN

Esta tesis se enmarca dentro del Análisis de Datos Funcionales (FDA) que adapta las técnicas estadísticas clásicas a situaciones en que los datos son funciones. El objetivo general es aplicar métodos del Análisis de Datos Funcionales al estudio de problemas de geoposicionamiento. En cierta medida, el trabajo que se presenta, tiene un enfoque de carácter metodológico, que queda de manifiesto en el capítulo dedicado a la red SPINA. El aplicar técnicas FDA a datos GPS se justifica, desde un punto de vista conceptual, en que, dichas mediciones se corresponden con una monitorización de una magnitud cuya concreción matemática es una curva.

Al comienzo de la tesis se revisan los elementos básicos del FDA, desde la elección de la métrica, hasta la adaptación de las técnicas descriptivas del caso vectorial al funcional. Así, analizamos métricas y semimétricas funcionales en espacios de Hilbert, incluyendo la semimétrica basada en derivadas y la semimétrica basada en componentes principales funcionales. Se realiza un análisis detallado del proceso de suavizado a partir de las distintas bases de funciones. Se repasan los estadísticos sobre una función y los estadísticos sobre una muestra de funciones, teniendo especial interés la covarianza (y la correlación) entre dos muestras de funciones de dos variables. Se detallan los principales aspectos del Análisis Funcional de Componentes Principales y, finalmente, se hace una revisión de los métodos de clasificación no supervisada en el caso funcional.

A continuación, se analizan la naturaleza y las particularidades de los datos geodésicos, en general, y el "sistema de posicionamiento GPS", en particular. Una parte importante está dedicada a la identificación y tratamiento de los errores, tanto sistemáticos como aleatorios. También se detalla el esquema de "Procesado de las observaciones GNSS-GPS" mediante el software de procesamiento de datos geodésicos Bernese. Finalmente se trata la depuración de datos GPS y la posterior reconstrucción o imputación de los mismos mediante filtros Kalman.

La siguiente parte de este trabajo está dedicada al análisis de los datos de posicionamiento suministrados por la red geodésica SPINA (Sur de la Península Ibérica y Norte de África). Como ya se ha comentado, es un trabajo con un enfoque claramente metodológico, donde se detallan todos los pasos a dar al aplicar técnicas FDA. El contenido de esta parte es una adaptación del artículo publicado en el año 2018 en la revista "Mathematical Geosciences" con el título "Analysis of a GPS network based on Functional Data Analysis". Es uno de los primeros trabajos publicados donde se tratan datos de carácter geodésico usando métodos FDA. Entre otras aportaciones, se hace una propuesta original de depuración de datos, para luego imputar en base

al error en las mediciones aportado por el software Bernese.

"Relación funcional entre deformación y sismicidad en El Hierro", constituye también una aplicación del FDA, en este caso a un problema que combina los movimientos sísmicos y la deformación geodésica. Al igual que en el caso anterior se trata de la transcripción de un artículo que, a fecha de hoy, se encuentra en segunda revisión. Los datos proceden del suceso que tuvo lugar en la isla de El Hierro alrededor de octubre de 2011 y cuyo evento principal fue la erupción submarina ocurrida el 10 de octubre de dicho año. El objetivo principal fue relacionar la sísmica con la deformación registrada en la zona, al objeto de modelar un sistema de alerta. En este trabajo se incluye una de las aportaciones originales de la tesis, que es la introducción de una medida de correlación funcional.

Por último, se recogen las principales conclusiones del trabajo, las limitaciones del estudio y la exposición de las principales líneas de trabajo abiertas, que se pretenden abordar en un futuro inmediato.

ABSTRACT

This thesis falls within the Functional Data Analysis (FDA), which adapts classical statistical techniques to situations where data are functions. The general objective of this work is to apply functional data analysis methods to the study of geolocation problems. The work presented has a methodological approach, which is evident in the chapter dedicated to the SPINA network. Applying FDA techniques to data from GPS measurements is justified, from a conceptual point of view, because these measurements correspond to a monitoring process of a magnitude whose mathematical concretion is a curve.

At the beginning of the thesis, the basic elements of the FDA are reviewed, from the choice of the metric to the adaptation of the descriptive techniques of the vectorial case to the functional case. Thus, we analyze functional metrics and semimetrics in Hilbert spaces, including the semimetric based on the derivatives and the semimetric based on the functional principal components. A detailed analysis of the smoothing process is performed from the different function bases. The statistics on a function and the statistics on a sample of functions are reviewed, with special interest in the covariance (and the correlation) between two samples of functions of two variables. The main aspects of the Functional Analysis of Principal Components are detailed and, finally, a review of the unsupervised classification methods in the functional case is made.

Next, the nature and particularities of geodetic data in general, and the "GPS positioning system" in particular, are analyzed. An important part is dedicated to the identification and treatment of errors, both systematic and random. The scheme of "Processing of GNSS-GPS observations" by using Bernese geodetic data processing software is also detailed. Moreover, the cleaning of GPS data and the subsequent reconstruction or imputation of the same by using a Kalman filter is treated.

The following part of the work is devoted to the analysis of the positioning data provided by the SPINA geodetic network (South of the Iberian Peninsula and North of Africa). As already mentioned, it is a work with a clearly methodological approach, which details all the steps to be taken when applying FDA techniques. The content of this part is an adaptation of the article published in 2018 in the magazine " Mathematical Geosciences " with the title " Analysis of a GPS network based on Functional Data Analysis. " It is one of the first published works where geodetic data is processed through FDA methods. Among other contributions, an original proposal for data cleaning is made, and then imputed based on the measurement error provided by the Bernese software.

"Functional relationship between deformation and seismicity in El Hierro", also constitutes an application of the FDA, in this case to a problem that combines the seismic movements and the surface deformation process. Like the previous case, it deals with the transcription of an article that, as of today, is in second revision. The data collects information about the event that took place on the island of El Hierro around October 2011 and whose main event was the submarine eruption that occurred on October 10 of that year. The main objective was to relate the seismic to the deformation registered in the area, in order to model an alert system. This work includes one of the original contributions of the thesis, which is the introduction of a functional correlation measure.

Finally, the main conclusions of the work, the limitations of the study and the exposition of the main lines of open work, that are intended to be addressed in the immediate future are collected.

Índice general

Introducción

Este trabajo de investigación se enmarca dentro del Análisis de Datos Funcionales, familiarmente conocido como FDA, su acrónimo en inglés. El objetivo fundamental del FDA es la aplicación de las técnicas estadísticas clásicas a situaciones en que los datos son funciones, univariantes o multivariantes. Las funciones son pues los elementos unitarios de análisis. Las formas funcionales o curvas, se obtienen mediante procedimientos de suavizado de las observaciones de una variable en un conjunto de individuos. Para realizar dicho suavizado debe elegirse una base de funciones. La dimensión teórica en análisis funcional es infinita, aunque en la práctica dicha dimensión se corresponde con el tamaño de la base de funciones de suavizado. Desde una óptica conceptual, en lo que se refiere a la dimensión, hay una equivalencia entre un problema multivariable vectorial y uno univariable funcional; mientras que desde el punto de vista de la aplicación de los procedimientos, el análisis funcional multivariante combina los aspectos de los análisis de datos funcionales y las técnicas multivariantes vectoriales.

El objetivo general de la tesis es aplicar métodos de análisis funcional al estudio de problemas de geoposicionamiento. En cierta medida, el trabajo que se presenta, tiene un enfoque de carácter metodológico, que queda de manifiesto en el Capítulo 3. El aplicar técnicas FDA a datos provinientes de mediciones GPS se justifica, desde un punto de vista conceptual, en que, dichas mediciones se corresponden con una monitorización de una magnitud cuya concreción matemática es una curva; y desde un punto de vista práctico, en que los modernos sistemas de medición y procesado de datos ponen a nuestra disposición secuencias de observaciones de alta calidad, en una rejilla lo suficientemente densa como para poder hacer una 'reconstrucción' eficiente de las curvas originales.

En el Capítulo 1 se revisan los elementos básicos de un análisis funcional, desde la elección de la métrica en un espacio funcional, hasta la adaptación de las técnicas descriptivas básicas del caso vectorial al funcional. Así, analizamos métricas y

semimétricas funcionales en espacios de Hilbert, incluyendo la semimétrica basada en derivadas y la semimétrica basada en componentes principales funcionales. Se realiza un análisis detallado del suavizado a partir de bases de funciones, incluyendo bases de Splines, Fourier, Wavelets y funciones Núcleo. Se pone de manifiesto la importancia de buscar un equilibrio entre el nivel de ajuste y la calidad del suavizado, con el propósito de que las curvas tengan buenas propiedades analíticas. El que las funciones sean de clase C^2 nos permitirá trabajar con las curvas de velocidad y de aceleración. Se repasan las coeficientes básicos, tanto estadísticos sobre una función como estadísticos sobre una muestra de funciones; especial interés tendrá la adaptación de la covarianza (correlación) entre dos muestras de funciones de dos variables. La covarianza o correlación cruzada jugará un papel fundamental en las técnicas de reducción de la dimensión (como el análisis de componentes principales funcionales) y en aquellas otras que analizan las relaciones entre las variables (como el análisis de correlaciones canónicas funcionales). En la última parte del capítulo se detallan los principales aspectos del Análisis Funcional de Componentes Principales, aprovechando la óptica que nos ofrece el 'esquema de dualidad' y poniendo de manifiesto las propiedades aplicativas derivadas del 'desarrollo de Karhunen-Loéve'. Finalmente, se hace una revisión de los métodos de clasificación no supervisada en el caso funcional.

El segundo Capítulo está dedicado a analizar la naturaleza y particularidades de los datos geodésicos, en general, y el "sistema de posicionamiento GPS", en particular. Una parte importante está dedicada a la identificación y tratamiento de los errores, tanto sistemáticos como aleatorios. También se detalla el esquema de "Procesado de las observaciones GNSS-GPS" mediante el software de procesamiento de datos geodésicos Bernese. En la última parte del capítulo se trata la depuración de datos GPS y la posterior reconstrucción o imputación de los mismos mediante filtros Kalman.

El tercer Capítulo está dedicado al análisis de los datos de posicionamiento suministrados por la red geodésica SPINA (Sur de la Península Ibérica y Norte de África). Como ya se ha comentado, es un trabajo con un enfoque claramente metodológico, donde se detallan todos los pasos a dar en un análisis funcional. El contenido del capítulo es una adaptación del artículo publicado en el año 2018 en la revista "Mathematical Geosciences" con el título "Analysis of a GPS network based on Functional Data Analysis". Es uno de los primeros trabajos publicados donde se tratan datos de carácter geodésico usando técnicas de análisis funcional. Entre otras aportaciones, se hace una propuesta original de depuración de datos, para luego imputar en base al error en las mediciones aportado por el software Bernese.

El Capítulo 4 constituye también una aplicación del FDA, en este caso a un problema

que combina los movimientos sísmicos y las deformaciones geodésicas. El objetivo era explicar la "Relación funcional entre deformación y sismicidad en El Hierro", al igual que el capítulo anterior se trata de la transcripción de un artículo que, a fecha de hoy, se encuentra en segunda revisión. Los datos proceden del suceso que tuvo lugar en la isla de El Hierro en un periodo de tiempo alrededor de octubre de 2011 y cuyo evento principal fue la erupción submarina ocurrida el 10 de octubre de dicho año. El objetivo principal fue relacionar la sísmica con la deformación registrada en la zona al objeto de modelar un sistema de alerta. Una de las aportaciones originales de la tesis es la introducción de una medida de correlación funcional.

Por último, en el Capítulo 5 se recogen las principales conclusiones del trabajo, las limitaciones del estudio y la exposición de las principales líneas de trabajo abiertas, que se pretenden abordar en un futuro inmediato.

Capítulo 1

Análisis de datos funcionales

1.1. Introducción

El análisis de datos funcionales, o FDA en inglés, es una rama de la Estadística que estudia y analiza la información contenida en curvas, superficies o cualquier elemento que varíe en un soporte continuo, frecuentemente el tiempo. Cada observación funcional es una curva que toma valores dentro de un recinto del espacio soporte.

Existen muchos campos experimentales en los que los individuos a estudiar vienen determinados por curvas (geodesia, biomedicina, ciencias ambientales, finanzas, etc.), aunque, por problemas obvios de accesibilidad a dichos datos, de lo que se dispone en la práctica es de una discretización o "esbozo" de las curvas de los individuos en un cierto rango $[t_{min}, t_{max}]$. La disponibilidad de medios tecnológicos para recolectar este tipo de datos en una rejilla suficientemente fina, junto con la capacidad de poder almacenarlos y tratarlos computacionalmente, nos permite proponer una batería de soluciones eficientes para este tipo de problemas. Dichas soluciones se basan en la reconstrucción de las curvas y la adaptación de las técnicas vectoriales clásicas al caso funcional.

En cualquiera de los campos experimentales en el que se esté trabajando, si, para un conjunto de n individuos, $\{I_i, i = 1, 2, \ldots, n; n \geq 1\}$, se tiene la posibilidad de medir en diferentes instantes de tiempo dentro de un intervalo, $\tau \equiv [t_{min}, t_{max}]$, una determinada magnitud, X, se acaba obteniendo, para cada individuo, una secuencia de mediciones de dicha magnitud en una rejilla, $(t_{min}, t_{(2)}, \ldots, t_{(m-1)}, t_{max})$, que puede ser tan fina como nos permitan las condiciones y los recursos disponi-

bles: $\{(x_i(t_1), x_i(t_{(2)}), \ldots, x_i(t_{(m-1)}), x_i(t_m)), i = 1, 2, \ldots, n\}$. Cuando $t_{(j)} - t_{(j-1)} = \frac{t_2 - t_1}{m-1}, j = 2, \ldots, m$, la rejilla es uniforme o regular, correspondiéndose la toma de datos con una monitorización; en otros casos, la elección de los momentos en que se toman las mediciones puede tener carácter discrecional o incluso aleatorio. Idealmente, al aumentar la densidad de puntos de la rejilla, ésta convergerá al propio intervalo, τ, y las secuencias de mediciones se transformarán en curvas que podemos identificar como funciones del tiempo: $\{X_i(t) : t \in \tau, i = 1, 2, \ldots, n\}$.

En muchas situaciones la naturaleza de la magnitud bajo estudio es intrínsecamente funcional, los individuos vienen dados por funciones y, por tanto, el espacio, E, del cual forman parte será un espacio funcional; mientras que en otras situaciones, la magnitud tiene naturaleza discreta y los individuos vienen dados por vectores dentro de un espacio $E \in \mathbb{R}^m$, que podríamos convertir, en según qué casos, en funciones de un "espacio funcional proyectado", E'. En cualquier caso, admitiendo que cada individuo viene dado por una curva, parece razonable aplicar técnicas de análisis de datos adaptadas al espacio funcional.

Una vez que, de una forma u otra, se dispone de una colección de individuos representados por funciones, el objetivo es tratarlos estadísticamente. Podríamos concluir que el análisis de datos funcionales, FDA[1] por sus siglas en inglés, nace de la adaptación de los conceptos y técnicas del análisis estadístico vectorial a un espacio métrico o pseudométrico funcional[2]. A lo largo de los últimos 30 años muchos autores han establecido las bases teóricas del análisis estadístico funcional explorando las posibilidades que ofrece trabajar en un espacio de funciones, entre las que destacan especialmente las aportaciones de Ramsay y Silverman ([1]) y de Ferrety y Vieu ([2]); otros muchos autores han establecido metodologías y adaptado muchas de las técnicas multivariantes vectoriales clásicas a un espacio funcional.

La obtención de las curvas a partir de las secuencias de observaciones de cada individuo, es un problema que deberá ser resuelto en función de las condiciones que se den. El espectro de posibles soluciones a este problema es muy amplio: métodos paramétricos de ajuste, métodos no paramétricos de suavizado y filtrado basados en bases de funciones, soluciones de suavizado local, etc.; en todo caso, siempre deberá exigirse que las curvas reúnan unas mínimas condiciones de regularidad, como por ejemplo que dispongan de primera y segunda derivada. La naturaleza de los datos, su posible carácter periódico, los objetivos del análisis y la densidad y regularidad de la rejilla, deberán tenerse en cuenta a la hora de elegir el procedimiento más eficiente para reconstruir las curvas.

[1] Functional Data Analysis

[2] Lo que va a depender de la manera de cuantificar las diferencias entre curvas.

Obsérvese que si los datos son intrínsecamente funcionales, el procedimiento de obtención de las curvas no es sino una reconstrucción de las mismas a partir de la información parcial de las mediciones obtenidas en la rejilla. El problema se complica si pensamos que las propias mediciones en la rejilla de tiempo podrían no ser exactas y venir dadas con un cierto grado de precisión, o, lo que es lo mismo, estar afectadas por un error de medición. Si los datos fueran de naturaleza discreta, las curvas se obtendrán como proyecciones de los individuos en un espacio funcional dual, lo que permitirá aplicar las técnicas FDA y aprovechar las posibilidades analíticas y computacionales del espacio funcional asociado. En cualquier caso, resulta evidente la importancia que adquiere la obtención o reconstrucción de las curvas en el análisis de datos funcionales, pues de ello dependerá la calidad de la información que será posteriormente tratada con FDA. Más adelante dedicaremos una sección a la obtención de las curvas asociadas a las secuencias de observaciones de los individuos.

Para formalizar conceptualmente la idea de dato funcional usaremos el enfoque propuesto por Ferraty y View ([2]). Siguiendo a estos autores, se dice que "una variable aleatoria $\mathcal{X}$ es una variable aleatoria funcional, si toma valores en un espacio infinito dimensional (o espacio funcional). Una observación χ de $\mathcal{X}$ es un dato funcional".

Conectando con el párrafo anterior, desde un punto de vista funcional una variable aleatoria, $\mathcal{X}$, puede expresarse como $\mathcal{X} = \{\mathcal{X}(t) : t \in \tau \subset \mathbb{R}\}$, es decir, la variable se corresponde con una curva aleatoria en el plano (t, x); mientras que las observaciones o realizaciones de la variable[3], determinadas por $\chi = \{\chi(t) : t \in \tau \subset \mathbb{R}\}$, serían curvas físicas en dicho plano. La variable funcional así definida es de soporte unidimensional. Si el soporte $\tau \subset \mathbb{R}^2$, la variable funcional toma valores sobre el recinto soporte τ del plano $\mathbb{R}^2$: $\mathcal{X} = \{\mathcal{X}(t_1, t_2) : (t_1, t_2) \in \tau \subset \mathbb{R}^2\}$, correspondiéndose con una superficie aleatoria en el espacio $(t_1, t_2, \mathcal{X})$; mientras que las realizaciones de esta variable determinadas por $\chi = \{\chi(t_1, t_2) : (t_1, t_2) \in \tau \subset \mathbb{R}^2\}$, serían superficies físicas en dicho espacio. La variable funcional así definida es de soporte bidimensional. Obsérvese que, en principio, el recinto soporte τ no tiene porqué ser conexo, pudiendo venir determinado por la unión de intervalos o rectángulos.

Para Ferraty y View ([2]), "una colección de n datos funcionales, $\{\chi_1, \ldots, \chi_n\}$, se obtiene a partir de la observación de n variables funcionales $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ idénticamente distribuidas como $\mathcal{X}$"[4].

[3] Las observaciones podrían corresponderse a las mediciones de una magnitud obtenidas por distintas máquinas u observadores de un mismo individuo o a las mediciones de dicha magnitud de distintos individuos de una población obtenidas por una máquina u observador.

[4] Puede ser una muestra de curvas, representando a individuos, si las variables son independientes; lo que nos permitirá analizar la evolución, o comportamiento dinámico, de la magnitud en

En las situaciones más habituales de soporte unidimensional real, las funciones $\chi(t)$ pertenecerán a un subespacio $\mathcal{V}$ del espacio $\mathcal{H}_\tau = \{f(t)|f \in L^2(\tau)\}$, que es un espacio de Hilbert[5]. En el caso de soporte multidimensional real, $\mathcal{H}_\tau$ será el espacio de funciones de clase C^2 definidas en un rectángulo euclídeo contenido en $\mathbb{R}^K$, con $K \geq 2$. En todos los casos, el espacio $\mathcal{H}_\tau$ tiene dimensión infinita.

Cada curva o dato funcional $\chi(t)$, obtenida a partir del vector de observaciones, puede ser evaluada para cualquier $t \in \tau$, ofreciéndonos información de cómo cambia la magnitud en cada instante del tiempo; además, la evaluación de la curva en los puntos de la rejilla soporte nos ofrecerá un criterio para la reconstrucción y posterior evaluación de la calidad de la misma, como veremos más adelante. Una de las condiciones deseables en las curvas es que existan las dos primeras funciones derivadas, $\chi^{(1)}(t)$ y $\chi^{(2)}(t)$, las cuales nos aportarán información sobre los cambios en velocidad y aceleración, respectivamente, de la magnitud estudiada.

El FDA admite también un enfoque multidimensional sin más que considerar p magnitudes, $X^1, \ldots, X^p$, con $p \geq 2$, de las que se dispone de información sobre el mismo recinto soporte para un conjunto de n individuos. El ajuste o suavizado para cada magnitud determinará las $n \times p$ funciones muestrales, de forma que el conjunto de los datos funcionales estará determinado por una matriz, cuyos elementos vendrán dados por las curvas[6]:

$$\chi_i^j = \{\chi_i^j(t)|\chi_i^j \in L^2(\tau); i = 1, \ldots, n; j = 1, \ldots, p\}$$

La matriz de individuos/variables sería:

$$\chi = \begin{pmatrix} \chi_1^1(t) & \chi_1^2(t) & \cdots & \chi_1^p(t) \\ \chi_2^1(t) & \chi_2^2(t) & \cdots & \chi_2^p(t) \\ \cdots & \cdots & \cdots & \cdots \\ \chi_n^1(t) & \chi_n^2(t) & \cdots & \chi_n^p(t) \end{pmatrix}$$

Desde el punto de vista algebráico, el modelo de datos es equivalente al de una matriz de datos multivariante vectorial, solo que cada individuo vendrá caracterizado, en lugar de por un vector de escalares por un vector de curvas.

estudio, X, en una población de individuos.

[5] $\mathcal{H}_\tau$ es el espacio de las funciones medibles definidas sobre el intervalo soporte unidimensional, $\tau = [t_{min}, t_{max}] \in \mathbb{R}$, es decir, $\int_{t_{min}}^{t_{max}} |f(t)|^2 dt < \infty$.

[6] Para simplificar la notación consideraremos un problema con soporte en un intervalo.

Llegados a este punto, nos gustaría llamar la atención sobre el hecho de que, para una cierta magnitud, cada individuo considerado viene representado por un único elemento: una curva con buenas características analíticas, en lugar de por una serie de valores en un soporte discreto. Piénsese en las ventajas que ello supone, gracias al amplio catálogo de recursos que nos proporciona el análisis funcional, merced a la adaptación de las técnicas vectoriales clásicas. No obstante, el FDA no pretende sustituir, por ejemplo, el análisis clásico de series temporales, sino realizar una propuesta de análisis complementaria. Obviamente no todos son ventajas, pues salvo en casos muy excepcionales donde la representación funcional es exacta, habrá que pagar el peaje de asumir los errores derivados del proceso de conversión a curvas; además habrá que gestionar las cuestiones derivadas de la alta dimensionalidad de los datos y su manejo computacional.

Aunque conceptualmente estamos trabajando en espacios funcionales de dimensión infinita, desde un punto de vista práctico las curvas que se obtienen a partir de la secuencia de observaciones obtenidas sobre la rejilla tendrán una representación computacional discreta. Detallaremos esta cuestión más adelante, en la sección dedicada a la obtención de las curvas.

1.2. Métricas y semimétricas en un espacio funcional

Antes de considerar otras cuestiones más operativas, analizaremos en este apartado las métricas o semimétricas que van a permitir cuantificar el grado de similitud o de dispersión entre diferentes elementos funcionales, como por ejemplo obtener...

- la diferencia entre un par de curvas observadas
- la diferencia entre una curva observada y el promedio del conjunto de curvas del que ésta forma parte
- la dispersión de un conjunto de curvas
- la curva media de una curva observada
- la correlación entre dos curvas

Como puede observarse en la relación anterior, la mayoría de los ítems hacen referencia al cálculo de estadísticos descriptivos uni y bivariantes entre curvas.

La elección de la métrica o semimétrica va a depender del objetivo del análisis concreto que se desee realizar; así, si lo que interesa es la comparación directa entre curvas, la métrica elegida podría ser la inducida por la norma l^2 en el espacio de Hilbert funcional $\mathcal{H}_\tau$, mientras que si lo que se desea es comparar las velocidades de cambio de las curvas o curvaturas, habrá que trabajar con la seminorma definida sobre las primeras derivadas en dicho espacio. En cualquier caso, se necesita tener definido un producto escalar o semiescalar en el espacio considerado, dicho producto define una norma o seminorma y, ésta, la métrica, o semimétrica, inducida. La notación algebraica proporciona un marco general para datos de cualquier naturaleza. Usaremos la caracterización de espacios de Hilbert dada por Aubin en su libro *Applied Functional Analysis* ([3]).

Definición 1.2.1 *Si consideramos dos elementos x e y de un espacio vectorial real V, un producto semiescalar, $\langle .,. \rangle$ en V, es una aplicación definida como:*

$$\begin{array}{rl} \langle .,. \rangle : & V \times V \longrightarrow \mathbb{R} \\ & (x,y) \longrightarrow \langle x,y \rangle \end{array} \tag{1.2.1}$$

que verifica:

I $\langle \sum_{i=1}^n \lambda^i x_i, y \rangle = \sum_{i=1}^n \lambda^i \langle x_i, y \rangle$ *(Linealidad respecto a x)*

II $\langle x, \sum_{j=1}^m \mu^j y_j \rangle = \sum_{j=1}^m \mu^j \langle x, y_j \rangle$ *(Linealidad respecto a y)*

III $\langle x,y \rangle = \langle y,x \rangle$ *(Simetría)*

IV $\langle x,x \rangle \geq 0 \ \forall x \in V$ *(Positividad)*

Llamaremos al par $(V, \langle , \rangle)$, formado por un espacio vectorial y un producto semiescalar, un espacio pre-hilbertiano no separable. Un producto escalar es una forma bilineal simétrica para la cual:

IV' $\forall x \neq 0, \langle x,x \rangle > 0$ *(Definida positiva)*[7]

y llamamos al par $\{V, \langle , \rangle\}$, donde $\langle .,. \rangle$ es un producto escalar, un espacio prehilbertiano.

[7] La condición IV' claramente implica la condición IV.

Un producto escalar define una norma y consecuentemente una distancia en el espacio V.

Proposición 1.2.1 *Si $\langle .,. \rangle$ en V es un producto semiescalar, entonces se verifica la igualdad de Cauchy-Schwarz:*

$$|\langle x, y \rangle| \leq \sqrt{\langle x, x \rangle}\sqrt{\langle y, y \rangle} \tag{1.2.2}$$

Proposición 1.2.2 *Si $\langle x, y \rangle$ es un producto semiescalar, entonces $||x|| = \sqrt{\langle x, x \rangle}$ es una seminorma y será una norma si $\langle x, y \rangle$ es un producto escalar.*

En efecto, se puede comprobar fácilmente que se verifica la desigualdad triangular:

$$||x + y|| \leq ||x|| + ||y||, \tag{1.2.3}$$

por lo que un espacio prehilbertiano es un espacio normado y por lo tanto un espacio métrico para la distancia $d(x, y) = ||x - y||$.

Definición 1.2.2 *Decimos que un espacio prehilbertiano es un espacio de Hilbert si es completo para la distancia asociada.*

Observación 1.2.1 *Una semimétrica es una métrica, excepto que $d(x(t), y(t)) = 0$ no implica que $x(t) = y(t), \forall t \in \tau$ [4].*

En la siguiente sección particularizaremos estas definiciones para el caso funcional dependiendo de un soporte unidimensional.

1.2.1. Métrica en un espacio de Hilbert funcional

Definición 1.2.3 *Definimos el producto escalar euclidiano entre dos funciones, x e y, pertenecientes al espacio $L^2(\tau)$ de funciones de cuadrado integrable en $\tau = [0, T]$, como:*

$$\langle x, y \rangle = \int_0^T x(t)y(t)dt, \tag{1.2.4}$$

dicho producto verifica las propiedades generales de un producto escalar.

Definición 1.2.4 *A partir del producto escalar anterior, la norma* $||.||$ *de una función* $x \in L^2(\tau)$ *se define como:*

$$||x|| = \sqrt{\langle x, x \rangle} = \sqrt{\int_0^T x(t)^2 dt}, \tag{1.2.5}$$

que verifica las propiedades:

1. $||x|| \geq 0$ y $||x|| = 0$ si y sólo si $x(t) = 0(t)$
2. $||ax|| = |a|\ ||x||$, para todo $a \in \mathbb{R}$
3. $|\langle x, y \rangle| \leq ||x||\ ||y|| = \sqrt{\langle x, x \rangle \langle y, y \rangle}$ (Desigualdad de Cauchy-Schwarz)
4. $||x + y|| \leq ||x|| + ||y||$ (Desigualdad triangular)

De la *Desigualdad de Cauchy-Schwarz* se deduce trivialmente la *Desigualdad del coseno*:

$$-1 \leq \frac{\langle x, y \rangle}{||x||\ ||y||} \leq 1 \tag{1.2.6}$$

que permite extender el concepto geométrico de ángulo al caso funcional. Si llamamos θ al "*ángulo funcional*", se tendría que:

$$cos\theta = \frac{\langle x, y \rangle}{||x||\ ||y||} = \frac{\int_0^T x(t)y(t)dt}{\sqrt{\int_0^T x(t)^2 dt}\sqrt{\int_0^T y(t)^2 dt}} \tag{1.2.7}$$

Puesto que en el caso vectorial, el coseno del ángulo entre dos vectores es un indicador normalizado del grado de relación lineal entre dichos vectores, podríamos interpretar que, en el caso funcional, el *coseno funcional*, obtenido a partir de la expresión (1.2.7), medirá el grado de relación funcional normalizado. De esta manera, un coseno próximo a cero indicaría inexistencia de relación entre las funciones, lo que implicaría independencia entre ellas, mientras que cuando dicho coseno se acerca a -1 o a 1, indicaría una fuerte relación inversa o directa, respectivamente. En cualquier caso, seguiremos usando el concepto de *ortogonalidad* entre funciones cuando su producto escalar sea nulo.

Es fácil comprobar que la "Bilinealidad" del producto escalar funcional, implica que la relación entre dos funciones, medidas a partir de su coseno funcional, es invariante por cambios de escala o traslaciones.

Definición 1.2.5 *Definimos la distancia l_2, o euclídea, entre las funciones $x(t)$ e $y(t)$, como:*

$$d(x,y) = ||x-y|| = \sqrt{\langle x-y, x-y\rangle} = \sqrt{\int_0^T (x-y)^2(t)dt} \tag{1.2.8}$$

Proposición 1.2.3 *$L^2(\tau)$ con la distancia inducida por la norma euclídea (1.2.8) es un espacio de Hilbert.*

1.2.2. Semimétricas en un espacio de Hilbert funcional

Podríamos denominar a la métrica euclidiana descrita en el apartado anterior como "natural", ya que nos da información del grado de proximidad entre dos curvas que representan a sendos individuos de una población; sin embargo, en ocasiones podemos estar interesados en analizar el parecido de los individuos desde otras perspectivas distintas, lo que nos lleva a considerar semimétricas. Ferraty y Vieu proponen en el Capítulo 3 de su libro ([2]) tres semimétricas para datos funcionales, una basada en componentes principales funcionales (FPC), otra en derivadas de orden $p \geq 1$ y una última en técnicas de mínimos cuadrados parciales (PLS). Pasamos a detallar las dos primeras semimétricas, aplicables a un problema general, mientras que el uso de la semimétrica PLS en problemas funcionales se recomienda en situaciones en la que se dispone de un factor que indica la pertenencia de cada curva a una determinada clase, conocidos como MPLSR (Multivariate Partial Least-Square Regression). La estrategia subyacente a los modelos MPLSR es la de maximizar la covarianza dentro de cada clase. En este tipo de situaciones los modelos MPLSR obtienen mejores resultados que los basados en componentes principales y son especialmente útiles cuando el objetivo del análisis funcional es la clasificación o la discriminación de las curvas.

Semimétrica obtenida a partir de componentes principales funcionales.

Supongamos que se dispone de una muestra de curvas, $\chi_1, \ldots, \chi_n$, idénticamente distribuidas, extraídas de una variable aleatoria funcional, $\mathcal{X} = \{\mathcal{X}(t) : t \in \tau \subset$

$\mathbb{R}\}$. En ocasiones, al objeto de representar las curvas se usa una proyección en el espacio generado por las componentes principales funcionales (FPCA)[8]. Supongamos que la base ortonormal de componentes principales asociada a la variable aleatoria funcional $\mathcal{X}$ viene dada por $\{v_k(t)\}_{k\in\mathbb{N}}$. Cualquier realización χ_i de $\mathcal{X}$, que verifica que $E\int \chi_i^2(t)dt < \infty$, tendrá una expansión en el espacio determinado por la base de componentes principales dada por Dauxois at al. [5]:

$$\chi_i(t) = \sum_{k=1}^{\infty} \left(\int_0^T \chi_i(t) v_k(t) dt \right) v_k(t) \tag{1.2.9}$$

En general, unas pocas componentes recogen un porcentaje muy alto de la inercia de la muestra, $(\chi_1, \ldots, \chi_n)$, por lo que podemos obtener una estimación de cada χ_i, proyectando sobre el espacio generado por las p primeras componentes:

$$\tilde{\chi}_i(t) = \sum_{k=1}^{p} \left(\int_0^T \chi_i(t) v_k(t) dt \right) v_k(t) \tag{1.2.10}$$

Definición 1.2.6 *A partir de la norma l^2 y considerando la proyección anterior, podemos definir la seminorma basada en FPCA como:*

$$||\chi_i||_{FPCA} = \sqrt{\int (\tilde{\chi}_i(t))^2 dt} = \sqrt{\sum_{k=1}^{p} \left(\int \chi_i(t) v_k(t) dt \right)^2} \tag{1.2.11}$$

Definición 1.2.7 *Dadas dos funciones, χ_i y χ_i', la disimilitud entre ellas, basada en la semimétrica inducida por la seminorma anterior, vendría dada por:*

$$d_{FPCA}(\chi_i, \chi_i') = ||\chi_i - \chi_i'||_{FPCA} = \sqrt{\sum_{k=1}^{p} \left(\int [\chi_i(t) - \chi_i'(t)] v_k(t) dt \right)^2} \tag{1.2.12}$$

Observación 1.2.2 *En la práctica no se dispone de la base ortonormal FPCA de $\mathcal{X}$, sino de una estimación de la misma, $\{\tilde{v_k}(t)\}_{k\in\mathbb{N}}$, obtenida a partir de la diagonalización de la muestra de curvas obtenida. Se demuestra que las autofunciones $\tilde{v_k}$ son estimadores consistentes de las v_k*

[8] Se detallará en una posterior sección cómo se obtienen e interpretan dichas componentes.

Semimétrica basada en derivadas.

Si se desea analizar el comportamiento de las curvas a través de sus curvaturas o derivadas de orden q, la semimétrica aconsejable es la que viene dada por la norma l^2 sobre las derivadas de orden $q > 1$[9].

Definición 1.2.8 *Dadas dos curvas, χ_i y χ_i', la semimétrica basada en la derivada de orden q, viene dada por:*

$$d^q_{deriv}(\chi_i, \chi_i') = \sqrt{\sum_{k=1}^{p} \left(\int [\chi_i^{(q)}(t) - \chi_i'^{(q)}(t)] v_k(t) dt \right)^2} \tag{1.2.13}$$

Observación 1.2.3 *Dado que la obtención de derivadas sucesivas es muy sensible computacionalmente hablando, es recomendable usar una base de funciones que de estabilidad a dicho problema computacional. Una base de Splines para obtener la estimación suavizada de las curvas consigue buenos resultados en este sentido.*

1.3. Estadísticos funcionales univariantes y bivariantes

Detallaremos en esta sección como se definen las características principales de los estadísticos funcionales básicos que involucran a una o varias curvas, mediante la extensión de las principales medidas de representación, escala y dependencia del caso vectorial al caso funcional. Distinguiremos entre `estadísticos sobre una función`, `estadísticos sobre una muestra de una función aleatoria` y `estadísticos sobre muestras de dos o más funciones aleatorias`. Seguiremos el esquema propuesto por Navarro Pérez ([6]).

Supongamos que disponemos de representaciones funcionales de uno o más individuos para una cierta magnitud, X, definidas en un intervalo $\tau \equiv [t_{min}, t_{max}]$ de amplitud $T = t_{max} - t_{min}$, por lo que podemos considerar sin pérdida de generalidad que el intervalo soporte es $[0, T]$. Comenzaremos proyectando los medidas univariantes y bivariantes de la Estadística clásica al caso funcional, aunque más que de una generalización, se trata de establecer una metodología de trabajo en la que los

[9] Para q=0 se obtendría la métrica euclídea.

elementos implicados podrían ser vectores, funciones, etc., que, de alguna manera, miden el estado de un conjunto de individuos en relación a la(s) magnitud(es) considerada(s).

1.-Estadísticos sobre una función

Supongamos que tenemos una función x, calcularemos su media, varianza y la covarianza con otra función y.

Definición 1.3.1 *Suponiendo que* $\mathbf{1}(t) = 1,\ \forall t \in [0,T]$, *definimos la* `media`, *o* `valor medio`, *de la función* x *en el intervalo* $[0,T]$ *como:*

$$\overline{x} = T^{-1}\langle x, \mathbf{1}\rangle = T^{-1}\int_0^T x(t)\mathbf{1}(t)dt = T^{-1}\int_0^T x(t)dt$$

es decir, el área normalizada por el tamaño del intervalo, encerrada en x *entre* 0 *y* T.

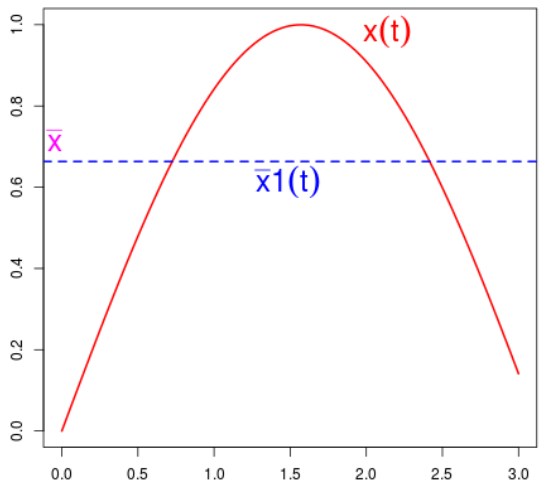


Figura 1.1: Media y función media de una función $x(t)$

La `función valor medio de` x es $f_{\bar{x}}(t) = \overline{x},\ \forall t \in [0,T]$ o de forma equivalente $f_{\bar{x}}(t) = \overline{x}\mathbf{1}(t)$ (Fig. 1.1). Se verifica que:

$$\int_0^T x(t)dt = \int_o^T f_{\overline{x}}dt = T\overline{x};$$

es decir, el área que encierra x es la misma que la que encierra la `función valor medio` (Fig 1.2).

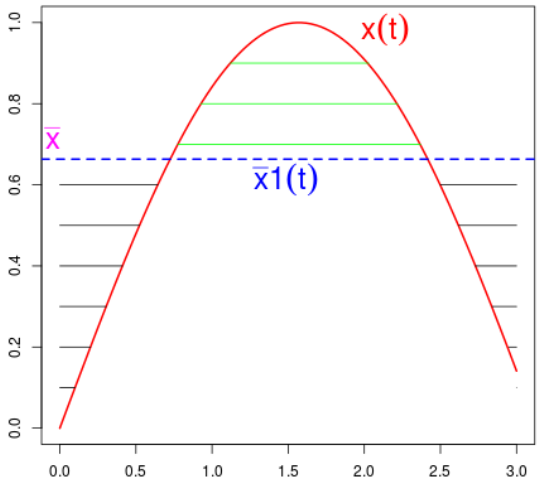


Figura 1.2: Propiedad de la función media de x

Definición 1.3.2 *Definimos la* ***varianza*** *de x como*

$$S_x^2 = T^{-1}\langle x - \overline{x}\mathbf{1}, x - \overline{x}\mathbf{1}\rangle = T^{-1}\int_0^T (x(t) - \overline{x}\mathbf{1}(t))^2 dt$$

La varianza representa el cuadrado de la variación media de los valores de la función respecto a su valor medio. Una varianza alta refleja la existencia de intervalos donde la función x se aleja mucho de su media. La raiz cuadrada de la varianza es la desviación típica de la curva x, S_x.

Definición 1.3.3 *Definimos la covarianza entre las funciones x e y como:*

$$\begin{aligned} S_{xy} = & \ T^{-1}\langle x - \overline{x}\mathbf{1}, y - \overline{y}\mathbf{1}\rangle = \\ = & \ T^{-1}\int_0^T (x(t) - \overline{x(t)}\mathbf{1}(t))(y(t) - \overline{y(t)}\mathbf{1}(t))dt \end{aligned}$$

Normalizando la covarianza se obtendría la correlación:

$$r_{xy} = \frac{S_{xy}}{S_x S_y},$$

que tomará valores en $[-1, 1]$. En ambos casos, la integral se puede descomponer en tantos intervalos como cortes tengan las funciones x e y con sus respectivas funciones medias. Dentro de cada intervalo las funciones diferencias pueden ser las dos del mismo signo, lo que aportará sumandos positivos, o de distinto signo, sumandos negativos; la suma final, interpretada en términos del coeficiente de correlación, nos permitirá evaluar la "relación lineal" entre las funciones.

El modelo funcional sería una generalización directa de uno puntual en el que las muestras se corresponderían con mediciones ordenadas de una o varias características o variables en los mismos instantes del tiempo. Cuando el producto escalar considerado sea general, la matriz de ponderaciones de las covarianzas sería la de los individuos o tiempos, no el de las variables.

2.-Estadísticos de la muestra de una función aleatoria

Supongamos ahora que se dispone de una muestra de n realizaciones, $\chi_1(t), \chi_2(t), ..., \chi_n(t)$, de una variable aleatoria funcional, $\mathcal{X}$, definida en un intervalo $\tau \equiv [0, T]$. El modelo de datos se recoge en el Cuadro 1.1.

Punto soporte	*Individuo* 1	*Individuo* 2	...	*Individuo n*
t_1	$x_1(t_1)$	$x_2(t_1)$	...	$x_n(t_1)$
t_2	$x_1(t_2)$	$x_2(t_2)$	...	$x_n(t_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_m	$x_1(t_m)$	$x_2(t_m)$	...	$x_n(t_m)$
Funciones suavizadas muestrales	$\chi_1(t)$	$\chi_2(t)$		$\chi_n(t)$

Cuadro 1.1: Modelo de datos de dos muestras de funciones aleatorias

Definición 1.3.4 *Se define la* ***función media muestral*** *como*

$$\overline{\chi}(t) = \frac{1}{n}\sum_{i=1}^{n}\chi_i(t),\ \forall t \in \tau$$

Es decir, la función media muestral es la media, punto a punto, dentro del intervalo $[0,T]$ de las n funciones muestrales. La media muestral será el estimador ideal de la media poblacional, $\tilde{\mu}_{\mathcal{X}} = \overline{\chi}(t)$; además, $\overline{\chi}(t)$ es de clase C^2.

Definición 1.3.5 *En las mismas condiciones anteriores, definimos la* ***función cuasi-varianza muestral*** *de $\mathcal{X}$, o simplemente* ***varianza muestral****, como:*

$$S^2_{\chi}(t) = \frac{1}{n-1}\sum_{i=1}^{n}(\chi_i(t) - \overline{\chi}(t))^2$$

Para cada t, la función varianza muestral evalua las desviaciones al cuadrado entre las $\chi_i(t)$ y su función media $\overline{\chi}(t)$. $S^2_{\chi}(t)$ es continua y, al igual que para la media, es el estimador ideal de la varianza poblacional, $\tilde{\sigma}^2_{\mathcal{X}} = S^2_{\chi}(t)$.

Definición 1.3.6 *Definimos la* ***función covarianza intra-puntos muestral*** *de la variable $\mathcal{X}$ entre dos puntos genéricos t y s del intervalo $[0,T]$ como:*

$$Cov_{\chi}(t,s) = \frac{1}{n-1}\sum_{i=1}^{n}(\chi_i(t) - \overline{\chi}(t))(\chi_i(s) - \overline{\chi}(s)),\ (t,s) \in [0,T]\times[0,T];$$

obviamente se tiene que:

$$Cov_{\chi}(t,t) = S^2_{\chi}(t)$$

Observación 1.3.1 *La covarianza intra-puntos nos permitirá medir el promedio de la covarianza, para la muestra considerada, para cada par de puntos dentro del intervalo soporte. Por otra parte,*

$$S = \{(t,s,Cov_{\chi}(t,s)),\ (t,s) \in [0,T]\times[0,T]\}$$

nos da la superficie que representa las auto-covarianzas de las funciones suavizadas muestrales.

Definición 1.3.7 *Para poder interpretar el tamaño de la relación entre pares de puntos definimos la* ***función correlación lineal intra-puntos muestral:***

$$Cor_\chi(t,s) = \frac{Cov_\chi(t,s)}{\sqrt{S^2_\chi(t)S^2_\chi(s)}}$$

La correlación-intrapuntos muestral es el estimador ideal de la correlación intra-puntos poblacional, $\tilde{\rho}_\chi(t_r,t_s) = Cor_\chi(t_r,t_s)$.

Observación 1.3.2 *La correlación intra-puntos nos permitirá medir el promedio de correlaciones, para la muestra considerada, para cada par de puntos dentro del intervalo soporte. Es evidente que:*

$$Cor_\chi(t,t) = 1.$$

Por otra parte,

$$R = \{(t,s,Cor_\chi(t,s)),\ (t,s) \in [0,T]\times[0,T]\}$$

nos da la superficie que integra las auto-correlaciones de las funciones suavizadas muestrales.

3.-Covarianza y correlación cruzadas.

Supongamos ahora que nos interesa analizar la relación entre dos características, X_1 e X_2, de una población, para lo que se dispone de una muestra de tamaño n de cada variable aleatoria funcional, $\mathcal{X}_1$ y $\mathcal{X}_2$, dadas por $[\chi_{1,1}(t),\chi_{1,2}(t),,...,\chi_{1,n}(t)]$ y $[\chi_{2,1}(s),\chi_{2,2}(s),...,\chi_{2,n}(s)]$[10], con t y s tomando valores en un soporte $[0,T]$. El modelo de datos sería el que viene dado en el Cuadro 1.2. Podemos considerar que cada individuo está representado por una función dada por:

$$\phi_i(t,s) = (\chi_{1,i}(t),\chi_{2,i}(s)),\ \forall i = 1,...,n$$

o, lo que es lo mismo, que $(\phi_1(t,s),\phi_2(t,s),...,\phi_n(t,s))$ es una muestra funcional de tamaño n de la variable aleatoria bidimensional $(\mathcal{X}_1,\mathcal{X}_2)$ en el soporte $[0,T]\times[0,T]$. Dicha muestra se ha obtenido a partir del suavizado de dos conjuntos de observaciones realizadas en m puntos del intervalo $[0,T]$.

[10] El primer subíndice identifica la variable y el segundo subíndice al individuo.

Punto soporte	*Individuo* 1		*Individuo* 2		...	*Individuo n*	
t_1	$x_{1,1}(t_1)$	$x_{2,1}(t_1)$	$x_{1,2}(t_1)$	$x_{2,2}(t_1)$	...	$x_{1,n}(t_1)$	$x_{2,n}(t_1)$
t_2	$x_{1,1}(t_2)$	$x_{2,1}(t_2)$	$x_{1,2}(t_2)$	$x_{2,2}(t_2)$	...	$x_{1,n}(t_2)$	$x_{2,n}(t_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
t_m	$x_{1,1}(t_m)$	$x_{2,1}(t_m)$	$x_{1,2}(t_m)$	$x_{2,2}(t_m)$	...	$x_{1,n}(t_m)$	$x_{2,n}(t_m)$
Procesos de suavizado	$\chi_{1,1}(t)$	$\chi_{2,1}(s)$	$\chi_{1,2}(t)$	$\chi_{2,2}(s)$		$\chi_{1,n}(t)$	$\chi_{2,n}(s)$
Funciones muestrales	$\phi_1(t,s)$		$\phi_2(t,s)$		...	$\phi_n(t,s)$	

Cuadro 1.2: Modelo de datos de dos muestras de funciones aleatorias

Definición 1.3.8 *En las condiciones descritas, se define la* ***función covarianza muestral cruzada*** *en los puntos* $t, s \in [0,T]$ *según Ramsay ([1]) como:*

$$Cov_{\chi_1\chi_2}(t,s) = \frac{1}{n-1}\sum_{i=1}^{n}(\chi_{1,i}(t) - \overline{\chi_1}(t))(\chi_{2,i}(s) - \overline{\chi_2}(s))$$

La función covarianza cruzada relaciona dos variables aleatorias en base a un diferimiento temporal, y lo hace como un promedio para el conjunto de individuos, del producto de las desviaciones entre las funciones y sus medias en cada variable.

Definición 1.3.9 *Normalizando la covarianza anterior se obtiene la* ***función de correlación muestral cruzada*** *en los puntos* $t, s \in [0,T]$ *como:*

$$r_{\chi_1\chi_2}(t,s) = \frac{Cov_{\chi_1\chi_2}(t,s)}{\sqrt{S^2_{\chi_1}(t)S^2_{\chi_2}(s)}} \tag{1.3.14}$$

1.4. Estimación de las curvas. Suavizado

La primera cuestión operativa que se plantea ante un problema funcional, es la conversión de las series de observaciones de una magnitud para un conjunto de individuos, $\{\mathbf{x}_i = (x_i(t_1),\ x_i(t_2),\ ..., x_i(t_m)),\ i = 1, ..., n\}$[11], en funciones, $\{\chi_i(t), i = 1, ..., n\}$. Si se dispone de información fidedigna de la relación funcional entre la magnitud y la variable soporte, dicha función podría ser sometida a evaluación.

[11] Supondremos que todos los individuos han sido observados en la misma rejilla.

Ramsay y Silverman en la Sección 1.3 de su libro [1] proponen para aproximar las series de temperaturas, dado su carácter sinusoidal y periódico en cuatro poblaciones en Canadá, el operador:

$$Temp_i(t) = c_{i1} + c_{i2}\sin(\pi t/6) + c_{i3}\cos(\pi t/6),\ i = 1, 2, 3, 4$$

Por otro lado, si se tiene garantía de que las mediciones son exactas, la obtención de las funciones se haría, en principio, mediante un proceso de interpolación, imponiendo la condición:

$$\chi_i(t_j) = x_i(t_j),\ j = 1, ..., m \tag{1.4.15}$$

No obstante, si las funciones obtenidas al obligar que se den las condiciones (1.4.15) son demasiados `rugosas`, o las mediciones no son exactas, lo que ocurre en la mayoría de las situaciones, es recomendable usar un `método de suavizado` para obtener las curvas. En cualquier caso, se trata de reconstruir la forma funcional para el conjunto de individuos a partir de sus observaciones discretas en un conjunto soporte.

Para Ramsay ([1]) "El término funcional en referencia a los datos observados se refiere a la estructura intrínseca de los datos más que a su forma explícita. En la práctica, los datos funcionales generalmente se observan y registran de forma discreta como m pares (t_j, x_j), siendo x_j una instantánea de la función en el tiempo t_j, posiblemente afectada por un error de medición. El tiempo es tan a menudo el soporte continuo sobre el que se registran los datos funcionales, que podemos caer en el hábito de referirnos a t_j como tal, pero sin duda pueden estar involucrados otros continuos, como la posición espacial, la frecuencia, el peso, etc."

La idea de suavizado tiene que ver con que los cambios que se producen entre puntos próximos del soporte son suficientemente pequeños, de forma que las funciones admitan al menos primeras y segundas derivadas. De aquí, que el proceso de suavizado no se limita únicamente al objetivo de obtención de las funciones, ofreciendo muchas más posibilidades al análisis funcional de las que se tendrían si nos moviéramos en el plano multivariante.

Por otra parte, como se ha comentado con anterioridad, los datos observados vienen afectados por un error de medición que podemos relacionar con una perturbación o ruido, de manera que uno de los objetivos del suavizado es el filtrado del ruido de la mejor manera posible. Además, suponiendo que se pudiera diseñar el muestreo para disponer de tantas observaciones como fueran necesarias, se tendría que contar con un mayor número de mediciones en aquellos entornos del soporte donde la curvatura, dada por la segunda derivada, fuera mayor.

Las técnicas de suavizado pueden clasificarse en dos grupos: suavizado por bases de funciones y suavizado por ponderaciones locales basado en kernel. Describiremos en el apartado siguiente las dos metodologías.

1.4.1. Método basado en bases de funciones

Una base de funciones, $\{\phi_k,\, k \geq 0\}$, definida en un soporte $[0,T]$ es un conjunto de funciones ortogonales, de modo que cualquier función del espacio funcional puede representarse como una combinación lineal de las funciones base. Dada una curva $\chi(t)$, conceptualmente se trata de encontrar coeficientes $c_0, c_1, \ldots, c_k, \ldots$, de forma que:

$$\chi(t) = c_0\phi_0(t) + \ldots + c_k\phi_k(t) + \ldots \tag{1.4.16}$$

Si consideramos el producto escalar de $\chi(t)$ por un elemento genérico de la base, se tendría, puestos que estos son ortogonales:

$$\begin{aligned} \langle \chi, \phi_k \rangle &= \int_0^T \chi(t)\phi_k(t)dt = \int_0^T (c_0\phi_0(t) + \ldots + c_k\phi_k(t) + \ldots)\phi_k(t)dt = \\ &= \int_0^T c_k\phi_k(t)^2 dt = c_k \int_0^T \phi_k(t)^2 dt \end{aligned}$$

de donde:

$$c_k = \frac{\langle \chi, \phi_k \rangle}{\int_0^T \phi_k(t)^2 dx} = \frac{\langle \chi, \phi_k \rangle}{||\phi_k||^2}$$

y sustituyendo:

$$\chi(t) = \sum_{k=0}^{\infty} c_k\phi_k(t) = \sum_{k=0}^{\infty} \frac{\langle \chi, \phi_k \rangle}{||\phi_k||^2}\phi_k(t)$$

Si la base es ortonormal, se tendría que:

$$\chi(t) = \sum_{k=0}^{\infty} c_k\phi_k(t) = \sum_{k=0}^{\infty} \langle \chi, \phi_k \rangle \phi_k(t)$$

Dado que el espacio funcional es de dimensión infinita, la propuesta dada en (1.4.16) no deja de ser teórica, de forma que para poder abordar un problema real necesita-

mos truncar dicha expresión para quedarnos con un número finito de términos. La determinación del número, K, de funciones base para proyectar o expandir la función de suavizado va a depender del criterio que se elija; en la literatura se ofrece una gran cantidad de opciones, siendo una de las más exitosas el Criterio de Validación Cruzada Generalizado (GCVC), que detallaremos más adelante. Así, la expansión empírica de la muestra de funciones vendrá dada por la expresión:

$$\chi_K(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \Phi(t)\mathbf{c} \tag{1.4.17}$$

El grado de aproximación deseado para la función $\chi(t)$, dado por (1.4.17), a su vector observado, $\mathbf{x}$, condiciona el número de elementos de la base de funciones. Suponiendo que la matriz $\Phi = \{\phi_k(t_j),\ j = 1, ..., m,\ k = 1, ..., K\}$, de orden $m \times K$, que evalua las funciones base en la rejilla soporte, es de rango completo, se puede obtener una representación exacta, o interpolación, con una base de tamaño $K = m$, de forma que se pueden obtener los coeficientes c_k de manera que $\chi(t_j) = x(t_j)$ ([1]). No obstante, no buscamos un ajuste de las observaciones, sino un suavizado de las mismas que posea buenas propiedades, por lo que el número de funciones bases será generalmente una pequeña fracción de m.

Una vez determinado K, el criterio habitual para la obtención de los coeficientes de expansión, c_k, es el de mínimos cuadrados, lo que nos lleva a resolver el problema de optimización:

$$\min_{\{c_k\}} \sum_{j=1}^{m} [x(t_j) - \chi_K(t_j)]^2 = \min_{\{c_k\}} \sum_{j=1}^{m} [x(t_j) - \sum_{k=1}^{K} c_k \phi_k(t_j)]^2 = \min_{\mathbf{c}} (\mathbf{x} - \mathbf{c}\Phi)'(\mathbf{x} - \mathbf{c}\Phi) \tag{1.4.18}$$

La solución al problema anterior es:

$$\mathbf{c} = (\Phi'\Phi)^{-1}\Phi'\mathbf{x} \tag{1.4.19}$$

En la rejilla soporte, la suavización de $\mathbf{x}$ viene dada por la proyección ortogonal:

$$\mathbf{S}\mathbf{x} = \Phi(\Phi'\Phi)^{-1}\Phi'\mathbf{x} = \Phi\mathbf{c} \tag{1.4.20}$$

donde la matriz de suavizado, $\mathbf{S} = \Phi(\Phi'\Phi)^{-1}\Phi'$, denominada `matriz sombrero`, es simétrica e idempotente.

Es importante ver que los coeficientes c_k caracterizan al vector de observaciones en la base, pudiendo usarse al conjunto de coeficientes como representante del mismo con el objetivo, por ejemplo, de clasificar un conjunto de curvas usando técnicas vectoriales.

1.4.2. Base de Fourier

Una de las base más utilizadas en análisis funcional es la base de Fourier, especialmente recomendable en situaciones en que los datos tengan un comportamiento periódico. La base ortonormal de Fourier en $[0, T]$ viene dada por:

$$\{\frac{1}{\sqrt{T}}, \frac{\cos(\frac{2\pi t}{T})}{\sqrt{T/2}}, \frac{\cos(\frac{4\pi t}{T})}{\sqrt{T/2}}, \ldots, \frac{\text{sen}(\frac{2\pi t}{T})}{\sqrt{T/2}}, \frac{\text{sen}(\frac{4\pi t}{T})}{\sqrt{T/2}}, \ldots\}$$

A partir de dicha base se obtiene el desarrollo de Fourier, truncado para un cierto K, de la función como:

$$\chi(t) = \frac{c_0}{T} + \frac{2}{T}\sum_{k=1}^{K}[c_k \cos(\frac{2k\pi t}{T}) + c'_k \,\text{sen}(\frac{2k\pi t}{T})]$$

1.4.3. Base de B-Splines

Si los datos no presentan periodicidad, la base de suavizado más habitual es la de B-Splines ([7]). Una base B-Splines de grado $p \geq 0$, es una colección de funciones continuas no negativas, cada una de ellas formada por un conjunto de $p+1$ trozos de polinomios de grado p conectados entre sí (cada polinomio tiene $p+1$ coeficientes a estimar). La base recubre un intervalo que contiene al Soporte, que se ha particionado a través de un conjunto de S nodos. Si la distancia entre nodos es constante los B-Spline son iguales y la distancia entre dos de ellos consecutivos es constante. La Figura 1.3 representa bases de Splines de grados 0 a 3.

Los B-Splines de grado p poseen las siguientes propiedades:

- Consisten en $p+1$ trozos de polinomios de grado p que se unen en p nodos internos.

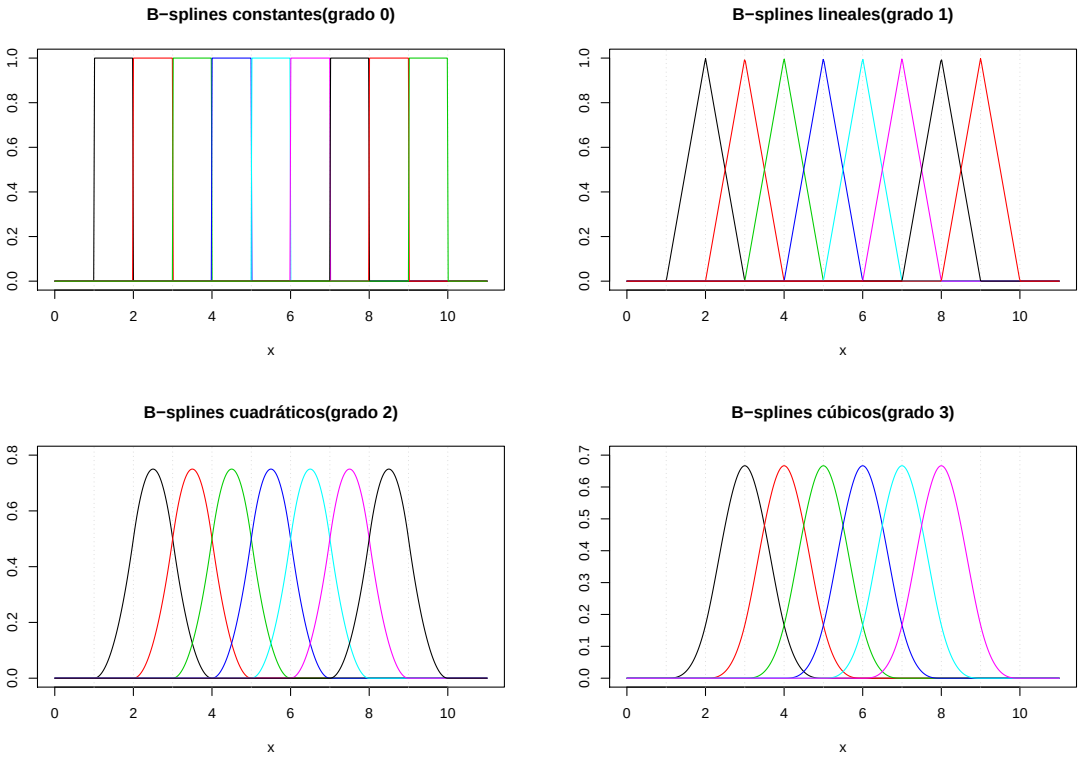


Figura 1.3: Bases de splines de grados 0, 1, 2 y 3

- Las derivadas hasta el orden $p-1$ son continuas en los puntos de unión.
- El B-spline es positivo en el dominio expandido por $p+2$ nodos y cero en el resto.
- Salvo en los extremos, se solapa con $2p$ trozos de polinomios de sus vecinos.
- Hay $p+1$ B-splines no nulos en cada $t \in [0, T]$.
- Para construir un B-spline de grado p son necesarios $p+2$ nodos.

Las bases B-Spline tienen buenas propiedades, ya que no padecen de los efectos frontera y tienen un buen comportamiento local. Para la obtención de los elementos de la base, consideraremos una partición del intervalo soporte determinado por un conjunto de K nodos, $\{t_1, t_2, ..., t_K\}$ y ampliaremos dicha partición para obtener los B-Splines cúbicos: $\{t_{-2}, t_{-1}, t_0, t_1, ..., t_K, t_{K+1}, t_{K+2}, t_{K+3}\}$. La base de B-Splines

cúbicos en esta partición de nodos se obtiene de forma recursiva mediante la fórmula de De Boor:

$$b_{k,0}(t) = \begin{cases} 1, & \text{si } t_{k-2} \leq t \leq t_{k-1} \\ 0, & \text{en el resto} \end{cases}, \quad k = 0, 1, ..., K+4 \tag{1.4.21}$$

$$b_{k,1}(t) = \frac{t - t_{k-2}}{t_{k-1} - t_{k-2}} b_{k,0}(t) + \frac{t_k - t}{t_k - t_{k-1}} b_{k+1,0}(t), \quad k = 0, 1, ..., K+3 \tag{1.4.22}$$

$$b_{k,2}(t) = \frac{t - t_{k-2}}{t_k - t_{k-2}} b_{k,1}(t) + \frac{t_{k+1} - t}{t_{k+1} - t_{k-1}} b_{k+1,1}(t), \quad k = 0, 1, ..., K+2 \tag{1.4.23}$$

$$b_{k,3}(t) = \frac{t - t_{k-2}}{t_{k+1} - t_{k-2}} b_{k,2}(t) + \frac{t_{k+2} - t}{t_{k+2} - t_{k-1}} b_{k+1,2}(t), \quad k = 0, 1, ..., K+1 \tag{1.4.24}$$

La expansión de nuestras observaciones en la base de B-Splines cúbicos obtenida vendría dada por:

$$\chi(t) = \sum_{k=0}^{K+1} c_k B_{k,3}(t)$$

donde $B_{k,3}(t)$ denota el valor del k-ésimo B-spline de grado 3 en el punto t.

1.4.4. Splines de penalización: P-Splines

La solución del problema de optimización 1.4.18, dada por 1.4.19, es óptima desde el punto de vista del sesgo, aunque no desde la óptica de la varianza. El sesgo está relacionado con la calidad del ajuste, mientras que la varianza lo está con las propiedades funcionales de la expansión. La reducción del sesgo y la varianza son fundamentales en cualquier problema de estimación, aunque en la situación que nos ocupa tienen carácter contrapuesto, y la reducción de uno de ellos conlleva el aumento del otro. Así, un suavizado con alto nivel de ajuste será muy rugoso,

mientras que, por el contario, la disminución de la varianza irá en detrimento de la calidad del ajuste.

Para solucionar este problema, O'Sullivan ([8]) introduce una corrección en el problema de optimización 1.4.18 que penaliza la segunda derivada de la expansión, formulando el problema como:

$$\min_{\mathbf{c}}(\mathbf{x}-\mathbf{c}\Phi)'(\mathbf{x}-\mathbf{c}\Phi)+\lambda\int_0^T(\Phi''\mathbf{c})^2dt \tag{1.4.25}$$

siendo λ el parámetro de penalización o suavizado. La expansión del suavizado vendrá dada por una función $\chi_{K,\lambda}(t)$. Los P-splines fueron introducidos por Eilers y Marx (1996) ([9]). Según Aguilera ([10]), las razones para usar P-splines son:

- Son splines de rango bajo, es decir, el tamaño de la base utilizada es mucho menor que la dimensión de los datos. Esto contrasta con lo que ocurre con los splines de suavizado, donde hay tantos nodos como datos, lo que provoca que haya que trabajar con matrices de grandes dimensiones. En el caso de los P-splines, el número de nodos no supera los 40, con lo cual son computacionalmente eficientes, sobre todo si se trabaja con gran cantidad de datos.

- La introducción de penalizaciones relaja la importancia de la elección del número y la localización de los nodos ([11]).

- La correspondencia entre los P-splines y el BLUP (mejor predictor lineal e insesgado) en un modelo mixto permite, en algunos casos, utilizar la metodología existente en el campo de los modelos mixtos y el uso de software estadístico, como la librería `nlme`, en S-PLUS y R.

Estimación del parámetro de suavizado

Para la obtención de λ se usa normalmente el modo de validación cruzada generalizada (GCV) ([12]). El método GCV obtiene el valor de λ optimizando:

$$\min_{\lambda} GCV = m^{-1}\sum_{j=1}^{m}\left(\frac{x(t_j)-\chi_{K,\lambda}(t_j)}{1-Traza((S)/m}\right)^2 \tag{1.4.26}$$

1.4.5. Base de Wavelets

Una base Wavelet en el espacio L^2 se construye a partir de una función ψ, llamada `wavelet madre`, definida en $(-\infty, \infty)$, a partir de traslaciones y dilataciones de ψ. Los Wavelets son una extensión del análisis de Fourier. "El objetivo del análisis con wavelets es convertir una señal en números–coeficientes– que pueden ser manipulados, almacenados, transmitidos, analizados o usados para reconstruir la señal original" ([13]). Como afirma Barbara Burke en su libro ([13]), una Wavelet es un microscopio matemático, en el sentido de que puede fijar la atención en cualquier detalle de la información que se desea analizar, simplemente considerando la translación y la dilatación adecuada. Por ello, los suavizados con wavelets tienen muy buenas propiedades locales. Los elementos de la base vienen dados por:

$$\Psi_{j,k}(t) = 2^{\frac{j}{2}}\psi(2^j t - k) \tag{1.4.27}$$

Se dice que $\psi \in L^2(\mathbb{R})$ es una `wavelet ortonormal` si el conjunto $\{\Psi_{j,k}; j,k \in \mathbb{Z}\}$ forma una base ortonormal.

La expansión funcional wavelet de un conjunto de observaciones $x(t_j), \quad j = 1, ..., m$, en el espacio generado por la base ortonormal es igual a:

$$\chi(t) = \sum_j \sum_k w(j,k)\Psi_{j,k}(t) \tag{1.4.28}$$

siendo los $w(j,k)$ las funciones de peso o coeficientes de $\chi(t)$ asociados a la base.

1.4.6. Elección del número de elementos de la base

La elección del número de elementos de la base para obtener la función suavizada, dependerá del criterio que se aplique. El criterio debe proponer un equilibrio entre el nivel de ajuste de la curva a los puntos observados y la calidad funcional de dicha curva. Una base muy larga generaría una curva muy ajustada, aunque un comportamiento muy rugoso y deficientes condiciones funcionales; mientras que una base muy corta generaría una curva muy suave y buenas condiciones funcionales, pero sería una mala representación de las observaciones. Uno de los criterios más utilizados en FDA para la determinación del tamaño óptimo de la base está basado

en la Validación Cruzada Generalizada, implementado en las librerías FDA y fda.usc de R .

1.4.7. Suavizado local mediante funciones kernel

Para que un método de suavizado sea consistente, es razonable exigir que el valor de la estimación de la función en un punto t esté influenciado principalmente por las observaciones cercanas a t, a mayor proximidad mayor influencia. Los métodos de suavizado de bases de funciones verifican básicamente ese `principio de ponderación local`. En este apartado iremos un poco más lejos e impondremos, de manera explícita, que la función de suavizado dependa de `funciones de peso locales`. Un suavizado local a través de una `regresión kernel` es un procedimiento no paramétrico para estimar el comportamiento de una variable dependiente (X) para un valor dado de la independiente ($T = t$), usando `funciones tipo kernel`. Para ello, se estima la esperanza condicional:

$$E[X|T = t] = m(t), \tag{1.4.29}$$

de manera que el estimador del regresor $m(t)$ tenga la forma:

$$\widetilde{m}(t) = \chi(t) = \sum_{j=1}^{m} \omega_j(t) x(t_j), \tag{1.4.30}$$

donde los pesos ω_j solo tomarán valores grandes cuando t esté muy cerca del punto de referencia t_j. La idea básica es determinar los pesos $\omega_j(t)$ a partir de funciones kernel, $K(t)$, realizando cambios de escala y traslaciones en el argumento de la misma. Es decir, $\omega_j(t) = f(K(\frac{t+k}{h}))$; h se denomina `ancho de la ventana` y es el parámetro fundamental del suavizado local. El parámetro h puede ser constante o variable, en este último caso el valor de h se obtiene en función del rango de los k-vecinos más cercanos.

Las funciones kernel están diseñadas para alcanzar su valor máximo en cero, decrecer rápidamente y anularse, generalmente, para valores donde $|t| \geq 1$. Los kernel son generalizaciones del concepto de histograma. Las propiedades de un kernel son [14]:

- $|K(t)| < \infty$

- $K(-t) = K(t)$
- $\int |K(t)|dt < \infty$
- $\lim_{|t|\to\infty} |tK(t)| = 0$
- $\int K(t)dt = 1$

En su forma univariable, los kernel más usados son:

- `Kernel uniforme o rectangular`. La utilización del kernel uniforme genera un suavizado de `histograma móvil`:

$$K(t) = \frac{1}{2}\mathbb{1}_{(-1,1]}(t) \tag{1.4.31}$$

- `Kernel triangular`. Tiene, como su nombre anuncia, forma triangular:

$$K(t) = 1 - |t|, \text{ para } |t| < 1; \quad 0, \quad \text{en otro caso} \tag{1.4.32}$$

- `Kernel gaussiano`. Viene dado por la densidad de una Normal de media cero y desviación típica 1:

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{1.4.33}$$

- `Kernel Epanechnikov`. Es el kernel más estudiado. Es un segmento del perfil de un arco de parábola que se coloca sobre cada punto:

$$K(t) = \frac{3}{4}(1 - t^2), \text{ para } |t| < 1; \quad 0, \quad \text{en otro caso} \tag{1.4.34}$$

Uno de los procedimientos más usado para obtener las ponderaciones de los $\omega_j(t)$ a partir de funciones kernel está basado en el estimador de Nadaraya-Watson ([15, 16]). Los pesos vienen dados por:

$$\omega_j(t) = \frac{K[(t_j - t)/h]}{\sum_{r=1}^{m} K[(t_r - t)/h]}, \tag{1.4.35}$$

que sustituidos en la expresión (1.4.30), nos da el suavizado por polinomios locales de Nadaraya-Watson:

$$\chi(t) = \frac{\sum_{j=1}^{m} K[(t_j - t)/h]x(t_j)}{\sum_{r=1}^{m} K[(t_r - t)/h]}, \tag{1.4.36}$$

El tamaño de h condiciona el grado de suavizado y, al igual que con la determinación del número de funciones base comentada anteriormente, debe buscarse un equilibrio entre el sesgo y la varianza del ajuste. Así, un h muy grande determinará un sobre-suavizado, es decir, mucho sesgo y la pérdida subsiguiente de información relevante aportada por la muestra de observaciones; por otra parte, un h muy pequeño se traducirá en un sobre-ajuste (en el límite el suavizado se convierte en una interpolación), lo que supondrá una alta varianza y mala calidad funcional del suvizado. Teniendo en cuenta lo anterior, hay muchos procedimientos de determinación de h que, entre otros criterios, buscan establecer un equilibrio entre sesgo y varianza; además, la elección de h también dependerá de la función kernel que se haya elegido. Ruppert, Sheather y Wand ([17]) proponen una metodología basada en `plug-in bandwidth selection` para obtener el ancho de banda en una regresión local con kernel gausiano. [12].

1.5. Componentes principales funcionales

Una vez obtenida la forma funcional de los individuos, $\{\chi_1, \chi_2, ..., \chi_n\}$, mediante el procedimiento adecuado en cada caso, dispondremos de un conjunto de curvas definidas en un soporte $[0, T]$. En lo que sigue, supondremos que las curvas constituyen una muestra aleatoria o que, al menos, conforman un conjunto razonablemente homogéneo peteneciente a una población. Aparte de esta adscripción al grupo, el nexo en común para el conjunto de individuos lo constituye el procedimiento de obtención de sus representaciones funcionales: en el caso de suavizado por una base de funciones, el elemento común será dicha base, en el caso de un suavizado local, compartirán kernel, polinomios regresores y ancho de banda, etc.

Tomando como punto de partida del analisis FDA el conjunto de funciones $\mathcal{X} = \{\chi_i\}_{i=1,...,n}$ y, admitiendo una cierta homogeneidad en dicho conjunto, uno de los recursos más interesantes del análisis funcional es la base de componentes principales funcionales, que podríamos considerar una "autobase" o base interna del conjunto $\mathcal{X}$. La idea es adaptar el procedimiento de obtención de componentes principales vectorial (PCA), de dimensión finita, al caso funcional, infinito-dimensional. Para realizar

[12] El método está implementado en la función `dpill{KernSmooth}` del software `R`.

esta adaptación, nos basaremos en el denominado `esquema de dualidad`, que permite la representación de individuos y variables en un mismo espacio. La primera parte del contenido de esta sección se basa en el trabajo de Ramsay ([18]) titulado "When de Data Are Functions". Por otra parte, una vez obtenidas las componentes principales funcionales (FPCA), veremos algunas de sus posibilidades aplicativas, derivadas en gran medida del `desarrollo de Karhunen-Loéve`.

Caso matricial. Supongamos que $\mathbf{X}$ representa una matriz de datos observados, $x_{ij},\ i = 1, \cdots, n,\ j = 1, \cdots, p$. $\mathbf{X}$ puede verse como un operador o una correspondencia entre dos espacios vectoriales, $\mathbf{E}$ y $\mathbf{F}$, de dimensiones respectivas p y n, de forma que cada uno de ellos tiene definido un producto interno, $\langle \mathbf{e}_j, \mathbf{e}_k \rangle$ y $\langle \mathbf{f}_j, \mathbf{f}_k \rangle$, y están dotados de una base ortonormal. El espacio $\mathbf{E}$ es el espacio de los individuos, mientras que $\mathbf{F}$ es el espacio de las variables.

Las aplicaciones que representa $\mathbf{X}$ son:

- $\mathbf{X} : \mathbf{E} \rightarrow \mathbf{F}$, que transforma un vector $\mathbf{e} \in E$ de la forma $\mathbf{f} = \mathbf{Xe}$. Si consideramos que $\mathbf{E}$ está generado por n vectores $\mathbf{x_i}$, entonces se verifica que:

$$f_i = \langle \mathbf{e}, \mathbf{x_i} \rangle = \sum_{j=1}^{p} e_j x_{ij} \qquad y \qquad \mathbf{f} = \langle \mathbf{e}, \mathbf{X} \rangle = \mathbf{Xe} \tag{1.5.37}$$

  es decir, al aplicar el operador $\mathbf{X}$ sobre un elemento de su espacio, $\mathbf{e}$, se obtiene un vector del espacio $\mathbf{F}$ cuya componente i-ésima mide el ángulo que forma $\mathbf{e}$ con $\mathbf{x_i}$. El operador $\mathbf{X}$ aplicado sobre los vectores canónicos del espacio $\mathbf{E}$ genera trivialmente las coordenadas de las magnitudes originales, $\mathbf{v_j}$.

- $\mathbf{X^t} : \mathbf{F} \rightarrow \mathbf{E}$, que transforma un vector $\mathbf{f} \in F$ de la forma $\mathbf{e} = \mathbf{X^t f}$. Si consideramos que $\mathbf{F}$ está conformado por p vectores $\mathbf{v_j}$, entonces se verifica que:

$$e_j = \langle \mathbf{f}, \mathbf{v_j} \rangle = \sum_{i=1}^{n} f_i x_{ij} \qquad y \qquad \mathbf{e} = \langle \mathbf{f}, \mathbf{X^t} \rangle = \mathbf{X^t f} \tag{1.5.38}$$

  es decir, al aplicar el operador $\mathbf{X^t}$ sobre un elemento de su espacio, $\mathbf{f}$, se obtiene un vector del espacio $\mathbf{E}$ que viene dado como una combinación lineal de las $\mathbf{x_i}$ ponderados por las componentes de $\mathbf{f}$. El operador $\mathbf{X^t}$ aplicado sobre los

vectores canónicos del espacio $\mathbf{F}$ genera trivialmente los puntos originales, $\mathbf{x_i}$. Por otra parte, la aplicación $\mathbf{X}$ y su traspuesta verifican que:

$$\langle \mathbf{f}, \mathbf{Xe} \rangle = \langle \mathbf{X^t f}, \mathbf{e} \rangle \tag{1.5.39}$$

Desde esta óptica, el PCA se convierte en el problema de obtener una secuencia de vectores, $\xi_\mathbf{1}, \xi_\mathbf{2}, \cdots$ de norma 1 del espacio $\mathbf{E}$, cuya transformación por el operador $\mathbf{X}$ tenga módulo máximo y sean ortogonales entre sí:

$$\begin{array}{c} \max\limits_{\xi_\mathbf{j}} \langle \mathbf{X}\xi_\mathbf{j}, \mathbf{X}\xi_\mathbf{j} \rangle \\ s.a. \qquad ||\xi_\mathbf{j}||^2 = 1 \\ \langle \xi_\mathbf{j}, \xi_\mathbf{k} \rangle = 0 \quad \forall k < j \end{array} \tag{1.5.40}$$

Además de las dos anteriores aplicaciones obtenidas a partir de $\mathbf{X}$, ésta también determina otras dos aplicaciones que transforman un espacio en sí mismo, resultantes de las composiciones de las anteriores:

- $\mathbf{V} = \mathbf{X^t} \circ \mathbf{X} : \mathbf{E} \rightarrow \mathbf{E}$. Cuando las columnas de $\mathbf{X}$ están centradas en cero, $(n-1)^{-1}\mathbf{V}$ es la matriz de varianzas covarianzas de $\mathbf{X}$. A partir de (1.5.40), el ACP puede expresarse como:

$$\begin{array}{c} \max\limits_{\xi_\mathbf{j}} \langle \xi_\mathbf{j}, \mathbf{X^t X}\xi_\mathbf{j} \rangle = \max\limits_{\xi_\mathbf{j}} \langle \xi_\mathbf{j}, \mathbf{V}\xi_\mathbf{j} \rangle \\ s.a. \qquad ||\xi_\mathbf{j}||^2 = 1 \\ \langle \xi_\mathbf{j}, \xi_\mathbf{k} \rangle = 0 \quad \forall k < j \end{array} \tag{1.5.41}$$

  Cuya solución son los autovectores de $\mathbf{V}$.

- $\mathbf{W} = \mathbf{X} \circ \mathbf{X^t} : \mathbf{F} \rightarrow \mathbf{F}$. Análogo al anterior.

Es posible decir que la matriz de datos determina un subespacio de $\mathbf{F}$ en el sentido de que cualquier vector $\mathbf{e} \in \mathbf{E}$ es transformado por $\mathbf{X}$ en un vector de $\mathbf{F}$. La transformación por $\mathbf{X}$ de una base ortonormal, $\mathbf{e}_j; j = 1, \cdots, p$, de $\mathbf{E}$ nos da un conjunto de vectores de $\mathbf{F}$ que generan la imagen de cualquier vector $\mathbf{e} \in \mathbf{E}$. La dimensión de este subespacio de $\mathbf{F}$, digamos $\mathbf{F'}$, es, en general, igual a $min(n, p)$. Si un elemento cualquiera, $\mathbf{f}$ de $\mathbf{F}$ no se encuentra en este subespacio, su imagen por $\mathbf{W}$ ciertamente lo estará, ya que $\mathbf{W}(\mathbf{f}) = \mathbf{X} \circ \mathbf{X^t}(\mathbf{f})$ primero aplica, mediante $\mathbf{X^t}$, $\mathbf{f}$ en $\mathbf{E}$ y luego mediante $\mathbf{X}$ se obtiene un elemento de $\mathbf{F'}$.

Caso funcional. En el caso funcional la aplicación $\mathbf{X}$ tiene dimensión $n \times \infty$, el espacio $\mathbf{E}$ es el espacio de los individuos, de dimensión infinita, representado por n funciones x_i definidas en un intervalo finito $[0, T]$, mientras que el espacio $\mathbf{F}$ es, generalmente, un espacio tiempo de dimensión n. Los dos espacios quedan caracterizados por:

- Supongamos que $\mathbf{E}$ es un espacio de dimensión infinita, que representa un proceso aleatorio $\{X(t) : t \in [0, T]\}$, del cual se dispone de una muestra aleatoria de n funciones x_i de cuadrado integrable. De hecho, $\mathbf{E}$ es un espacio de Hilbert separable $L^2[0, T]$, lo que implica que cualquier función del espacio puede expresarse como combinación lineal de un conjunto numerable de funciones ortonormales. El producto escalar entre dos funciones del espacio $\mathbf{E}$ viene dado por:

  $$\langle e_j, e_k \rangle = \int_0^T e_j(t) e_k(t) dt$$

  El producto escalar es una medida de la relación entre dos elementos del espacio. En la Figura (1.4) pueden observarse las dos situaciones extremas, a la izquierda los productos escalares entre pares de funciones son altos, negativos o positivos según el caso, mientras que el producto entre las dos funciones de la derecha estará próximo a cero.

  La norma-$\mathcal{L}^2$ se define como:

  $$\| e \| = \sqrt{\langle e, e \rangle} = \sqrt{\int_0^T e(t)^2 dt}$$

  Esta norma nos da la magnitud de la función $e(t)$ en el espacio $\mathcal{L}^2$. Extendiendo algunas de las propiedades del caso vectorial, el "coseno del ángulo" entre dos funciones, e_j y e_k, sería:

  $$\cos(\theta) = \frac{\langle e_j, e_k \rangle}{\| e_j \| \| e_k \|} = \frac{\int_0^T e_j(t) e_k(t) dt}{\sqrt{\int_0^T e_j(t)^2 dt} \sqrt{\int_0^T e_k(t)^2 dt}}$$

  Decimos que las funciones e_j y e_k son ortogonales si $\cos(\theta) = 0$, es decir si son "perpendiculares", para lo que debe verificarse que

  $$\int_0^T e_j(t) e_k(t) dt = 0$$

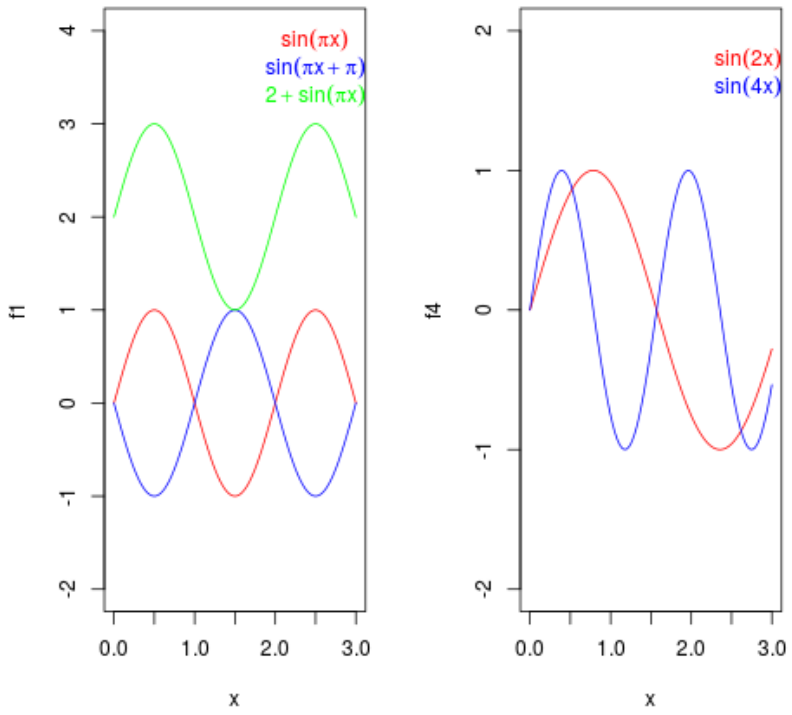


Figura 1.4: Funciones proporcionales y ortogonales

- **F** es el espacio tiempo, de dimensión n. Cualquier punto $t \in [0, T]$ está representado en este espacio por un vector $(x_1(t), x_2(t), \cdots, x_n(t))^T$

Las aplicaciones definidas ahora por la matriz de datos **X** son:

- $\mathbf{X} : \mathbf{E} \rightarrow \mathbf{F}$, que transforma una función $e(t) \in E$ en un vector **f** de **F**. Por equivalencia con (1.5.37), la coordenada i-ésima de **f** es igual a:

$$f_i = \langle e, x_i \rangle = \int_0^T x_i(t) e(t) dt \tag{1.5.42}$$

es decir, al aplicar el operador **X** sobre un elemento, $e(t)$, de su espacio se obtiene un vector del espacio **F** cuya componente i-ésima mide el ángulo que

forma e con x_i. Si $e(t)$ es ortogonal a algunas de las $x_i(t)$ las componentes correspondientes serán nulas.

- $\mathbf{X^t} : \mathbf{F} \rightarrow \mathbf{E}$, que transforma un vector $\mathbf{f} \in F$ en una función de $\mathbf{E}$:

$$e(t) = \mathbf{X^t f} = \sum_{i=1}^{n} f_i x_i(t) \tag{1.5.43}$$

es decir, al aplicar el operador $\mathbf{X^t}$ sobre un elemento, $\mathbf{f}$, de su espacio se obtiene una curva del espacio $\mathbf{E}$ que viene dada como una combinación lineal de las $x_i(t)$ ponderados por las componentes de $\mathbf{f}$. El operador $\mathbf{X^t}$ aplicado sobre los vectores canónicos del espacio $\mathbf{F}$ genera trivialmente las curvas originales, x_i. Es fácil ver que $\langle \mathbf{X}e, \mathbf{f} \rangle = \langle e, \mathbf{X^t f} \rangle$.

Desde esta óptica, el `FPCA` se convierte en el problema de obtener una secuencia de curvas, $\xi_1, \xi_2, \cdots$ de norma 1 del espacio $\mathbf{E}$, cuya transformación por el operador $\mathbf{X}$ tenga módulo máximo y sean ortogonales entre sí:

$$\begin{array}{cc} & \max_{\xi_j} \langle \mathbf{X}\xi_j, \mathbf{X}\xi_j \rangle \\ s.a. & ||\xi_j||^2 = 1 \\ & \langle \xi_j, \xi_k \rangle = 0 \quad \forall k < j \end{array} \tag{1.5.44}$$

Al igual que en el caso matricial, obtendremos las dos aplicaciones endomórficas que determina la aplicación X.

- $\mathbf{V} = \mathbf{X^t} \circ \mathbf{X} : \mathbf{E} \rightarrow \mathbf{E}$. Componiendo, se tendrá:

$$\mathbf{V}(e(t)) = \sum_{i=1}^{n} x_i(t) [\int_0^T x_i(u) e(u) du] = \int_0^T [\sum_{i=1}^{n} x_i(t) x_i(u)] e(u) du \tag{1.5.45}$$

$\mathbf{V}$ posee una estructura del tipo $\int K(t,u) e(u) du$, siendo $K(t,u)$ el núcleo de la transformación $\mathbf{V}$. Obsérvese que $K(t,u)/(n-1)$ es la covarianza intrapuntos del conjunto de funciones que determinan $\mathbf{E}$.

- $\mathbf{W} = \mathbf{X} \circ \mathbf{X^t} : \mathbf{F} \rightarrow \mathbf{F}$.

$$\mathbf{W}(\mathbf{f}) = \int_0^T x_i(t) \sum_{l=1}^{n} f_l x_l(t) dt \tag{1.5.46}$$

A partir de (1.5.44), el `FPCA` puede expresarse como:

$$\begin{array}{c} \max_{\xi_j} \langle \xi_j, \mathbf{X^t X} \xi_j \rangle = \max_{\xi_j} \langle \xi_j, \mathbf{V} \xi_j \rangle \\ s.a. \qquad ||\xi_j||^2 = 1 \\ \langle \xi_j, \xi_k \rangle = 0 \quad \forall k < j \end{array} \tag{1.5.47}$$

Cuyas soluciones son las autofunciones de $\mathbf{V}$, que se obtienen resolviendo la ecuación integral de Fredholm:

$$(V\xi)(t) = \int_0^T K(u,t)\xi(u)du = \langle K(t,\cdot), \xi \rangle = \lambda \xi(t) \tag{1.5.48}$$

Cada autofunción viene acompañada de su autovalor asociado, que están ordenados de mayor a menor, indicando la inercia de su autofunción. Según el Teorema de Mercer, el núcleo $K(t,u)$ admite el desarrollo:

$$K(t,u) = \sum_{i=1}^{\infty} \lambda_i \xi_i(t) \xi_i(u) \tag{1.5.49}$$

Por otra parte, cualquier función $e(t)$ puede expresarse, siguiendo el desarrollo de Karhunen-Loéve, como:

$$e(t) = \sum_{i=1}^{\infty} b_i \xi_i(t) \tag{1.5.50}$$

Donde la serie converge en media cuadrática en $[0,T]$ y los b_i se definen como:

$$b_i = \langle \xi_{\mathbf{i}}, \mathbf{e} \rangle = \int_0^T \xi_i(t) e(t) dt, \tag{1.5.51}$$

y representan la proyección de $e(t)$ en la i-ésima autofunción.

El `FPCA` permite expresar el conjunto de funciones muestrales, definidas en un espacio vectorial $\mathbf{L}^2$, como una combinación lineal de un número finito de `funciones propias`, haciendo una extrapolación del caso puntual. La contrapartida de los individuos, $\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})_{i=1,\ldots,n}$, del caso puntual son ahora funciones $x_i(t)_{i=1,\ldots,n}$, en el que el subíndice j, que identificaba la medición del individuo i-ísimo en la

variable j-ésima (x_{ij}), ha sido reemplazado por el argumento continuo t, que nos da información de la evolución de una magnitud a largo de un intervalo $[0, T]$.

Teniendo en cuenta (1.5.50), las funciones originales x_i pueden estimarse en función de la base constituida por las K primeras funciones principales como:

$$\widetilde{x_i}(t) = \sum_{k=1}^{K} b_{ik}\xi_k(t) = \sum_{k=1}^{K} \xi_k(t) \int_0^T \xi_k(t)x_i(t)dt, \quad i =, 1, \ldots, n \tag{1.5.52}$$

siendo b_{ik} la proyección de la curva observada $x_i(t)$ en la k-ésima función principal. Dicho de otra manera, $\widetilde{x_i}(t)$ es la proyección ortonormal de la curva $x_i(t)$ en el subespacio conformado por las K primeras componentes principales.

Algunas aplicaciones del FPCA

Como comentamos al principio de la sección, el FPCA además de tener valor por sí mismo, como técnica de reducción de la dimensión, es una herramienta muy útil en análisis funcional. Algunas de sus aplicaciones y utilidades son:

- `Métrica inducida por las Componentes Principales Funcionales.`

  Anteriormente, dentro del apartado `Semimétricas en un espacio de Hilbert funcional`, vimos como se podía obtener la semidistancia entre dos curvas a partir de la expresión (1.2.13). Por otro lado, Piza ([19]) propone una métrica inducida por las componentes principales funcionales vectoriales que denomina d_M:

  $$d_M(\underline{x}_i, \underline{x}_j) = \sum_{k=1}^{p} \lambda_k(\hat{x}_i^k - \hat{x}_j^k)^2, \tag{1.5.53}$$

  donde λ_k es el autovalor del k-ésimo eje principal del PCA y $\hat{x}_i^k$ y $\hat{x}_j^k$ son las coordenadas de los individuos $\underline{x}_i$ y $\underline{x}_j$ proyectados sobre el k-ésimo eje principal. Trasladando esta idea al campo funcional, podemos definir la medida inducida por las Componentes principales funcionales entre dos funciones χ_i y χ_j como:

  $$d_{M\infty}(\chi_i, \chi_j) = \sum_{k=1}^{\infty} \lambda_k(b_{ik} - b_{jk})^2, \tag{1.5.54}$$

siendo λ_k el autovalor del k-ésimo eje principal del FPCA y b_{ik} y b_{jk} el coeficiente k-ésimo de la proyección de las variables $\chi_i(t)$ y $\chi_j(t)$ en la base de componentes principales. Dicha distancia puede ser aproximada, usando una juiciosa elección de un número K de términos, como:

$$d_{MK}(\chi_i, \chi_j) = \sum_{k=1}^{K} \lambda_k (b_{ik} - b_{jk})^2, \tag{1.5.55}$$

- `Clasificación no supervisadas a partir del FPCA.`

  El desarrollo de Karhunen-Loéve dado por la expresión (1.5.50), permite la proyección de las curvas $\{\chi_i\}_{i=1,\ldots,n}$ en el espacio de las K primeras componentes principales. Peng y Müller ([20]), por un lado, y Abraham y otros ([21]), por otro, proponen una clasificación no supervisada vectorial, usando como representación del conjunto de curvas la matriz de coeficientes $\{b_{ik}\}_{i=1,\ldots,n;k=1,\ldots,K}$.

- `Estimación de la Volatilidad financiera a partir del FPCA.`

  Müller y otros ([22],[23]) en sus trabajos proponen una estimación de un proceso de volatilidad funcional en un mercado financiero, a partir del desarrollo de Karhunen-Loéve, como:

$$V(t) = \mu_V(t) + \sum_{k=1}^{K} \xi_k \phi_k(t), \tag{1.5.56}$$

  donde las ξ_k son variables incorreladas que satisfacen:

$$\xi_k = \int (V(t) - \mu_V(t))\phi_k(t)dt, \quad E[\xi_k] = 0, \quad V[\xi_k] = \lambda_k. \tag{1.5.57}$$

1.6. Clasificación funcional no supervisada

El Análisis Cluster, de Conglomerados o Clasificación no supervisada, proporciona una herramienta analítica para encontrar subgrupos diferenciados, no predefinidos de antemano, para un conjunto de individuos u objetos. Existen dos tipos principales de algoritmos Cluster: jerárquico y particional. Los algoritmos jerárquicos se basan en la definición de una jerarquía y tienen dos variantes: aglomerativos y divisivos; por su parte, los algoritmos particionales necesitan fijar el número de grupos, realizándose la clasificación de los individuos en los diferentes grupos a partir de la

optimización de un criterio (función objetivo), que está basado en la minimización de la variabilidad dentro del grupo y la maximización de la variabilidad entre los grupos. En esta sección estableceremos los principios básicos del análisis Cluster funcional univariable, a partir de una adaptación, bastante literal, del Cluster vectorial sobre un conjunto de variables.

El Análisis Cluster implica agrupar individuos en base a una razón de semejanza. Esta cuestión es importante, ya que puede mostrar información relevante sobre los individuos. Además del valor intrínseco de la clasificación, el cluster puede: detectar valores anómalos, determinar la dimensionalidad de los datos o poner de manifiesto las relaciones entre las variables ([24]). En este tipo de análisis, en general no hay información previa sobre el número y las características de los grupos y, la asignación de los elementos se realiza en base a una medida de similitud. Además, dependiendo del algoritmo, el número ideal de grupos a menudo lo determina el propio algoritmo. Se han desarrollado una gran cantidad de soluciones para producir la mejor clasificación, para un determinado criterio, de un conjunto de observaciones. Por todo lo anterior, consideramos que el análisis Cluster constituye una herramienta básica del análisis exploratorio funcional, lo que justifica su inclusión, junto con el FPCA, en este capítulo, donde estamos haciendo una revisión de los conceptos y recursos básicos del análisis funcional de datos.

Dos cuestiones relevantes que debemos analizar a fondo a la hora de aplicar un procedimiento Cluster funcional, y que van a determinar el resultado final de la clasificación, son la elección de la medida de similitud y del criterio de agrupación; además, si el algoritmo de clasificación es particional, habrá que establecer el número óptimo de grupos a formar. En general, los algoritmos de clasificación funcional no supervisados se basan en la elección de un vector representante de cada curva y en la obtención de la matriz de disimilaridades. A partir de dicha matriz, se suele aplicar alguno de los criterios de agrupación utilizados en la clasificación vectorial.

En la sección que sigue propondremos una nueva pseudodistancia funcional, basada en la distancia de Lipschitz, que puede ser de utilidad en ciertos contextos cuando aplicamos Cluster funcional.

1.6.1. Pseudométrica basada en la distancia de Lipschitz

La mayoría de los algoritmos de clasificación necesitan evaluar la similitud entre pares de individuos para realizar la agrupación. Hay muchas formas diferentes de definir disimilitudes entre los objetos, y la elección de esta medida de disimilaridad

depende en gran medida del tipo de datos con el que se trabaje. En el caso de datos funcionales la medida más usada es la distancia L^2. Dadas dos funciones f y g en el espacio $L^2(\tau)$, con $\tau = [T_1, T_2] \subset \mathbb{R}$ se tiene la distancia:

$$d(f,g) = \int_{T_1}^{T_2} (f(t) - g(t))^2 dt$$

Sin embargo, existen ocasiones en que esta medida no representa fielmente la idea intuitiva de disimilitud entre curvas. Por ejemplo, en algunos escenarios de análisis funcional parecería razonable concluir que dos funciones paralelas deberían estar muy próximas en la métrica empleada; así, si las curvas f y g que vienen dadas por $f(t) = t$ y $g(t) = t + c$, representan a dos individuos, es evidente que su comportamiento es bastante similar, salvo un sesgo o desplazamiento en el instante inicial, por lo que la disimilaridad entre ellos debería ser sensible a este hecho y arrojar un valor nulo o muy próximo a cero. Sin embargo, haciendo uso de la distancia usual, esto es la L^2, se tendría:

$$||f - g|| = \sqrt{|T_2 - T_1| * c^2} = c * \sqrt{(T_2 - T_1)} \neq 0$$

Otro ejemplo. Si se tienen las funciones $f_1(t) = t$ y $g_1(t) = -t$ en $[-2, 2]$, parece claro que los individuos a los que representan se diferencian mucho más en su comportamiento que otros par de individuos representados por las curvas f_2 y g_2 tales que $f_2(t) = t$ y $g_2(t) = t + 10$, definidas en el mismo intervalo. Haciendo uso de la distancia L^2, se tendría que:

$$d(f_1, g_1) = 8/\sqrt{3} < d(f_2, g_2) = 10 * \sqrt{4} = 20.$$

Para corregir esta distorsión en la evaluación de las diferencias entre dos curvas, en las situaciones que así lo demanden, vamos a introducir una pseudodistancia para datos funcionales basada en la métrica de Lipschitz.

Métrica de Lipschitz

Dada una métrica $d : X \times X \to [0, \infty)$, una función $f : X \to \mathbb{R}$ es λ-Lipschitz para un λ real no-negativo si $|f(x) - f(y)| \leq \lambda d(x, y)$, para cualesquiera x e y de X. Las funciones de la clase λ-Lipschitz las denotamos como $Lip_\lambda(X)$ y llamamos

$Lip(X)$ a la clase de funciones reales continuas que son λ-Lipschitz para algún $\lambda > 0$, siendo $Lip(X)$ un espacio vectorial. Para $f \in Lip(X)$ hay un mínimo valor $\lambda > 0$ para el cual $f \in Lip_\lambda(X)$, conocida como la constante de Lipschitz para f, la cual denotamos como $\alpha(f)$ ([25]). Es evidente que:

$$\alpha(f) = sup\{\frac{|f(x) - f(y)|}{|x - y|} : x \neq y\}$$

Es fácil ver que $\alpha : Lip(X) \to \mathbb{R}$ es una seminorma, ya que $\alpha(f) = 0$ si y solo si f es una función constante, además, si $\alpha(f) = \infty$, f no sería λ-Lipschitz. Si consideramos ahora las funciones de Lipschitz acotadas, $Lip^b(X)$, se define la norma de Weaver ([26]) de la función f como:

$$||f||_W = max\{||f||_\infty, \alpha(f)\},\ \forall f \in Lip^b(X)$$

Esta definición es equivalente a tomar $||f||_W = \alpha(f)$, siempre que f se anule al menos en un punto. Esta condición se añade precisamente para que se verifique la propiedad $||f - g|| = 0 \Leftrightarrow f = g$, necesaria para que $||.||_W$ sea norma; pero en el caso que nos interesa no es cierta, ya que dos rectas paralelas deberían tener distancia 0.

Así, vamos a definir $||f||_W = \alpha(f)$. Tenemos entonces que $||.||_W$ es una seminorma y por tanto $d_L(f, g) = ||f - g||_W$ es una pseudométrica. Para dos datos funcionales f y g, definimos la pseudodistancia de Lipschitz como:

$$d_L(f, g) = \alpha(f - g) = sup\{\frac{|(f - g)(x) - (f - g)(y))|}{|x - y|} : x, y \in X\}$$

.

Si $f, g : X \to \mathbb{R}$, haciendo $h = f - g$, se tendría:

$$d_L(h) = \alpha(h) = sup\{\frac{|h(x) - h(y))|}{|x - y|} : x, y \in X\}$$

Por el teorema del valor medio se sabe que existe un valor $k \in X$ tal que:

$$|h'(k)| = \frac{|h(x) - h(y)|}{|x - y|},$$

por lo que:

$$d_L(h) = \alpha(h) = sup\{|h'(k)|, k \in X\}$$

Volviendo a los ejemplos presentados anteriormente, se tendría que para dos curvas funcionales paralelas, $f = x(t)$ y $g = x(t) + c$, que representan a individuos con igual comportamiento, sería $h = c$ y $d_L(h) = sup\{|c'|\} = 0$. O considerando ahora las funciones $f_1 = x(t)$ y $g_1 = -x(t)$ en $[-2, 2]$, que intuitivamente se diferencian mucho más en su comportamiento que las funciones $f_2 = x(t)$ y $g_2 = x(t) + 10$, las pseudodistancias de Lipschitz en ambos casos serían:

$$d_L(2x(t)) = sup\{(2x)'\} = 2 \quad y \quad d_L(10) = sup\{(10)'\} = 0.$$

Propiedades.

Es fácil comprobar que la medida de Lipschitz, tal cual se ha definido es una pseudodistancia; es decir, verifica:

a) $\forall$ x,y, $d_L(x, y) \geq 0$

b) $\forall$ x, $d_L(x, x) = 0$

c) $\forall$ x,y, $d_L(x, y) = d_L(y, x)$

d) $\forall$ x,y,z, $d_L(x, y) \leq d_L(x, z) + d_L(z, y)$

1.6.2. Clasificación de técnicas de Cluster funcional univariable

Los métodos Cluster aplicados a datos funcionales se han desarrollado en los últimos años. El objetivo, como hemos indicado, es clasificar un conjunto de n curvas, $x_1, x_2, \ldots x_n$, que son observaciones de una variable funcional $\chi(t)$, obtenidas a partir del suavizado de mediciones en m instantes de tiempo. Jacques ([27]) clasifica las téecnicas de clasificación funcional en cuatro métodos: *raw−data clustering*, *two stage methods*, *non−parametric clustering* y *model−based clustering*.

- **Raw−data clustering**. Este tipo de métodos, llamados `métodos de regularización` mediante la discretización del intervalo de tiempo, consiste en usar directamente las "curvas" en los puntos en los que han sido observadas,

con lo que el enfoque funcional solo actúa como procedimiento de suavizado, o eliminación del ruido, de los datos observados. Se aplican técnicas de clustering para alta dimensionalidad vectorial (ver [28]). Los vectores de datos resultantes están autocorrelados y tienen una alta dimensionalidad, lo que conduce a estimaciones inestables de la matriz de covarianzas a menos que se imponga alguna forma de regularización ([29]).

- **Two-stage methods**. Son, como su nombre indica, *métodos de dos fases*, aunque es la primera de las fases la que los caracteriza como `métodos de filtrado`. Como hemos dicho, en la fase de filtrado *se reduce la dimensión de los datos* y en la segunda *se aplican las técnicas clásicas de clustering* para datos finitos. Una vez elegida la base de funciones, $\{\phi_i\}_{i=1,\dots,K}$, las curvas tienen una representación en la misma que viene dada por:

$$x_i(t) = \sum_{k=1}^{K} b_{ik}\phi_k(t)$$

  Cada función $x_i(t)$ queda caracterizada por el vector de coeficientes básicos, $b_{i1}, b_{i2}, \dots, b_{iK}$, que constituye una proyección única sobre un espacio K-dimensional. A partir de ahí, se obtiene la matriz de distancias euclídeas entre los vectores de coeficientes, para cada par de curvas. Esta es la fase del "filtering". La base de funciones puede ser genérica, como las de B−splines, Fourier, etc. o bien, la propia base de Componentes principales funcionales. Este último método es quizás el más extendido.

- **Non-parametric clustering**. Estos métodos se dividen en dos categorías. En la primera de ellas se aplica una técnica de Cluster particional, unido al uso de una distancia o disimilaridad específica entre curvas. Según la forma usada para calcular la distancia entre las "curvas", estos métodos pueden ser similares a los métodos raw-data clustering o a los métodos two-stage.

  Tarpey ([30]) demuestra en *"Linear transformation and the k−means clustering algorithm: Applications to Clustering Curves"* que aplicar el método de las k−means sobre datos funcionales con la métrica del espacio L^2, es equivalente a aplicar el método k−means clásico sobre los coeficientes básicos, después de aplicar una transformación lineal apropiada.

  Si x es una variable funcional y μ_j la variable funcional media de uno de los grupos, se medirá la distancia L_2 de x al grupo de centro μ_j , sobre un intervalo $[T_1, T_2]$ como:

$$||x - \mu_j|| = \sqrt{\int_{T_1}^{T_2} (x(t) - \mu_j(t))^2 dt}$$

En la segunda categoría se incluyen aquellos métodos que desarrollan nuevas heurísticas específicas para Cluster funcional. (Ver [31])

- **Model−based clustering**. Los métodos anteriores son válidos cuando los datos se han observado en una retícula densa, están igualmente espaciados y registrados en los mismos instantes para todos los individuos; básicamente, dichos métodos de clasificación discretizan el problema y reproducen los esquema de la clasificación vectorial. Los métodos de clasificación basados en modelos son aplicables a situaciones donde hay una escasez de observaciones, no necesariamente igualmente espaciadas ni observadas en los mismos instantes para todos los individuos.

  En la clasificación basada en modelos vectoriales ([29]), se asume que las observaciones son generadas a partir de una mixtura de K distribuciones: si $f_k(\mathbf{x}|\theta_k)$ es la densidad del k-ésimo cluster parametrizado por θ_k y $\mathbf{z_i} = (z_{i1}, ..., z_{in})$ es el vector de pertenencia de la i-ésima observación, $z - ik = 1$ si $\mathbf{x_i}$ pertenece al cluster k y 0 en caso contrario. Los $\mathbf{z_i}$ pueden ser tratados de dos maneras, como parámetros del modelo, que se estimarían maximizando la función de verosimilitud:

$$L_C(\theta_1, \ldots, \theta_K, \mathbf{z_1}, \ldots, \mathbf{z_n}|\mathbf{x_i}, \ldots, \mathbf{x_n}) = \prod_{i=1}^{n} f_{\mathbf{z_i}}(\mathbf{x_i}|\theta_{\mathbf{z_i}}) \tag{1.6.58}$$

que produce la solución de las k- medias cuando $f_{\mathbf{z_i}}(\mathbf{x_i}|\theta_{\mathbf{z_i}})$ es una normal multivariante con la matriz de covarianzas unitaria; este procedimiento se conoce como "clasificación de verosimilitud". Por otra parte, la pertencia al cluster puede ser tratado como un problema de imputación, donde $\mathbf{z_i}$ es una multinomial con parámetros $\pi_1, \ldots, \pi_K$, siendo pi_k la probabilidad de que $\mathbf{z_i}$ pertenezca al grupo k. En este caso, los parámetros son estimados maximizando:

$$L_M(\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K|\mathbf{x_1}, \ldots, \mathbf{x_n}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\mathbf{x_i}|\theta_k) \tag{1.6.59}$$

Este procedimiento se conoce como "mixtura de verosimilitudes". También se ha usado con éxito en este caso una distribución normal multivariante. La

diferencia entre ambos métodos es que en el primero se clasifica las curvas en un grupo, mientras que en el segundo, dada una curva, se obtiene la probabilidad de pertenencia de la misma a cada grupo.

James y Sugar ([32]) introducen una aproximación funcional a la clasificación basada en modelos a partir de los métodos de dos etapas, integrando las dos etapas en el mismo proceso; además, en contra con lo que ocurre con los métodos de regularización y de filtrado que no pueden aplicarse cuando las curvas están muestreadas en diferentes instantes de tiempo y que necesitan ciertas condiciones de regularidad en los datos, los métodos de clusterización basados en modelos se pueden en escenarios mucho menos restrictivos, puesto que trata cada curva de manera independiente.

Los métodos de clusterización basados en modelos, al igual que los de filtrado, convierten el problema original infinito dimensional en uno finito dimensional, usando una base de funciones; sin embargo, en lugar de tratar los coeficientes básicos como parámetros y ajustar una curva spline separada para cada individuo, James y Sugar usan un modelo de efectos aleatorios para los coeficientes. Ello les permite obtener mejores resultados sin importar lo irregulares que se muestren las curvas, exigiendo solo que el número de observaciones sea suficientemente grande para obtener las estimaciones de los parámetros. Aemás, las estimaciones de las curvas individuales son óptimas en términos de ECM.

Supondremos que parametrizamos nuestros modelos en términos de t, aunque la aproximación es válida en otros contextos. Partimos de un conjunto de funciones $\{x_1, \ldots, x_n\}$, "ocultas" tras los datos. Para estimar cada una de estas funciones, x_i, se dispone de m observaciones, $\{y_{i1}, \ldots, y_{im}\}$, registradas en un soporte discreto, $t_1, t_2, \ldots, t_m$. Como estas observaciones están registradas con errores, se tendrá que $y_{ij} = x_i(t_j) + \epsilon_{ij}$.

Supongamos ahora que queremos clasificar nuestros datos funcionales $\{x_1, \ldots, x_n\}$ en K grupos. Nos plantearemos para ello que una cierta función x_i, elegida aleatoriamente, pertenece al k-ésimo cluster, $k = 1, \ldots, K$, de centro $\mu_k(t)$ y constituido por n_k funciones. Asumimos que x_i sigue una distribución Normal. En la práctica x_i está observada con errores y solo en un conjunto finito de puntos. Si $x_i(t)$ pertenece al k-ésimo cluster se tendrá que:

$$\begin{pmatrix} x_i(t_1) \\ \ldots \\ x_i(t_m) \end{pmatrix} \sim N\left(M_k,\ \Omega_k \right)$$

donde:

$$M_k = \begin{pmatrix} \mu_k(t_1) \\ \dots \\ \mu_k(t_m) \end{pmatrix} \quad \text{y} \quad \Omega_k = \begin{pmatrix} \omega_k(t_1,t_1) & \dots & \omega_r(t_1,t_m) \\ \dots & & \\ \omega_k(t_m,t_1) & \dots & \omega_k(t_m,t_m) \end{pmatrix}$$

y

$$\omega_k(t_i,t_j) = cov_k(t_i,t_j) = \frac{1}{n_k-1}\sum_{l=1}^{n_k}(x_l(t_i)-\mu_k(t_i))(x_l(t_j)-\mu_k(t_j)),\ i,j=1,\dots,m$$

.

Ω_k es la matriz de covarianzas dentro del cluster k y suele ser inestable debido a su alta dimensionalidad. Además, los errores, $\epsilon_{ij} \sim N(0,\sigma^2)$, por lo que $\begin{pmatrix} \epsilon_{i1} \\ \dots \\ \epsilon_{im} \end{pmatrix} \sim N(0,\sigma^2.I_m)$, y como se tiene que $y_{ij} = x_i(t_j) + \epsilon_{ij}$ resulta que:

$$Y_i = \begin{pmatrix} y_{i1} \\ \dots \\ y_{im} \end{pmatrix} = \begin{pmatrix} x_i(t_1) \\ \dots \\ x_i(t_m) \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \dots \\ \epsilon_{im} \end{pmatrix} \sim N \begin{pmatrix} \mu_k(t_1) \\ \dots \\ \mu_k(t_m) \end{pmatrix}, \Omega_k + \sigma^2.I_m \Bigg)$$

En este tipo de métodos se utiliza la fase de regularización para estimar los parámetros del modelo, que son $\mu_k(t_j)$, $\omega_k(t_i,t_j)_{\{k=1,\dots,K\}}$ y σ^2, en una fina retícula de instantes de tiempo. Generalmente, no se realizan suposiciones sobre la forma funcional de los $\mu_k(t)$, pero sí se hacen algunas restricciones sobre la estructura de $\omega_k(t,t')$ para dotarla de estabilidad. En la fase de filtrado $x_i(t)$ se representa en una base de funciones, $\Phi(t) = (\phi_1(t),\dots,\phi_P(t))$, $x_i(t) = \Phi(t)\mathbf{b}$. Los vectores $\mathbf{b}$ son estimados de manera independiente para cada individuo y proyectan las curvas en el espacio de los coeficientes de dimensión P. Al conjunto de vectores básicos, $\{\mathbf{b_i}\}_{i=1,\dots,n}$, se le aplica un método Cluster con K grupos y las medias de los cluster resultantes son multiplicados por $\Phi(t)$ para obtener las estimaciones de los $\mu_k(t)$. Las estimaciones de los $\omega(t,t')$ son obtenidos de manera similar. La estimación de los parámetros del modelo se logra maximizando la verosimilitud de clasificación dada por (1.6.58) o la verosimilitud de la mixtura dada por (1.6.59), en ambos casos ello implica un proceso iterativo. Las curvas se asignan primero a un grupo (clasificación) o se le asigna una probabilidad de pertenecer a un grupo (mixtura), luego los parámetros se reestiman dadas las asignaciones actuales y el proceso se repite. En ([32]) pueden consultarse los detalles del algoritmo.

Elección del número de grupos

Uno de los principales problemas en Análisis Cluster es que la mayoría de los métodos de agrupación necesitan la elección del número de grupos antes de aplicar el procedimiento ([32]). Existen diferentes coeficientes utilizados para este propósito. Entre los más relevantes se encuentra el coeficiente Silhouette. Este coeficiente fue introducido por Rousseeuw en 1987 ([33]). Para cada elemento i del grupo A se define $a(i)$ como la distancia media de i a todos los elementos de A. Para cualquier otro grupo C se define $d(i, C)$ como la distancia media de i a todos los elementos de C y $b(i) = min\ d(i, C)$ variando C. En estas condiciones se define el coeficiente Silhouette en el punto i como:

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

Una vez obtenidos los $s(i)$, se define el coeficiente Silhouette de la clasificación como:

$$s = \frac{1}{n} \sum_{i=1}^{n} s(i)$$

Rousseeuw introdujo este coeficiente con la idea de dividir los datos en k grupos, siendo k aquel que determine el coeficiente s máximo.

Capítulo 2

Datos geodésicos

2.1. Estructura y significado de los datos

Sistema de Posicionamiento Global ([34])

NAVSTAR/GPS (NAVigation System with Timing and Ranging/ Global Positioning System) es un sistema de posicionamiento espacial desarrollado por el Departamento de Defensa de los Estados Unidos en 1973 a partir de los proyectos TIMATION y 621B, diseñados éstos con el objetivo de establecer un sistema de navegación pasiva utilizando medidas de distancia.

El sistema NAVSTAR/GPS se define como un sistema pasivo de navegación basado en satélites emisores de radiofrecuencias, que proporciona un marco de referencia espacio-temporal con cobertura global, independiente de las condiciones atmosféricas y de forma contínua en cualquier lugar de la Tierra.

El sistema GPS consta de segmentos principales:

- Un segmento espacial (formado por satélites que transmiten señales en dos frecuencias moduladas a partir de relojes atómicos altamente estables instalados a bordo).
- Un segmento de control (formado por satélites encargados del seguimiento del segmento espacial y de la actualización de sus señales).
- Un segmento de usuarios (formado por receptores pasivos que utilizan la señal de los satélites, obteniendo posición y tiempo).

El objetivo fundamental del Sistema de Posicionamiento Global es dotar de coordenadas absolutas a estaciones situadas en la Tierra o en su entorno espacial. Según los requisitos de precisión y el carácter, móvil o estático, de la estación se utilizarán métodos de posicionamiento absoluto en tiempo real, de posicionamiento relativo o GPS diferencial.

Las medidas GPS están ligadas al sistema de tiempo definido por los satélites y por el receptor. Por tanto, errores en los osciladores, tanto de los satélites como de los receptores, se propagarán a los resultados obtenidos en la resolución del modelo.

La no homogeneidad de la atmósfera, medio por donde se propaga la señal GPS, provoca errores en las medidas observadas, lo que hará necesario realizar una corrección Troposférica e Ionosférica por refracción atmosférica. Otro factor a tener en cuenta son las reflexiones que afectan a la señal por cuerpos reflectantes próximos al receptor. Este efecto se manifiesta en un aumento de la trayectoria de la señal y, consecuentemente, en un retardo en el tiempo de recepción de dicha señal. Aunque se utilizan antenas que disminuyen estas reflexiones, este efecto se considera error aleatorio en los diferentes modelos de tratamiento de datos observacionales.

En Geodesia, donde es necesaria una gran precisión, es obligado recurrir al posicionamiento relativo entre las estaciones de una red geodésica, para a partir del establecimiento de estaciones fijas, dotar a todas las estaciones de la red de coordenadas absolutas.

El posicionamiento relativo consiste en la determinación de la azimut, de la distancia relativa y de la diferencia en altura, o también los incrementos en las coordenadas cartesianas geocéntricas entre las estaciones que conforman la red geodésica. El Sistema de Posicionamiento Global permite resolver este problema mediante la observación simultanea, desde varias estaciones, de la señal emitida por los satélites.

Este método, basado en la construcción de diferencias entre las medidas realizadas en diferentes estaciones o diferentes satélites, en un mismo instante o en instantes sucesivos, elimina o reduce muchos de los efectos que limitan el sistema GPS. Dentro de este método se usan modelos de simples, dobles y triples diferencias.

El error observacional existente en las medidas realizadas por estos métodos se propagará al modelo establecido y se indicará en la correspondiente matriz de varianzas/covarianzas.

Supondremos que aunque exista una correlación física entre las observaciones realizadas, pues proceden de un mismo satélite, la única correlación posible es la derivada del propio modelo matemático. Supondremos también que los errores aleatorios pre-

sentes en las medidas de fase se distribuyen según una normal de media 0 y varianza σ^2 y que todas las observaciones realizadas tienen igual peso y son independientes y, por tanto, incorreladas.

Método de las dobles diferencias ([35])

La metodología empleada para el procesado de las observaciones GPS está basado en lo que se denomina el método de las dobles diferencias de la fase de la portadora.

La medida de la fase de la portadora resulta de la comparación de la fase de la señal portadora recibida en el receptor y la generada por el oscilador de dicho receptor. El modelo de la fase de la portadora, $L_i^k(t)$, para un receptor i y un satélite k es:

$$L_i^k(t) = \rho_i^k(t) + \lambda(\phi_i(t_0) - \phi^k(t_0)) + c[dt_i(t) - dt^k(t)] + \lambda N_i^k(t_0) - I_i^k(t) + T_i^k(t) + \epsilon_i^k, \tag{2.1.1}$$

donde:

- $\rho_i^k(t)$ es la distancia geométrica entre la posición del satélite k y el receptor i en los instantes de emisión y recepción de los códigos respectivamente, tal que $\rho_i^k(t) = \sqrt{(x^k(t) - x_i(t))^2 + (y^k(t) - y_i(t))^2 + (z^k(t) - z_i(t))^2}$, siendo $(x^k(t), y^k(t), z^k(t))$ y $(x_i(t), y_i(t), z_i(t))$ las coordenadas geocéntricas del satélite y del receptor en el instante t.
- $dt^k(t)$ representa el desfase entre el sistema de tiempo GPS y el del reloj del satélite.
- $dt_i(t)$ representa el desfase entre el sistema de tiempo GPS y el del reloj del receptor.
- $I_i^k(t)$ y $T_i^k(t)$ representan los retardos ionosféricos y troposféricos respectivamente.
- ϵ_i^k es un término de ruido que contiene todos los efectos no modelados.
- λ es la longitud de onda.
- $N_i^k(t_0)$ es la función entera ambigüedad de fase, debido a que cuando se adquiere la señal se tiene una ambigüedad en un número entero de longitudes de onda.

- $\phi_i(t_0) - \phi^k(t_0)$ es constante y representa la fase inicial del oscilador del receptor y del satélite.

En este modelo quedan por determinar el error del oscilador del receptor, la ambigüedad inicial, que se mantiene constante para cada satélite y las coordenadas geodésicas de la estación que aparecen en la distacia geométrica.

Las medidas de fase y las pseudodistancias de código están afectadas por errores sistemáticos y aleatorios, siendo necesaria su modelización para eliminar o minimizar su efecto. Para ello se utilizan combinaciones de los observables, que se denominan simples, dobles y triples diferencias ([36, 37])

El modelo de dobles diferencias se define como la diferencia entre dos lecturas simultáneas de la fase de la portadora de un mismo satélite k en dos receptores distintos, i y j, cuya posición relativa trata de determinarse. Es decir

$$\begin{aligned} \phi^k_{i,j}(t) \quad &= \phi^k_i(t) - \phi^k_j(t) = \\ &= \rho^k_{i,j}(t) + \lambda(\phi^k_{i,j}(t_0) + c(dt_i(t) - dt^k(t))) + \lambda = \\ &= N^k_{i,j}(t_0) + I^k_{i,j}(t) + Ti, j^k(t) + \epsilon i, j^k \end{aligned}$$

En este modelo se elimina el efecto de la fase inicial del oscilador del satélite y el error del oscilador de cada uno de los satélites con respecto al tiempo GPS.

Procesado de las observaciones GNSS-GPS

Las mediciones GPS se utilizan en este trabajo para cuantificar la deformación superficial y para obtener desplazamientos tridimensionales en función del tiempo.

El procesado de los datos se ha efectuado con el software Bernese 5.0 ([38]), desarrollado por la universidad de Berna, el cual utiliza el método de las dobles diferencias como principal técnica para la obtención de soluciones GPS, obtenéndose coordenadas cartesianas para cada estación en un intervalo de tiempo específico (en este caso, se han utilizado sesiones de 24 horas 30 segundos). Se ha utilizado el marco de referencia ITRF2008.

Una de las principales características de este software es su estructura modular. El formato de ficheros usados se corresponde con estándares de IGS o bien permiten su transformación desde éstos. El procesado de datos con Bernese 5.0 responde al siguiente esquema:

1. Un paso previo para preparar y transformar los ficheros orbitales al formato requerido.

2. Un pre-procesamiento, que consiste en la preparación de los datos para poder efectuar el cálculo final de las dobles diferencias receptores-satélites por época. Para lo cual se utilizan los siguientes programas:

   a) Programa CODSPP: obtiene las correcciones necesarias en los relojes de los receptores para sincronizar todas las observaciones con la escala de tiempo GPS utilizando las medidas de código. Se efectúa un ajuste estándar por mínimos cuadrados basado en la ecuación fundamental de observación de código para obtener los parámetros necesarios. Se estiman unas primeras coordenadas aproximadas para los distintos vértices, cuyas coordenadas a priori no son conocidas.

   b) Programa SNGDIF: crea un fichero con las simples diferencias de fase para cada enlace y día, esto es, prepara los ficheros de observación formando las simples diferencias tanto de código como de fase.

   c) Programa MAUPRP: comprueba todas las observaciones y obtiene a partir de los residuales de la solución mínimos cuadrados de las triples diferencias de fase por cada línea procesada, fijando las coordenadas de una de las estaciones, las coordenadas de la otra. Además, detecta los saltos de ciclo y errores groseros para su corrección.

3. Procesado y ajuste, para la estimación de parámetros y ajuste de soluciones.

   a) Programa GPSEST: permite la estimación de distintos tipos de parámetros: parámetros de troposfera, ionosféricos, órbitas, cálculo y estimación de ambigüedades, obtención de coordenadas y salida de ficheros.

   b) Programa ADDNEQ: realiza el ajuste de los ficheros anteriores.

2.2. Depuración de los datos GPS

El sistema GNSS-GPS es una herramienta fundamental en el estudio de los modelos de deformación de la superficie terrestre. La observación de los satélites que cubren las estaciones permanentes de una red de posicionamiento permite obtener series de datos de las coordenadas geodésicas topocéntricas, Norte, Este y Altitud, de cada una de las estaciones.

El estudio detallado de estos datos aporta información relevante sobre el movimiento de las placas tectónicas en la zona. La forma de obtención, el procesado posterior y la propia naturaleza de este tipo de observaciones hace que las series presenten valores faltantes y otros fuera de rango, que podrían resultar atípicos. Por ello se precisa de un procedimiento de "limpieza" y reconstrucción que corrija las series originales.

Los valores fuera de rango pueden tener un origen endógeno, asociados al método de obtención y procesado de datos, o exógeno, como los que se derivan de movimientos sísmicos ([39]). Hay que tener en cuenta que en la observación de los datos pueden producirse errores derivados del propio sistema GPS, de los movimientos no controlados de las estaciones, del procesado de datos en relación a la estaciones de referencia, etc. Dichos errores son pues de distinta naturaleza y tamaño, por lo que nos planteamos un doble arreglo de datos: en primer lugar un filtrado, basado en una medida de los errores asociados a los datos en cada instante del tiempo y, en segundo lugar, un suavizado discreto, para lo que se usará un filtro Kalman ([40]).

Son numerosas las técnicas usadas en la detección de atípicos en series temporales. Las más populares se pueden agrupar en tres tipos: métodos basados en estadísticos de razón de verosimilitudes, métodos basados en medidas de influencia y métodos bayesianos. Otros interesantes enfoques, planteados en los últimos años, están orientados al análisis de valores atípicos en conjuntos de datos multivariantes ([41]) y en conjuntos de datos funcionales ([42]). En el caso que nos ocupa, tratamos con series de datos que incorporan, como información adicional, una medida de error de procesado de los valores de la serie, proporcionada por el software Bernese. Rosado et al. en ([43]), usan esta información realizando un filtrado a través de un ajuste lineal de mínimos cuadrados ponderados, en el que se asignan pesos a las observaciones en función de las medidas de error del procesado. En general, los resultados que se obtienen son aceptables, puesto que el movimiento en el plano se ajusta bien mediante una recta; sin embargo, por el efecto de los movimientos sísmicos se puede romper esta tendencia ([39]). Además, la componente Elevación no es lineal, sino periódica sinuosidal.

Por estos motivos, en este trabajo se propone el uso de un criterio alternativo, que tiene en cuenta la información proporcionada por el error, pero que no asume la linealidad de las series de datos en función del tiempo. Nuestra propuesta es hacer uso de la variable obtenida de la aplicación del software Bernese, que recoge el error en cada medición, de forma local, eliminándose los datos que superen dos veces el error medio de cada estación GPS.

2.2.1. Filtro Kalman

El filtro Kalman es un algoritmo desarrollado por Rudolf E. Kalman en 1960 que describe una solución recursiva para determinar el estado no medible de un modelo lineal dinámico (MLD) a partir de las observaciones que se obtienen de él.

El modelo MLD más sencillo es el modelo lineal de paseo aleatorio con ruido blanco o *modelo polinómico de orden 1*. Este modelo es apropiado para series temporales que no muestren tendencia. Si se asume que los datos están afectados de una velocidad que depende del tiempo se dispone de un *modelo MLD polinómico de orden 2*. Este caso sí es apropiado para datos que presentan tendencia.

El filtro Kalman construye la mejor estimación posible de una variable oculta dentro de una medición a partir de la información suministrada por los sensores de medida, la acción de control y el estado del sistema en un instante previo. En el filtro Kalman para describir un sistema lineal se supone que:

- El ruido de medición ν_k tiene una distribución normal de media cero y la matriz de covarianzas es diagonal. Así $E[\nu_k] = \bar{\nu_k} = 0$ y la matriz de covarianzas V_k del ruido de medición es:

$$Cov(\nu_k; \nu_k) = Vk = v\delta_{n,m} \tag{2.2.2}$$

- El ruido de procesamiento k tiene una distribución normal de media igual a cero y su matriz de covarianzas es diagonal. $E[\omega_k] = \bar{\omega_k} = 0$ y la matriz de covarianzas Q_k viene definida por:

$$Cov(\omega_k; \omega_k) = Qk = q\delta_{n,m} \tag{2.2.3}$$

- Los ruidos de medición y procesamiento son independientes, con lo que:

$$Cov(\nu_k; \omega_k) = 0 \tag{2.2.4}$$

El filtro es un algoritmo recursivo en dos fases: predicción−corrección. Se pronostica un nuevo estado a partir de su estimación previa y, a continuación se añade un término de corrección proporcional al error de predicción, siendo este último minimizado estadísticamente. Se parte de la hipótesis de normalidad del vector de estado inicial y de las perturbaciones del sistema. De tal forma que es posible calcular la

función de verosimilitud sobre el error de predicción, con lo cual se lleva a cabo la estimación de los parámetros no conocidos del sistema.

En el filtro Kalman, el estado x_k de un sistema en el instante de tiempo k se define como el conjunto de variables que caracterizan el sistema en k. Este sistema queda descrito en el instante k a través de la expresión:

$$x_k = A_k x_{k-1} + B_k u_k + \omega_k, \tag{2.2.5}$$

donde x_{k-1} es el estado en el instante $k-1$, u_k es la acción de control realizada sobre el sistema para modificar el estado x_{k-1}, ω_k es el ruido de procesamiento, A_k es la matriz de transición de estados del sistema y B_k es la matriz que regula la acción de control. Además hay que considerar la ecuación que relaciona el estado del sistema con la información medida por los sensores en el instante k:

$$z_k = H_k x_k + \nu_k, \tag{2.2.6}$$

siendo x_k el estado en el instante k, z_k la medición de los sensores, ν_k el ruido aleatorio durante la medición y H_k la matriz de relación entre la medición y el estado del sistema. El procedimiento de estimación completo es el siguiente:

El modelo es formulado para un conjunto inicial de parámetros dados y los errores de predicción del modelo son generados por el filtro. Estos son utilizados para evaluar recursivamente la función de verosimilitud hasta maximizarla.

- En el primer paso el filtro Kalman predice el estado de sistema $\hat{x}_{\bar{k}}$ mediante el cálculo del valor esperado de x_k en la ecuación 2.2.5, obteniéndose así una primera estimación del estado actual del sistema:

  $$\hat{x}_{\bar{k}} = A_k \hat{x}_{k-1} + B_k u_k, \tag{2.2.7}$$

  donde $\hat{x}_{\bar{k}}$ es el estado predicho del sistema en el tiempo k a partir del estado corregido $\hat{x}_{k-1}$. La matriz de covarianzas $P_{\bar{k}}$ para la estimación será:

  $$P_{\bar{k}} = A_k P_{k-1} A_k^t + Q_k, \tag{2.2.8}$$

  siendo P_{k-1} la matriz de covarianzas del estado $\hat{x}_{k-1}$ y Q_k es la matriz de covarianzas asociada al ruido de procesamiento definida por la ecuación 2.2.2.

- En la fase de corrección la predicción mejora calculando la ganancia de Kalman K_k,

$$K_k = Cov(x_k, z_k)(Cov(z_k, z_k))^{-1}. \tag{2.2.9}$$

La ganancia de Kalman relaciona las mediciones con el estado predicho $\hat{x}_{\bar{k}}$ para reducir la incertidumbre del cálculo. Kalman ([40]) encontró esta expresión al minimizar el error cuadrático de la matriz de covarianzas del estado corregido. En términos de las matrices del sistema, la ganancia de Kalman es:

$$K_k = (P_{\bar{k}} H_k^t) \left(H_k P_{\bar{k}} H_k^t + V_k\right)^{-1}, \tag{2.2.10}$$

donde V_k es la matriz del ruido de medición definida en la ecuación 2.2.2.

El estado corregido $\hat{x_k}$ del sistema es:

$$\hat{x_k} = \hat{x}_{\bar{k}} + K_k \left(z_k - H_k \hat{x}_{\bar{k}}\right), \tag{2.2.11}$$

donde z_k es la medición realizada en el instante de tiempo k. La matriz de covarianzas P_k del estado corregido será:

$$P_k = (I - K_k H_k) P_{\bar{k}}. \tag{2.2.12}$$

En esta fase de trabajo, el estado y la matriz de covarianzas del sistema se corrigen con la información de las mediciones.

- El estado x_{k+1} de un sistema en el instante de tiempo k+1 queda descrito a través de la expresión:

$$x_{k+1} = A_k x_k + B_k u_k + \omega_k, \tag{2.2.13}$$

y por la ecuación que relaciona el estado del sistema con la información medida por los sensores en el instante k:

$$z_k = H_k x_k + \nu_k \tag{2.2.14}$$

En el caso de usar un modelo de segundo orden, como se verifica:

$$A_k = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},\ x_k = \begin{pmatrix} \mu_k \\ \beta_k \end{pmatrix} y w_k = \begin{pmatrix} w_{k,1} \\ w_{k,2} \end{pmatrix}$$

las ecuaciones (2.2.13) y (2.2.14) se transforman en:

$$\begin{aligned} z_k &= \mu_k + v_k; \quad v_k \sim N(0, V) \\ \mu_{k+1} &= \mu_k + \beta_k + w_{k,1}; \quad w_{k,1} \sim N(0, \sigma_\mu^2) \\ \beta_{k+1} &= \beta_k + w_{k,2}; \quad w_{k,2} \sim N(0, \sigma_\beta^2) \end{aligned} \tag{2.2.15}$$

Capítulo 3

Análisis de la red SPINA desde un enfoque funcional

3.1. Confluencia de placas en Andalucía

Desde el punto de vista geológico, la interacción entre la Península Ibérica y África da lugar a una compleja región localizada en la parte oeste del borde de las placas Euroasiática y Africana. El borde de placas está muy bien delimitado en la parte oceánica mientras que en la región Iberia-África, el borde es más difuso y constituye una amplia área de deformación ([44]). Esta zona de subducción, en la que una placa desciende bajo la otra, no tiene los límites bien definidos y se caracteriza por una actividad sísmica moderada. Por ello, es importante estudiar la geodinámica en esta región (ver Fig. 3.1)

3.2. Datos. Red SPINA

La red geodésica SPINA (Sur de la Península Ibérica y Norte de África) está compuesta por estaciones geodésicas permanentes pertenecientes, a su vez, a distintas redes GPS permanentes. Estas redes son RENEP (Red Nacional de Estaciones Permanentes de Portugal), RAP (Red Andaluza de Posicionamiento, situada en Andalucía, Sur de España), MERISTEMUM (Murcia, Sureste de España), IGS (Sistema Internacional GNSS), IGN (estaciones CORS del Instituto Geográfico Nacional Español), REGAM (Red Geodésica Activa de Murcia, al sureste de España) y ERVA

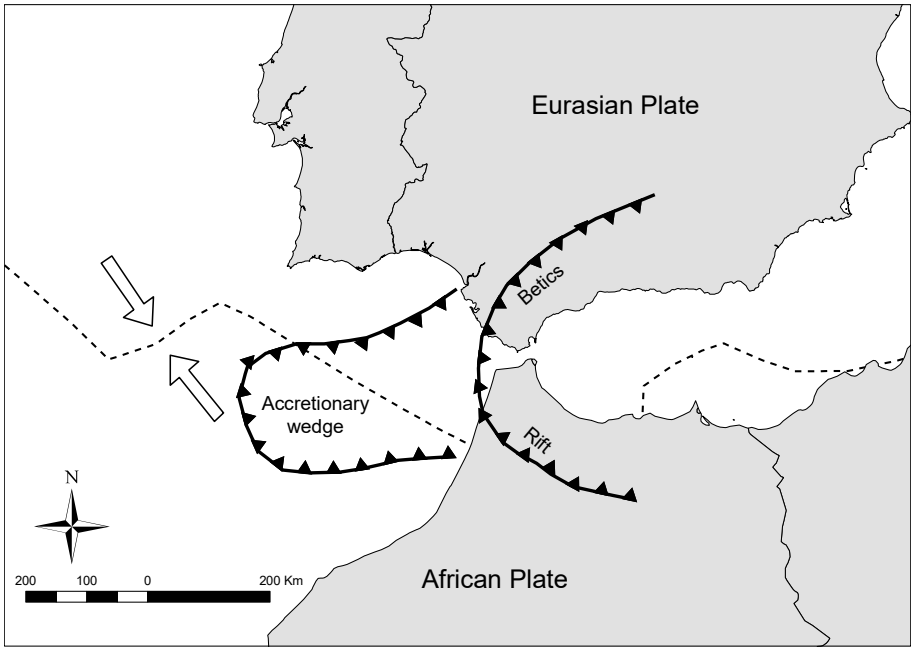


Figura 3.1: Límite entre la placa Euroasiática y la placa Africana

(Red de Estaciones de Referencia de Valencia, situada al Este de España). (Ver fig. 3.2.)

En este trabajo se utilizarán series de datos de 54 estaciones GPS pertenecientes a la red SPINA. Para cada una de estas estaciones se dispone de una observación diaria en cada componente Este, Norte y Altitud (ENU) entre los años 2011 y 2013. Estas series de datos se convertirán en muestras de funciones continuas de posicionamiento.

Las series temporales geodésicas han sido obtenidas mediante procesado respecto a la estación de referencia de Villafranca (VILL, IGS network, Madrid, Spain), que no está afectada por el movimiento específico de la geodinámica propia de la confluencia de las placas. El resultado de este procesado nos proporciona unas coordenadas geodésicas precisas (sub-centimétricas) para cada estación y cada día, que vienen expresadas en un sistema topocéntrico local ENU.

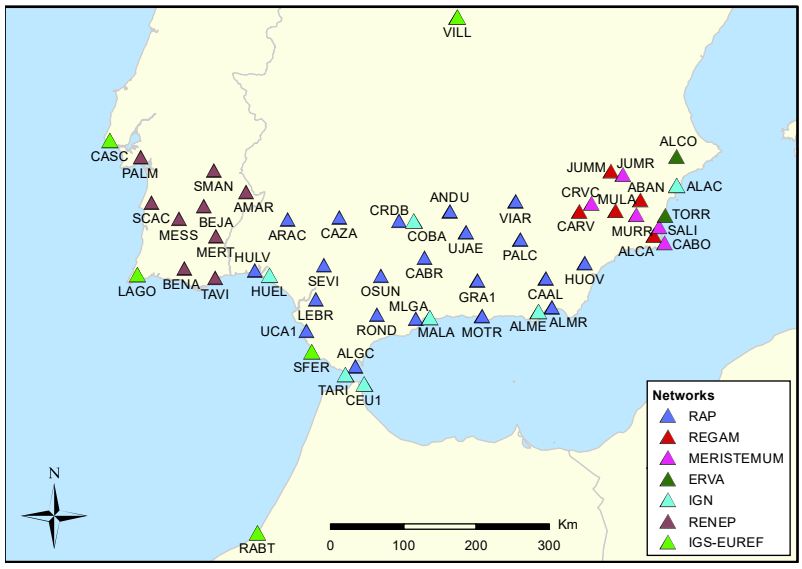


Figura 3.2: Red SPINA

3.3. Depuración y suavizado de los datos

En esta sección analizaremos las técnicas utilizadas para regularizar los datos. La aplicación de estas técnicas hará posible el uso posterior de procedimientos propios del análisis de datos funcionales. La figura 3.3 muestra un esquema de los procesos de regularización utilizados.

3.3.1. Filtrado 1 sigma y 2 sigma

El método usado normalmente para determinar valores atípicos en datos de tipo geodésico se basa en un ajuste de tipo lineal. Esta técnica se basa en el criterio de mínimos cuadrados ponderados, en el que se asignan pesos a las observaciones en función de sus errores. Se obtienen las series 1-sigma o 2-sigma (para $\sigma = 1$ o $\sigma = 2$), siendo este último el caso más frecuente.

Esta técnica es aplicada bajo la hipótesis de linealidad de la serie de datos, sin embargo la componente Altitud no es lineal y, aunque en principio el movimiento en el plano sí lo es, la ocurrencia de terremotos podría romper esta tendencia ([39]).

Nuestros datos iniciales se corresponden con el conjunto de coordenadas topocéntricas que indican el desplazamiento Norte, Este y en Altitud de las 54 estaciones de

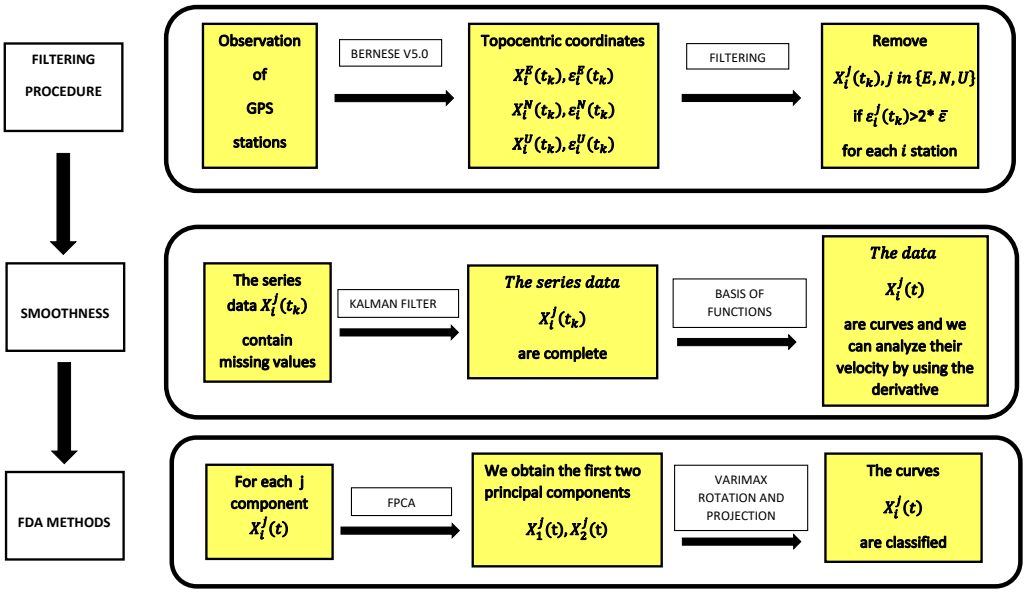
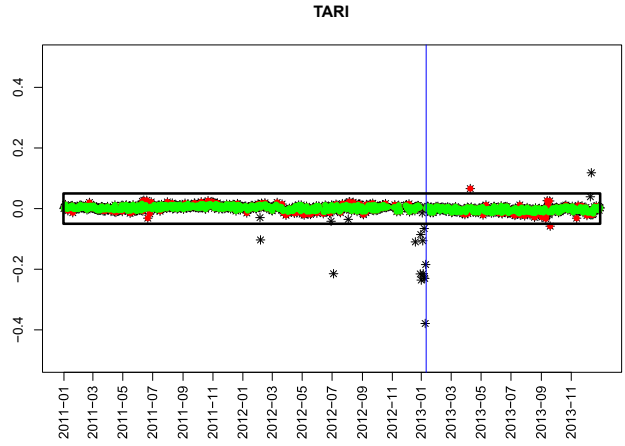


Figura 3.3: Regularización de los datos

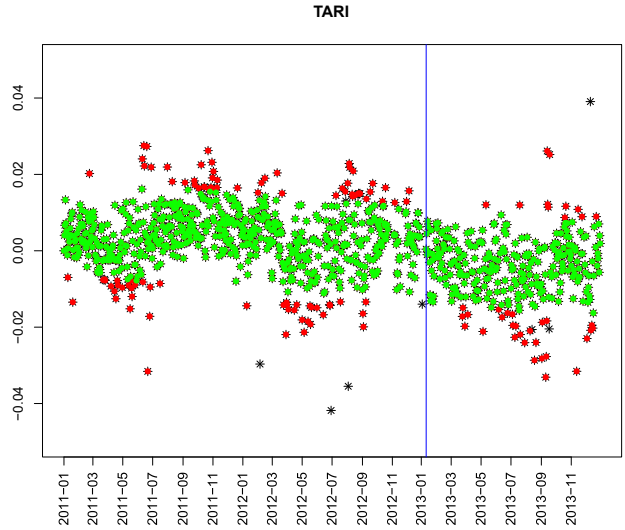
la red SPINA entre 2011 y 2013. Estas coordenadas han sido procesadas con el software Bernese (versión 5.0). El procesamiento con este software aporta al conjunto de datos una nueva variable que recoge el error en cada medición.

Comenzaremos planteando un modelo lineal, fijando $\sigma = 2$ ([43]), y compararemos los resultados con un nuevo método que propondremos a continuación basado en depuraciones locales. Como se ha comentado, el modelo lineal recoge bien el comportamiento de las componentes Norte y Este, pero no así la Altura; como consecuencia de ello, se estarían eliminando valores que no son verdaderos atípicos.

Las figuras 3.4 y 3.5 muestran el modelo de depuración lineal para $\sigma = 2$, en las estaciones GPS, TARI y UJAE para la componente Altitud. En estas figuras los valores que no han sido eliminados por el criterio lineal están en verde, mientras que el rojo representa a aquellos otros que no serían eliminados por el criterio que proponemos pero sí por el modelo lineal.

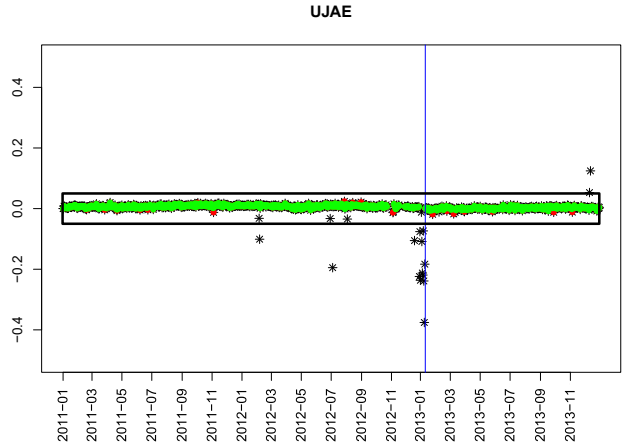


(a)

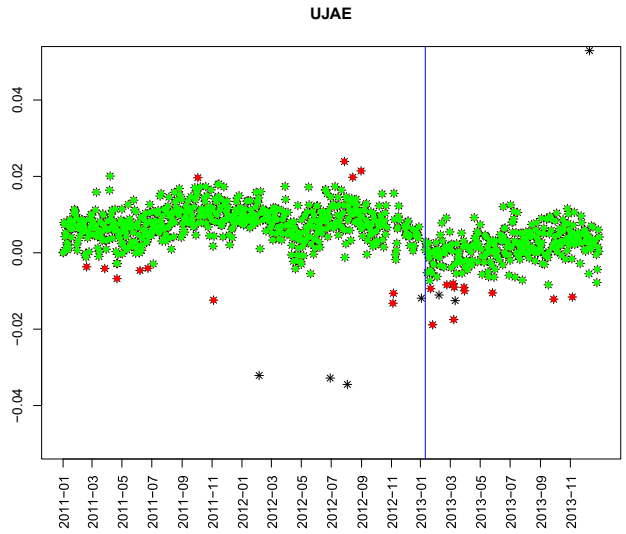


(b)

Figura 3.4: Comparación del procedimiento de filtrado en la componente Altitud para la estación GPS TARI. Los puntos rojos representa los datos que serían eliminados con el criterio lineal pero no con el modelo propuesto. Los valores representados en verde corresponden a los datos que no se eliminarían con el criterio lineal



(a)



(b)

Figura 3.5: Comparación del procedimiento de filtrado en la componente Altitud para la estación GPS UJAE. Los puntos rojos representa los datos que serían eliminados con el criterio lineal pero no con el modelo propuesto. Los valores representados en verde corresponden a los datos que no se eliminarían con el criterio lineal

Como se ha comentado, aunque el movimiento en el plano puede considerarse lineal, los movimiento sísmicos podrían romper esta tendencia y, por ello, al considerar el filtro basado en el modelo lineal podría eliminarse también información relevante sobre los eventos sísmicos producidos.

La figura 3.6 muestra el número de eventos sísmicos acontecidos entre 2011 y 2013 en Andalucía. Entre diciembre de 2012 y marzo de 2013 el número de eventos sísmicos es muy elevado. Comparando esta figura con las figuras 3.4 y 3.5 podría pensarse que se podría estar eliminando información sísmica relevante al considerar el filtro lineal.

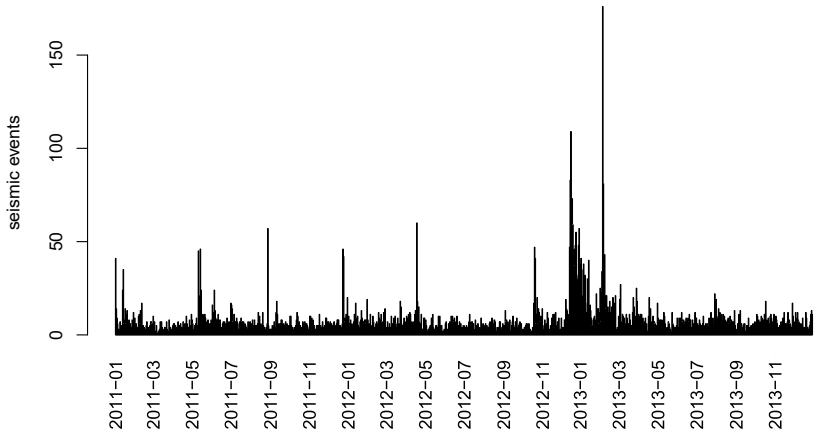


Figura 3.6: Número de eventos sísmicos producidos entre 2011 y 2013 en Andalucía. Puede observarse un incremento considerable de la actividad sísmica entre diciembre de 2012 y marzo de 2013

Por los motivos presentados se propone en este trabajo el uso de un criterio de eliminación de valores atípicos que no asuma la hipótesis de linealidad de los datos. La propuesta que realizamos se basa en el uso de la variable, proporcionada por el software Bernese, que mide el error de cada estimación, eliminándose para cada estación GPS aquellos datos cuya medida de error supere dos veces el error medio de la estación.

Con el nuevo criterio que proponemos el número de datos eliminados sería similar para cada componente, $1,389$, $1,429$, y $1,402$, respectivamente para Este, Norte y Altura. Estas cantidades se corresponden con el 2,6 %, 2,7 %, y 2,6 % sobre el total.

Sin embargo, con el criterio lineal los datos eliminados serían 1081, 985, y 1815, que se correspondonderían con el 2 %, 1,8 %, y 3,4 % de los datos. En este caso la cantidad de datos depurados sería mucho mayor en la componente Altitud, al no presentar ésta un comportamiento lineal.

3.3.2. Imputación Kalman

Las series de datos GPS originales contenían valores ausentes. Este conjunto de valores, que identificamos con el código NA (Not Available) ha aumentado una vez eliminados los valores atípicos con el criterio anteriormente descrito. El siguiente paso en la regularización de los datos es el uso de un filtro Kalman para imputar los valores ausentes.

En el caso en el que nos encontramos es recomendable el uso de un filtro Kalman de orden 2. La razón es que éste se comporta mejor con series de datos en los que hay un intervalo más o menos grande de valores omitidos, y con datos GPS éste es un problema habitual. La figura 3.7 ilustra el caso de la componente Este de la estación GPS MULA.

En resumen, el filtro Kalman nos proporcionará un primer suavizado discreto de los datos, así como la imputación, tanto de los datos originalmente faltantes como de los depurados en la fase de filtrado.

La figura 3.8 muestra los resultados obtenidos tras aplicar un filtro Kalman de orden 2 sobre las tres componentes en todas las estaciones de la red SPINA.

3.3.3. Suavizado mediante una base de funciones

La obtención de la forma funcional completará el proceso de suavizado de los datos y permitirá la posterior aplicacion de técnicas FDA. En primer lugar nos plantearemos cual es la base de funciones más apropiada para este conjunto de datos, así como la elección del número óptimo de funciones base y del parámetro de penalización de suavizado.

Para responder a la primera cuestión, la elección de la base más adecuada, probaremos con los tres ajustes más comunes, Splines, Fourier y Wavelets y compararemos los resultados obtenidos a través del error cuadrático medio MSE (mean square error).

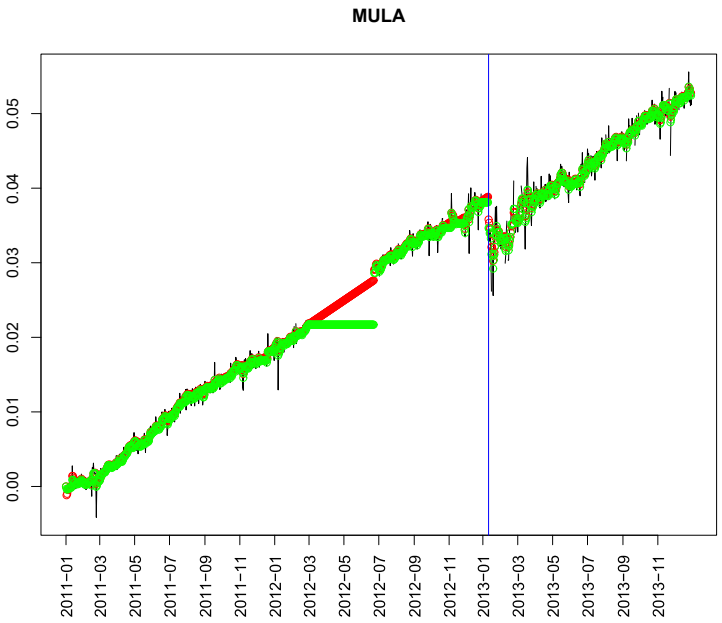


Figura 3.7: Comparación entre un filtro Kalman de orden 1 (verde) y de orden 2 (rojo) para la componente Este en la estación MULA. El filtro Kalman de orden 1 parece ineficiente cuando existen saltos en los datos

El ajuste de las tres componentes en las tres bases puede verse en la figura 3.9. Se aprecia en esta figura como las máximas diferencias entre las curvas de Fourier y los suavizados p-splines se producen al comienzo y al final, debido al comportamiento anómalo del suavizado de Fourier. Por esta razón, al comparar los tres ajustes a través del coeficiente MSE, los 35 primeros y últimos datos de las series no serán utilizados. Los resultados obtenidos para el coeficiente en las tres bases pueden verse en la tabla 3.1. En la figura 3.10 una comparativa de los distintos ajustes a través de los valores MSE obtenidos en las distintas estaciones GPS.

En los resultados de la tabla 4.1 puede verse como, tras eliminar los extremos de las series, los errores en las series Este y Norte son algo mayores al usar la base de Fourier. Sin embargo en la componente Altitud los errores son similares en las tres bases, siendo incluso algo menor en la base de Fourier. Esto es fácil de entender por las fluctuaciones que se producen en esta serie. De todas formas, en general los mejores resultados son producidos con la base de p-splines y por ello esta base será la elegida para realizar el suavizado.

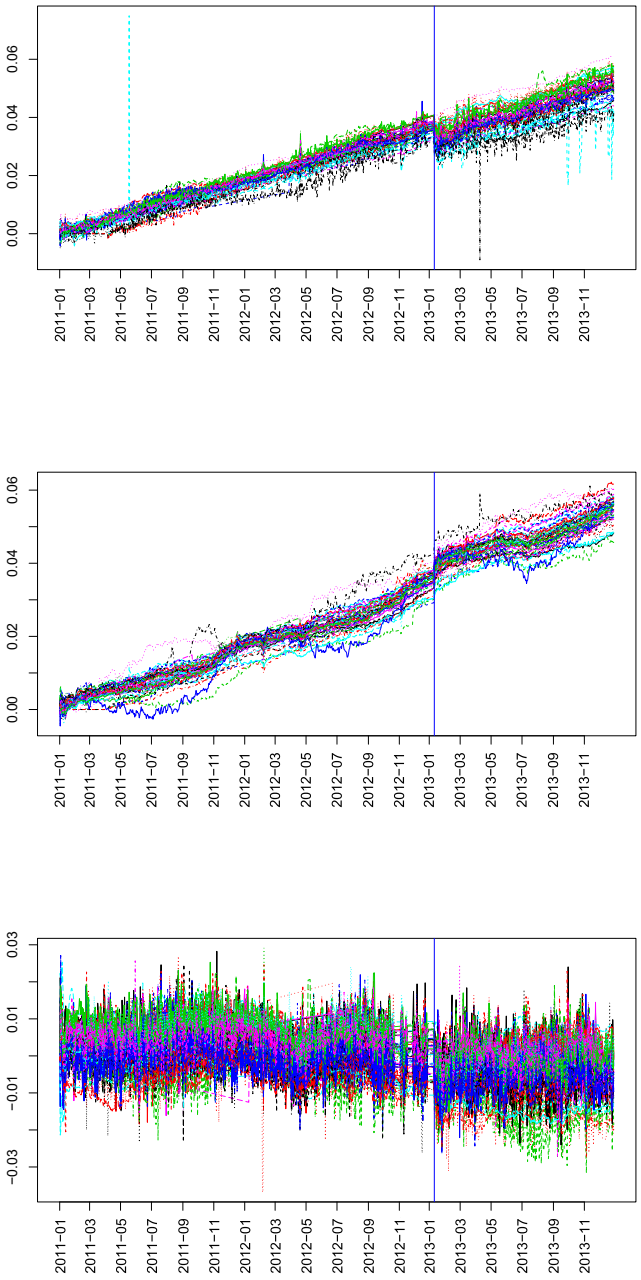


Figura 3.8: Datos depués de aplicar un filtro Kalman de orden 2 a las componentes Este (a), Norte (b) y Altitud (c). Se aprecia un comportamiento más homogéneo en las componentes Este y Norte

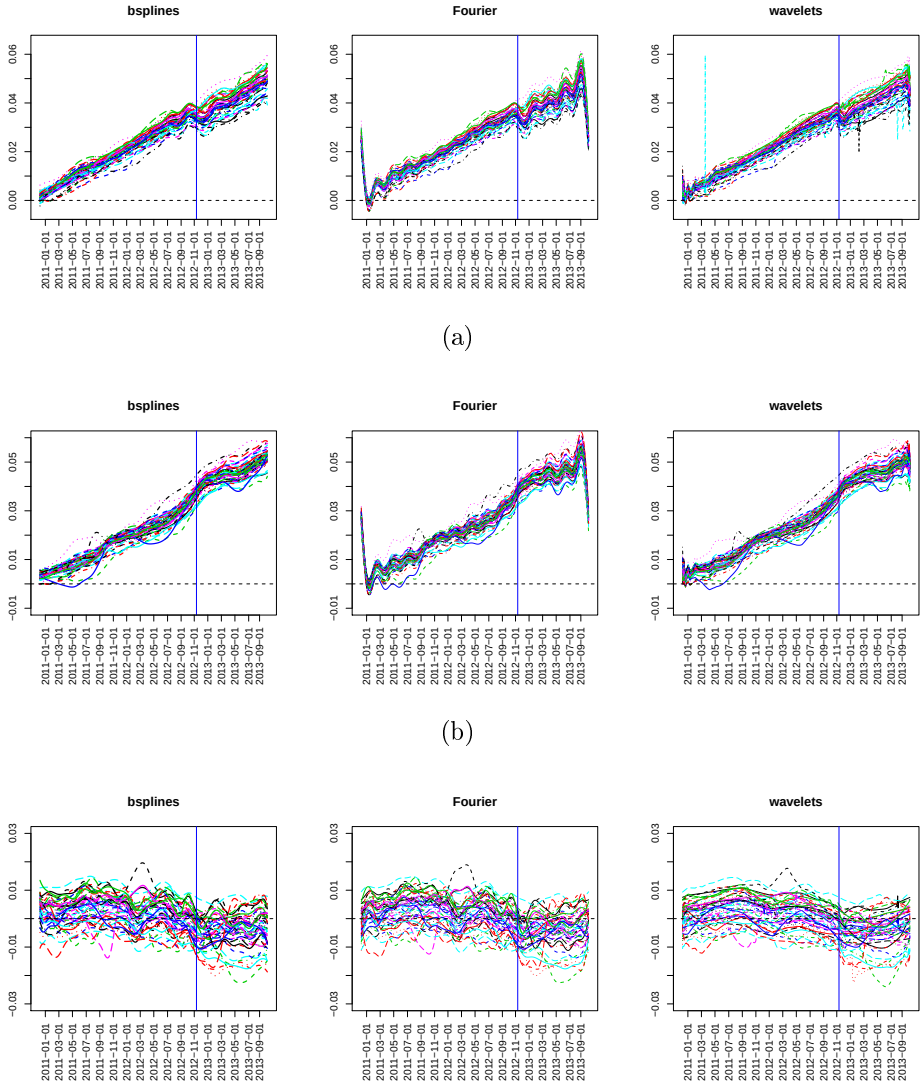
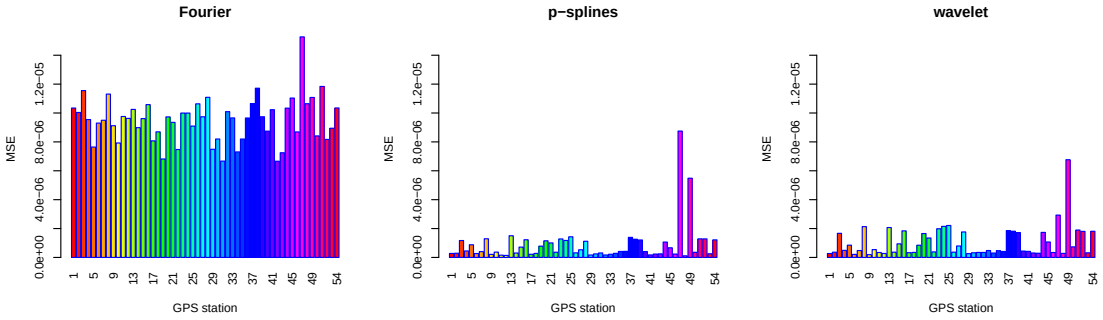
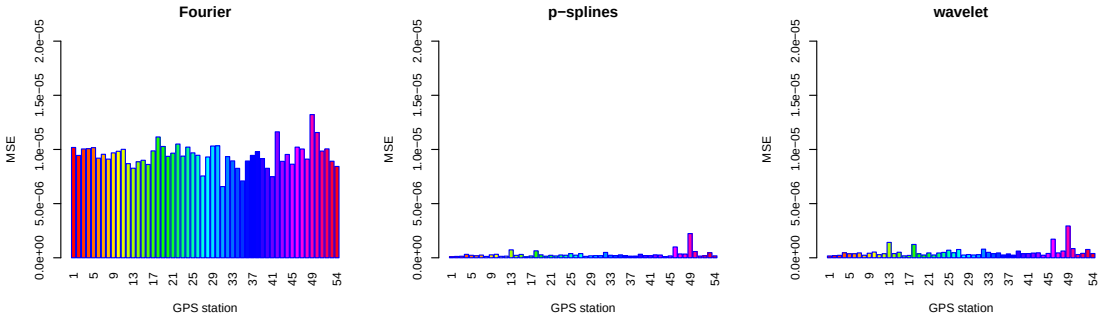


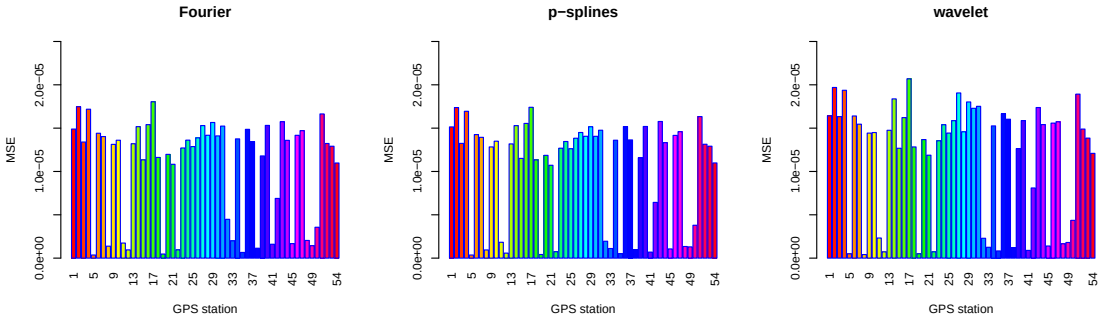
Figura 3.9: Comparación del proceso de suavizado en las bases de p−splines, Fourier y wavelets en las componentes Este (a), Norte (b) y Altitud (c). Se aprecia una tendencia lineal en todas las estaciones, en las componentes norte y este, sobre todo en la base de splines y en la de wavelets. Las curvas obtenidas con las bases p−splines y de wavelets son similares, el problema sería que al usar wavelets las curvas no serían diferenciables. Al utilizar la base de Fourier las curvas presentan un comportamiento periódico, como se esperaba. Por otro lado, la componente altitud, en las tres bases, presenta un comportamiento errático y heterogéneo



(a)



(b)



(c)

Figura 3.10: Comparativa en los distintos ajustes de valores MSE, obtenidos por estación, en las componentes Este (a), Norte (b) y Altitud (c)

Cuadro 3.1: MSE en cada estación GPS después de eliminar los primeros y los últimos valores de las series

Este	amp; mean	amp; sd
Fourier	amp; $9{,}5 * 10^{-6}$	amp; $1{,}5 * 10^{-6}$
p-splines	amp; $8{,}7 * 10^{-7}$	amp; $1{,}3 * 10^{-6}$
wavelets	amp; $1{,}1 * 10^{-6}$	amp; $1{,}1 * 10^{-6}$
Norte	amp; mean	amp; sd
Fourier	amp; $9{,}5 * 10^{-6}$	amp; $1{,}1 * 10^{-6}$
p-splines	amp; $3{,}0 * 10^{-7}$	amp; $3{,}1 * 10^{-7}$
wavelets	amp; $5{,}1 * 10^{-7}$	amp; $4{,}4 * 10^{-7}$
Altitud	amp; mean	amp; sd
Fourier	amp; $1{,}0 * 10^{-5}$	amp; $5{,}8 * 10^{-6}$
p-splines	amp; $1{,}0 * 10^{-5}$	amp; $5{,}9 * 10^{-6}$
wavelets	amp; $1{,}2 * 10^{-5}$	amp; $5{,}9 * 10^{-6}$

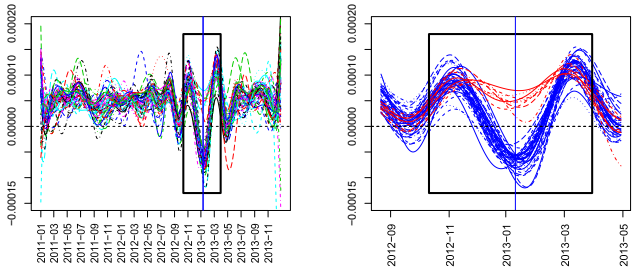
3.4. Análisis de los datos

3.4.1. Análisis de las curvas derivadas

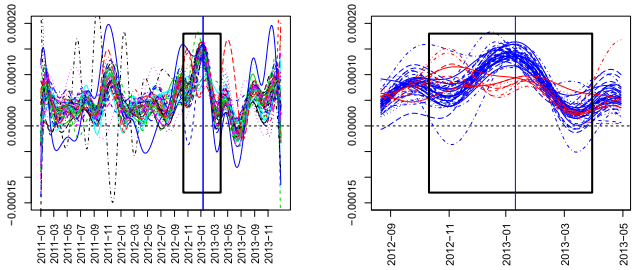
Bajo la hipótesis inicial de que las estaciones GPS situadas a un mismo lado de la zona de subdución tendrán una velocidad de desplazamiento similar, nos planteamos el análisis del conjunto de curvas derivadas. Al representar estos elementos (Figura 3.11) se observan dos aspectos relevantes. El primero es que en el periodo comprendido entre Noviembre de 2012 y Marzo de 2013 se produce una clara sincronización entre las curvas en el plano horizontal (Este-Norte). El segundo aspecto llamativo es que en las componentes Este y Norte se distinguen dos grupos de curvas con un comportamiento diferenciado. Se han representado en rojo las curvas correspondientes a la red portuguesa y en azul al resto de estaciones.

Para analizar la existencia de estos distintos grupos de curvas y analizar luego su posible relación con la zona de subducción, aplicaremos Componentes principales funcionales (FPCA) sobre las curvas derivadas.

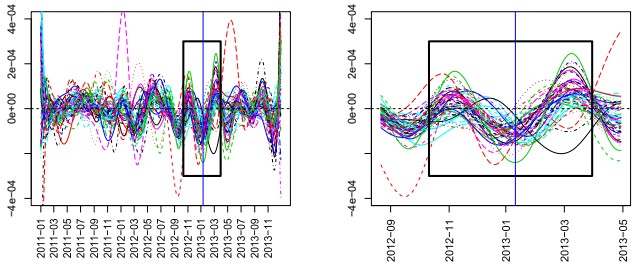
En la figura 3.12 se pueden ver las dos primeras componentes principales de cada coordenada. Puede comprobarse como éstas recogen alrededor del 85 % de la variabilidad total explicada, siendo la componente Este la que presenta un porcentaje



(a)



(b)



(c)

Figura 3.11: Curvas derivadas en las componentes (a) Este, (b) Norte y (c) Altitud. Entre Noviembre de 2012 y marzo de 2013 hay una sincronización de las curvas en el plano horizontal (Este-Norte). Las figuras de la derecha muestran la misma representación, en una ventana ampliada. Se pueden ver un comportamiento diferenciado en el plano horizontal en dos grupos de curvas: en rojo, las curvas de las estaciones portuguesas y en azul el resto de las estaciones

de inercia más elevado. Además esta componente Este presenta un único cambio de variabilidad, observable en el comportamiento de la segunda componente. Otro aspecto a tener en cuenta es que este cambio de variabilidad sucede durante un periodo en que el número de movimientos sísmicos fue elevado. Además, debido a la situación de las placas tectónicas en la zona, es fácil suponer que el desplazamiento hacia el Este registrará las mayores diferencias en las estaciones GPS ubicadas a ambos lados de la zona de subducción.

Por todos estos motivos, la componente Este será analizada con más profundidad. Se aplica en primer lugar una rotación VARIMAX sobre las componentes de esta coordenada Este y a continuación los datos son proyectados sobre las dos primeras componentes principales de las curvas derivadas. El resultado puede verse en la figura 4.16, que representa la distribución geográfica obtenida sobre el mapa de la red SPINA; se observan cuatro grupos de curvas diferenciados, siendo uno de ellos el correspondiente a las estaciones de la red Portuguesa, aunque el grupo verde, conformado por las estaciones 15 y 31 se comporta de manera atípica, correspondiéndose con las estaciones CABR y MALA.

3.5. Resultados

Como hemos visto existen campos experimentales como ocurre en el campo de la geodesia en el que los elementos que se analizan vienen determinados por curvas. Sin embargo, en la práctica, por problemas obvios de accesibilidad a dichos datos, se dispone únicamente de una discretización de las curvas en un periodo $[t_{min}, t_{max}]$. La disponibilidad de medios tecnológicos para recolectar este tipo de datos en una rejilla suficientemente fina, junto con la capacidad de poder almacenarlas y tratarlos computacionalmente, nos permite proponer soluciones eficientes para este tipo de problemas; dichas soluciones pasan, generalmente, por la adaptación de las técnicas vectoriales clásicas al caso funcional.

En este sentido, tras establecer una metodología adecuada para depurar los datos geodésicos que no eliminara datos que puedieran ser relevantes, se han obtenido las curvas que van a representar los desplazamientos de la estaciones GPS de la red SPINA, en este caso, en una base de p-splines.

El análisis de estas curvas a traves del análisis ACPF ha permitido determinar dos grupos diferenciado de curvas que podrían pertenecer a ambos lados de la zona de subducción, que era el objetivo de este trabajo.

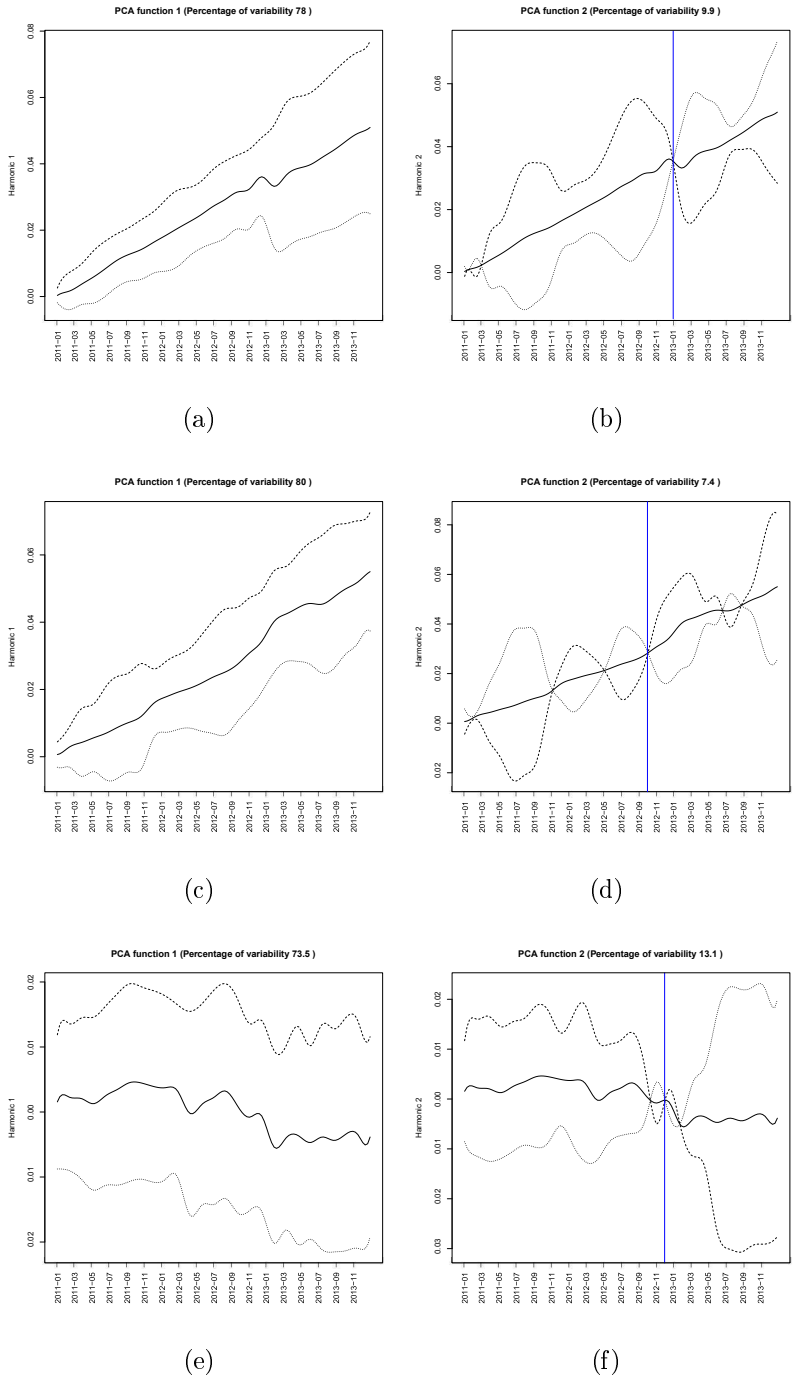


Figura 3.12: Las dos primeras componentes principales sobre las curvas derivadas en las componentes Este ((a) y (b)), Norte ((c) y (d)) y Altitud ((e) y (f)). La línea azul en los gráficos de la derecha marca un cambio de variabilidad

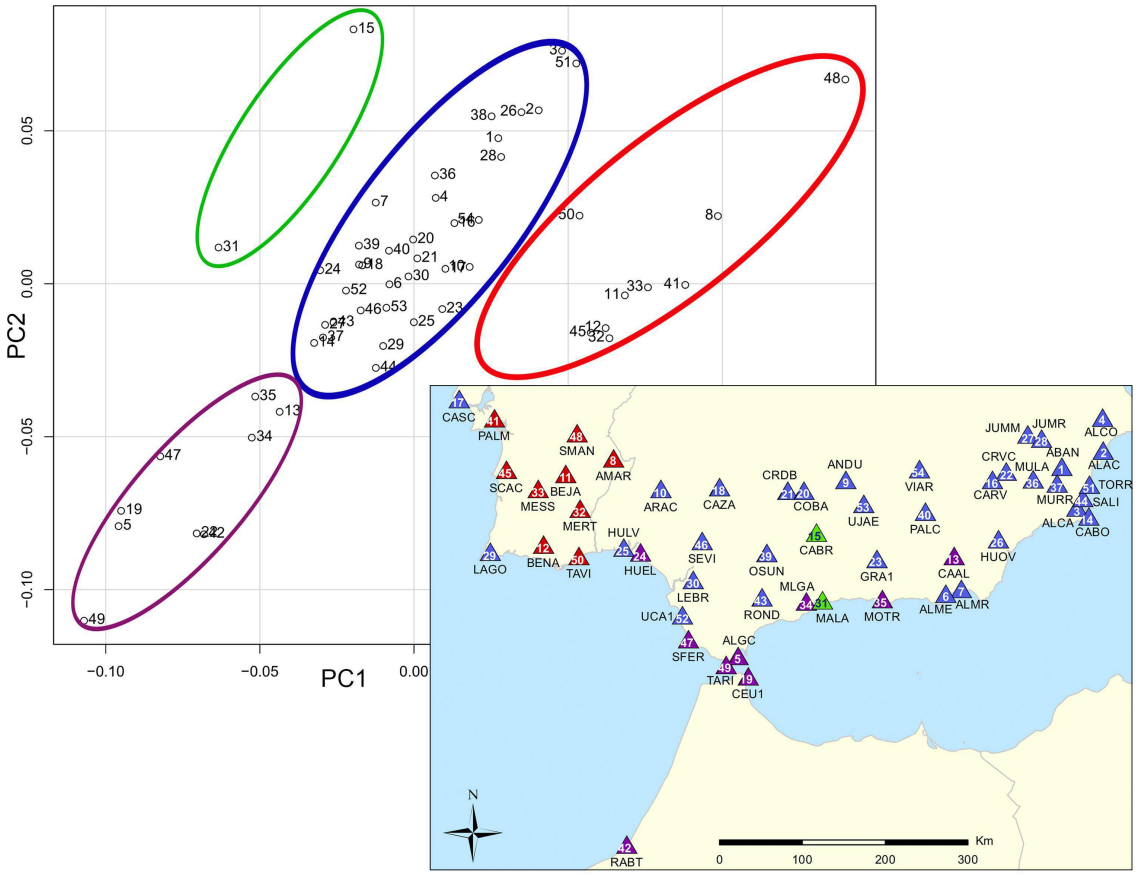


Figura 3.13: Proyecciones sobre las dos primeras componentes principales de las curvas derivadas para la componente Este, después de aplicar una rotación VARIMAX, y la distribución geográfica de los grupos obtenidos sobre la red SPINA. Los colores de las estaciones representan los diferentes grupos. Pueden verse tres grupos diferenciados y dos estaciones con un comportamiento atípico que podrían formar parte de otro grupo

Además el análisis de las curvas derivadas ha permitido detectar la sincronización de dos grupos de curvas en un periodo de tiempo en el que el número de eventos sísmicos era muy elevado.

Capítulo 4

Relación funcional entre deformación y sismicidad en El Hierro

4.1. El Hierro

El Hierro es la isla más occidental del archipiélago de las Canarias. La formación de este archipiélago comenzó hace más de 40 millones de años ([45, 46, 47]). No está claro si su origen está relacionado con modelos basados en un punto caliente ([46]) o con intrusiones magmáticas controladas por compresión regional intercambiable y regímenes tectónicos de extensión ([47]). La morfología subaérea de El Hierro (27.7 °N; 18.0 °W; 278.5 km^2 y 1.501 m de altura máxima) está formada por de tres brazos ([46]) con abundantes diques ([48]) y grandes cicatrices de deslizamientos de tierra ([49]) (Ver Fig. 4.1).

De acuerdo con Stroncik ([50]), la actividad volcánica en El Hierro está controlada por una compleja variedad de pequeõs reservorios de magma, aislados a profundidades del manto.

La erupción submarina ocurrida en la isla de El Hierro el 10 de octubre de 2011, supuso la reactivación del proceso eruptivo, después de más de 40 años de inactividad en las Islas Canarias. Según Hernandez ([51]) y Carracedo ([52]), la última erupción conocida en El Hierro sucedió hace alrededor de 200 años. Las primeras señales del proceso eruptivo analizado en este trabajo se detectaron en Julio de 2011. Se produjo entonces un incremento en la sismicidad, un desplazamiento significativo en las estaciones GNSS-GPS y comenzó la detección de CO_2 difuso. Aunque la erupción

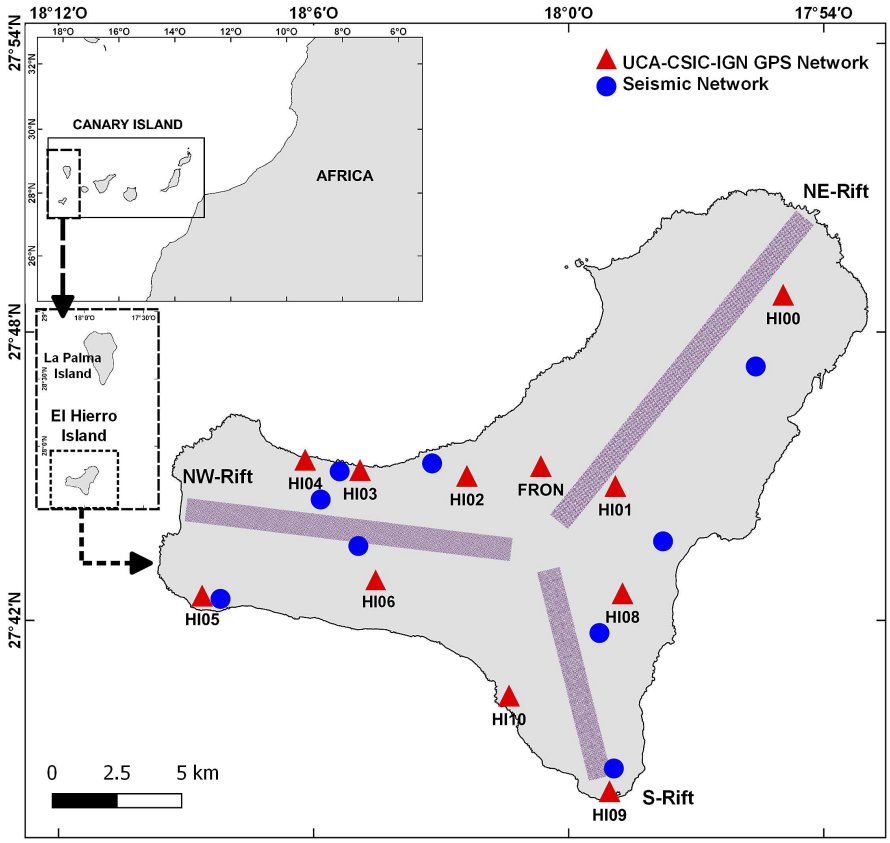


Figura 4.1: Mapa de El Hierro. En la figura pueden verse en rojo las estaciones GPS de la red UCA-CSIC-IGN, entre las que se encuentra la estación FRON y en azul los puntos de control de la red sísmica de El Hierro

submarina se detuvo en marzo de 2012, nuevas señales similares a las que comenzaron en julio de 2011 volvieron a ser detectadas. Estos nuevos episodios liberaron mas energía sísmica y/o causaron deformaciones varios centímetros mayores a las producidas por el episodio anterior ([53]).

Entre 2011 y 2014 hubo al menos siete episodios de intrusión magmática en el Hierro, pero sólo el primero (2011-2012) llegó a una erupción submarina ([53, 54]). La erupción, que duró casi 5 meses, comenzó el 10 de octubre de 2011 y fue precedida por casi 3 meses de inquietud. Sin embargo, la actividad magmática en la isla no cesó tras la erupción ya que, entre junio de 2012 y marzo de 2014, tuvieron lugar al menos seis episodios de intensa sismicidad y deformación. La Figura 4.2 muestra la gran cantidad de movimientos sísmicos que se produjeron entre 2011 y 2014 y su

ubicación.

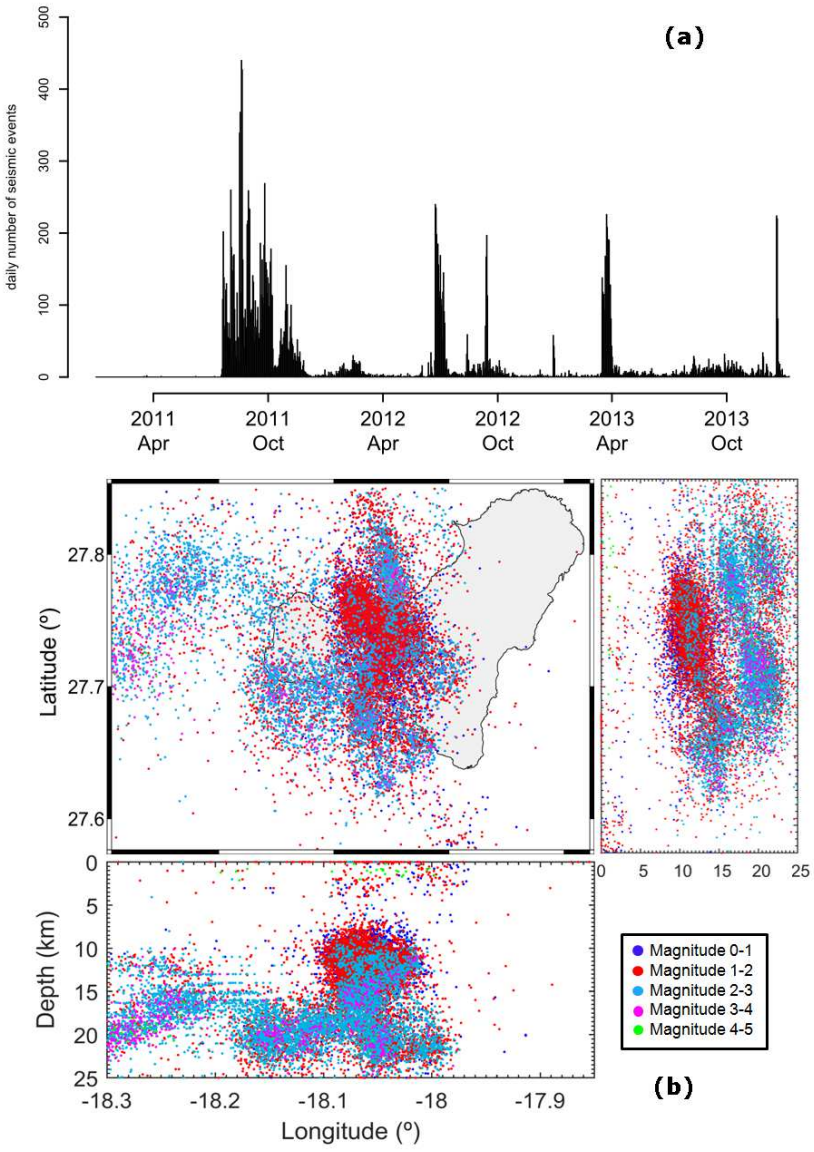


Figura 4.2: Número diario de eventos sísmicos y su localización 3D sobre la isla de El Hierro entre 2011 y 2014

No es sorprendente que el proceso de intrusión magmática afecte profundamente tanto al magma ascendente como a la roca circundante. Las variaciones en la cristalinidad del magma, la composición, la temperatura, el contenido volátil y la

vesicularidad producidas después de la erupción fueron lo suficientemente importantes como para causar cambios significativos en la viscosidad del magma, que debería haber causado, a su vez, cambios significativos en la reología y dinámica del magma ([55]). El efecto del proceso de intrusión en la roca circundante es igualmente profundo. Para acomodar el magma intruso, la roca anfitriona debe deformarse. La magnitud de la deformación de la roca huésped es inevitable y los modelos existentes generalmente coinciden en que el efecto debería ser medible en la superficie en muchos casos ([56]).

4.2. Deformación superficial y sismicidad

La deformación superficial ha sido un fenómeno esencial para describir el comienzo y la evolución del proceso eruptivo ocurrido en la isla de El Hierro, así como para pronosticar cambios en la actividad sísmica y volcánica durante un periodo de crisis ([57]).

A partir de las observaciones GNSS-GPS analizadas, se ha podido detectar rápidamente la reactivación, debido al cambio producido en la deformación de la geodinámica regional de la isla. Además, se ha podido observar que los cambios en la deformación superficial se han producido antes que la actividad sísmica, convirtiéndose así el parámetro deformación superficial en un importante precursor de la actividad volcánica ([58, 53]).

Esta relación ha sido analizada de manera puntual en [54]. Nuestro objetivo es analizar globalmente dicha relación, teniendo en cuenta las diferentes fases sísmicas incluidas en el periodo de tiempo analizado (2011-2014).

Así, el objetivo central de este capítulo será evaluar, a través del uso de la correlación funcional, cómo la deformación superficial puede explicar la actividad sísmica en la isla de El Hierro. Para ello se introduce una medida de similitud funcional que será aplicada tanto al análisis global, como a un análisis más localizado, realizado por fases. En este último, se trata de evitar la mezcla de fenómenos sísmicos que podrían afectar al análisis global.

Otro aspecto relevante de este trabajo es el análisis del periodo de tiempo en el que la deformación superficial precede a la actividad sísmica, por la importancia que este fenómeno presenta en problemas reales. También será analizado este aspecto tanto de manera global como por fases. Después del largo período sin actividad sísmica en la isla de El Hierro comprobamos que el lapso de tiempo entre el inicio del proceso

de deformación y el comienzo de la actividad sísmica es de aproximadamente un mes. Propondremos, en caso de producirse una situación similar, un método para predecir en tiempo real el comienzo de la actividad sísmica. Este sistema de alerta, estará basado en los cambios producidos en las curvas derivadas cuando hay un rápido descenso en la curva de deformación.

4.3. Datos. Procesado

Para alcanzar los objetivos planteados, disponemos de dos conjuntos de datos. Por un lado, para medir la deformación superficial se manejan datos de observaciones GNSS-GPS procedentes de la estación FRON (GRAFCAN), ubicada en el municipio de Frontera. En este caso, se mide en sesiones de 4 horas, la variación en distancia con la estación de referencia LPAL, situada en la isla de La Palma. Estos datos han sido procesados a través del software Bernese v5.0 ([38]). En la estrategia empleada, se usan métodos de dobles diferencias de la fase de la portadora, en modo de retardo ionosférico libre en el proceso de estimación de parámetros. Los errores troposféricos se modelan mediante el uso de una combinación del modelo Saastamoinen a priori ([59]) y las funciones de mapeo de Niell ([60]). Los parámetros troposféricos se calculan por horas y las ambigüedades se vuelven a resolver para cada baselínea de forma independiente, utilizando el observable de ionósfera libre con un modelo ionosférico a priori para determinar la ambigüedad de amplio camino ([61]). También se considera la carga de la marea oceánica proporcionada por el Observatorio de Onsala. Las ecuaciones normales se determinan para cada solución diaria. Finalmente, la solución anual se obtiene combinando las ecuaciones normales diarias en la época media de cada campaña, utilizando las órbitas precisas del IGS ([62]) y considerando como punto de referencia la estación LPAL en la isla de La Palma, con respecto el marco de referencia ITRF2008 ([63]).

Por otro lado, para evaluar la actividad sísmica se ha considerado el catálogo de terremotos del Instituto Geográfico Nacional (IGN). Haciendo uso de las mediciones diarias proporcionadas por este catálogo y la relación entre la energía diaria acumulada y la magnitud absoluta de los terremotos en la escala de Ritcher, establecida por Choy ([64]):

$$E = 10^{1,5*Mg+4,4}, \tag{4.3.1}$$

se ha obtenido la serie de energía diaria acumulada.

Ambas series de datos, de desplazamiento y de energía acumulada, han sido tratadas hasta obtener datos apropiados para ser analizados desde un punto de vista funcional.

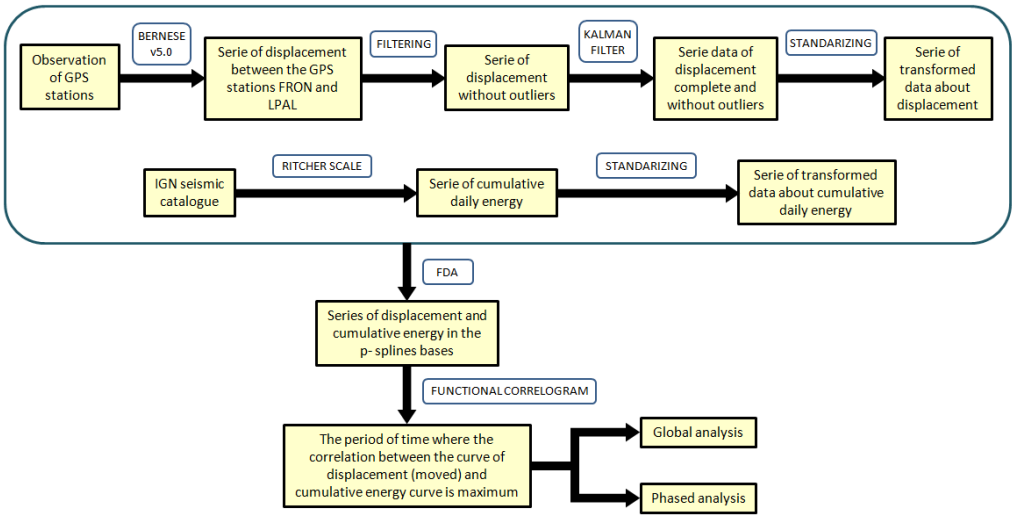
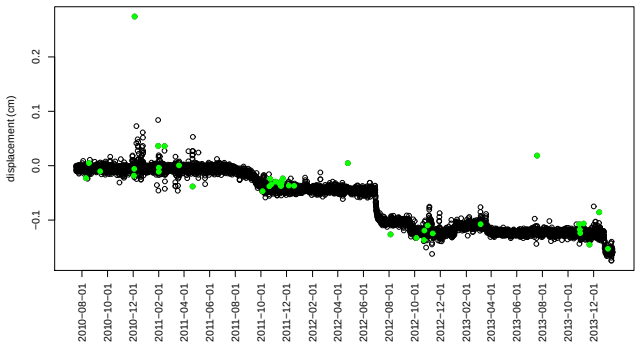


Figura 4.3: Metodología utilizada en la regularización de las series de datos sobre desplazamiento y energía sísmica acumulada

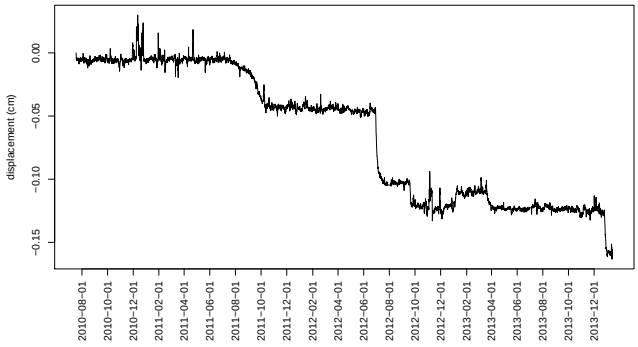
En el esquema mostrado en la figura 4.3 puede observarse el conjunto de procedimientos utilizados sobre estas series de datos.

En el caso de la serie de desplazamiento, el primer paso en el proceso de limpieza ha sido la eliminación de valores atípicos. Al tratarse de nuevo de datos sobre estaciones GPS, el criterio considerado ha sido el mismo usado en el capítulo anterior, que considera como anómalos aquellos datos cuyo error supere dos veces el error medio de la estación. El resultado puede verse en la figura 4.4 (a). La eliminación de atípicos se ha completado de nuevo con un procedimiento de imputación realizado por un filtro Kalman de orden 2 (4.4 (b)).

La serie sobre energía diaria acumulada no ha necesitado un proceso de eliminación ni de reconstrucción. Sin embargo ambas series han sido posteriormente estandarizadas para permitir su comparación. La figura 4.5 muestra la serie original sobre energía sísmica acumulada. En la figura 4.6 se pueden ver ambas series una vez tratadas y normalizadas.



(a)



(b)

Figura 4.4: a) Valores eliminados como atípicos y b) datos con filtro Kalman

4.4. Fases y subfases. Análisis del punto de cambio

Al estudiar la relación existente entre la energía acumulada y la deformación superficial es importante considerar los fenómenos sísmico-volcánicos existentes en el periodo de tiempo analizado. Teniendo en cuenta estos fenómenos, cuatro fases han sido detectadas en este periodo. Para determinar las diferentes fases, se ha considerado que cada fase sólo contenga un fenómeno sismo-volcánico relevante y que cada fase finalice el día de ocurrencia del sismo de mayor magnitud.

Las fases obtenidas al considerar estos aspectos pueden verse en la figura 4.7. En este gráfico están además representados los sismos de magnitud $> 2{,}5$. En particular

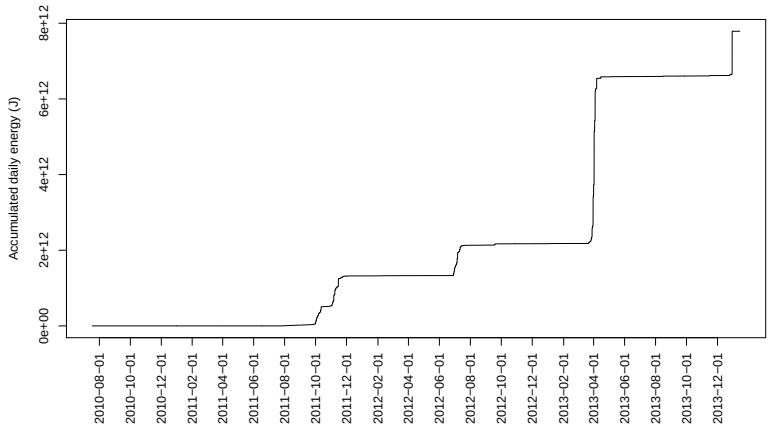


Figura 4.5: Serie original de energía sísmica acumulada entre 2010 y 2013

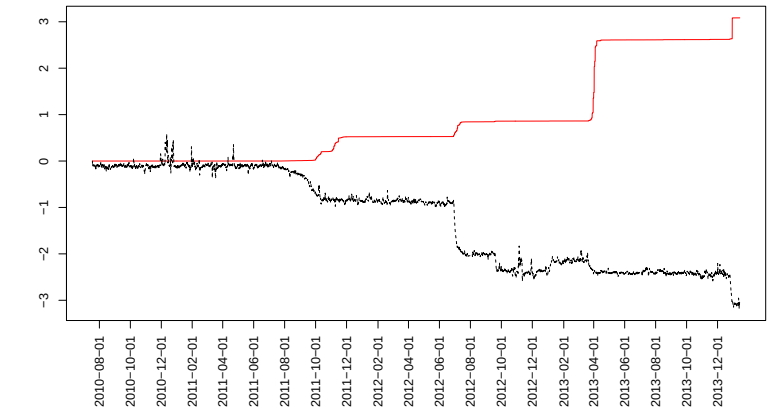


Figura 4.6: En la figura pueden verse las curvas sobre energía sísmica acumulada (en rojo) y sobre desplazamiento (en negro) una vez tratadas y normalizadas

aparecen representados en rojo los sismos de magnitud superior a 4, en azul los superiores a 3 y en rosa los superiores a 2,5.

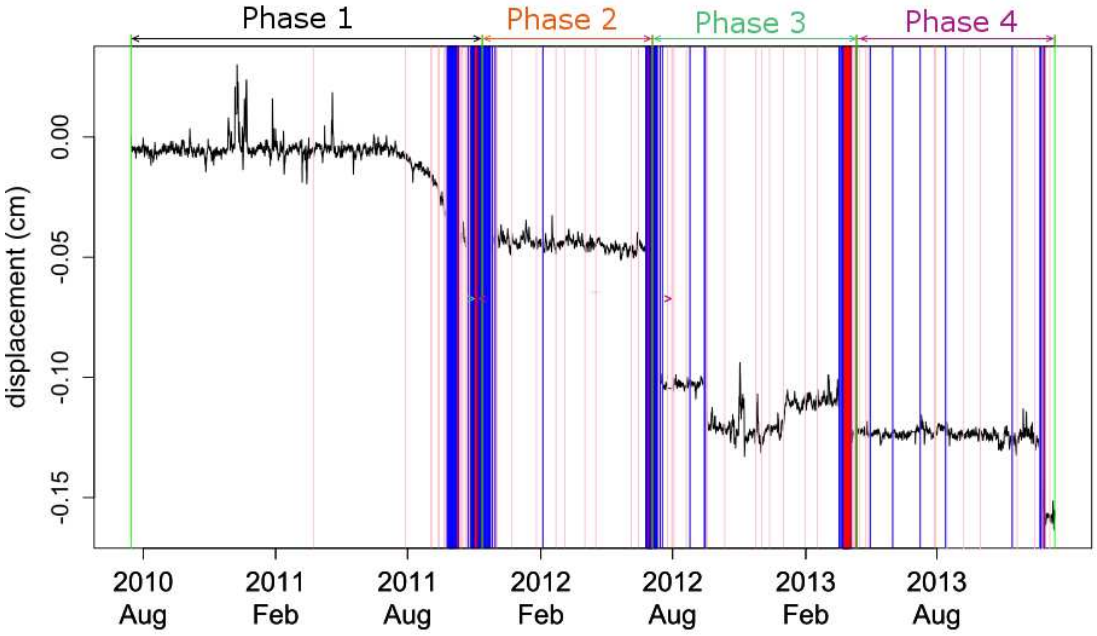


Figura 4.7: Serie de datos sobre desplazamiento depurada y normalizada en las 4 fases consideradas. En la imagen pueden verse en rojo los sismos de magnitud superior a 4, en azul los superiores a 3 y en rosa los superiores a 2,5 en la escala de Ritcher

Las principales características de estos periodos son:

Fase 1: Desde el comienzo de la serie de datos (14/07/2010) hasta el día 11/11/11.

La primera parte de este periodo está caracterizada por la ausencia de movimientos sísmicos y de deformación. Sin embargo, en agosto de 2011 la deformación superficial alcanza los 24 cm/año y la energía sísmica acumulada comienza a crecer. El 10 de octubre tiene lugar una erupción submarina y el 11 de noviembre se produce un terremoto de magnitud 4.6 que determina el fin de la fase 1.

Fase 2: Desde el fin de la fase 1 (12/11/11) hasta el 03/07/12.

Inicialmente la deformación se estabiliza y se producen algunos eventos sísmicos de magnitud inferior a 3.5. Desde final de junio hasta mediados de julio de 2012 ocurre una reactivación del proceso de deformación y se producen sismos de magnitud superior a 3.5. Este periodo acaba con un terremoto de magnitud 4.2 el 3 de julio.

Fase 3: Desde el fin de la fase 2 (04/07/12) hasta el 11/04/13.

Este periodo se caracteriza porque en él se produce un fenómeno de inflación- deflación entre septiembre de 2012 y enero de 2013. Esta fase acaba con una alta concentración de terremotos de magnitud superior a 4, siendo el último de magnitud 4.1 (11/04/13).

Fase 4: Desde el fin de la fase 3 (12/04/13) hasta el final de la serie de datos (11/01/14).

Despues de un largo periodo de estabilidad, en diciembre de 2013, se produce la reactivación del proceso de deformación. Un terremoto de magnitud superior a 5 ocurre en la isla el 27 de diciembre de 2013.

4.4.1. Subfases. Análisis del punto de cambio

El proceso de inflación-deflación de la tercera fase es un fenómeno inusual que podría ocultar la relación existente entre la deformación y la energía sísmica acumulada. Por ello es conveniente subdividir esta fase en subperiodos, de forma que sólo uno de ellos contenga el fenómeno y que éste no influya en el análisis planteado.

Existen diferentes métodos para segmentar una serie de datos ([65, 66]). En este caso, para que el fenómeno de inflación-deflación pueda ser aislado se utilizará la técnica de Análisis del punto de cambio ([66]). El objetivo de este método es detectar cuando se producen cambios significativos en la serie de datos.

Los métodos de detección de punto de cambio incluyen una gran cantidad de procedimientos destinados a dividir una serie de datos de forma óptima ([65, 66]).

En este trabajo, el procedimiento ha sido aplicado a través del paquete "changepoint" del programa *R*. En particular se ha utilizado un método que utiliza la segmentación binaria para identificar puntos de cambio en una serie dada, fijada una función de costo y de penalización. En *R*, se lleva a cabo a través del método "BingSeg". Los resultados obtenidos al aplicar este procedimiento a la serie de datos correspondiente a la tercera fase pueden verse en la gráfica 4.8.

Tras ello, la tercera fase queda dividida en 3 subfases a, b y c de forma que sólo la subfase b contiene al fenómeno de inflación-deflación.

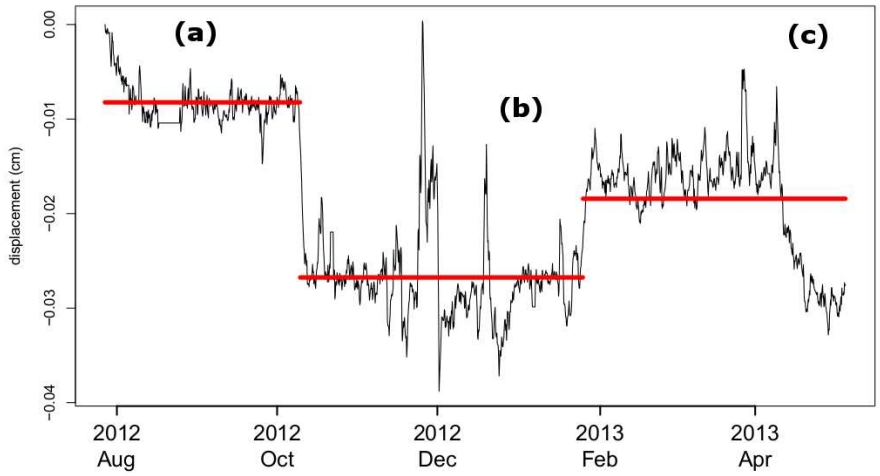


Figura 4.8: Resultado de aplicar el método de segmentación "BingSeg" sobre la tercera fase para encontrar con puntos óptimos que aislan el fenómeno de inflación-deflación en la subfase (b)

4.5. FDA. Correlación funcional

Para poder realizar un tratamiento y análisis funcional de los datos hay que ajustar las series de datos a un conjunto de curvas. El primer paso es seleccionar la base de datos funcionales más adecuada. En este caso la base elegida es la base de p-splines. El número óptimo de funciones base, así como el parámetro óptimo de suavizado han sido estimados a través del criterio de validación cruzada generalizada a través de la función *min.basis* del paquete *fda.usc* ([67]) del programa *R*.

Una vez aplicados los criterios anteriores sobre depuración de los datos, las curvas con parámetros óptimos obtenidas para los datos sobre desplazamiento y energía sísmica acumulada pueden verse en la figura 4.9.

Aplicando este mismo procedimiento sobre los datos divididos en fases e inicializando la energía acumulada en cada fase, se obtienen las curvas representadas en la figuras 4.10, 4.11, 4.12 y 4.13. A partir de estos datos funcionales trataremos de medir la relación entre sismicidad y deformación.

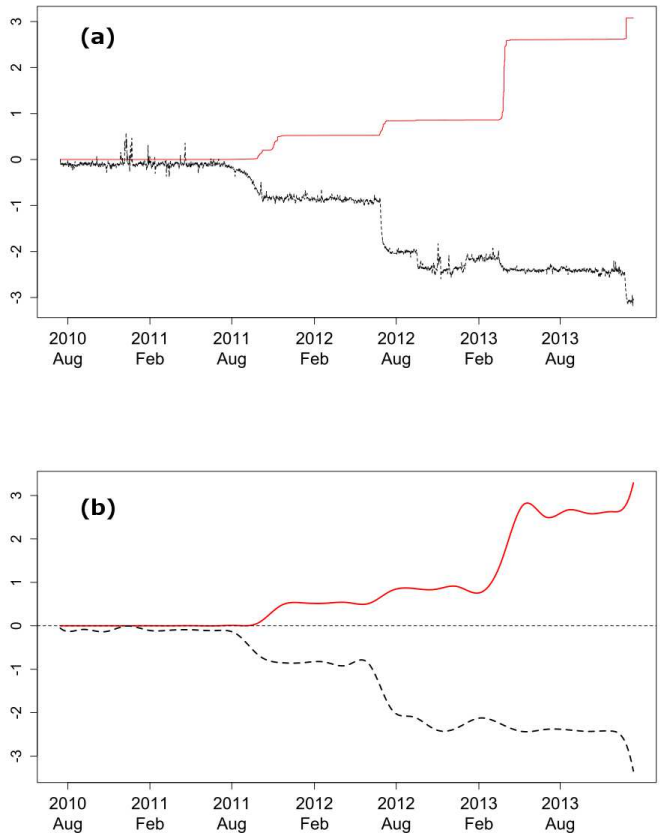


Figura 4.9: Series depuradas y normalizadas y curvas con parámetros óptimos

4.5.1. Correlación funcional

Para cada curva o dato funcional x(t), calcularemos su media, su varianza y la covarianza con otra función y(t). Es decir, obtendremos lo que se ha definido como "Estadísticos sobre una función". En el intervalo [0,T], como se ha visto en el Capítulo 1, estos estadísticos vienen dados de la siguiente manera:

`Función valor medio de` $x(t)$ es $f_{\bar{x}}(t) = \overline{x(t)}$, $\forall t \in [0,T]$, donde:

$$\overline{x(t)} = T^{-1}\langle x(t), 1(t)\rangle = T^{-1}\int_0^T x(t)1(t)dt = T^{-1}\int_0^T x(t)dt$$

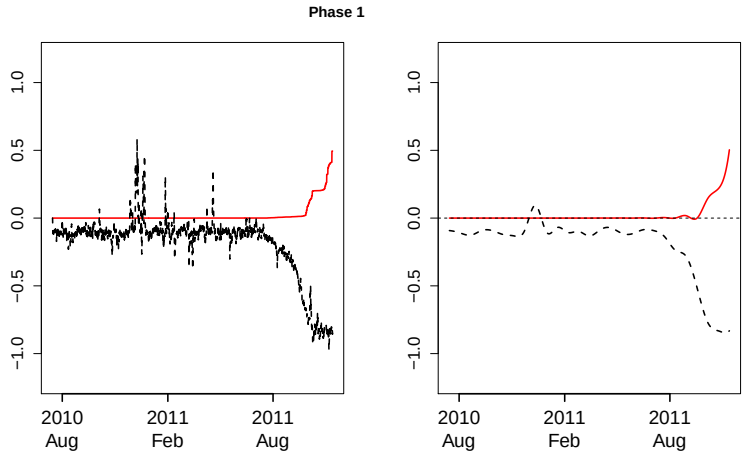


Figura 4.10: Curvas suavizadas y normalizadas sobre deformación superficial y energía sísmica acumulada con base de p-splines y parámetros óptimos para la fase 1

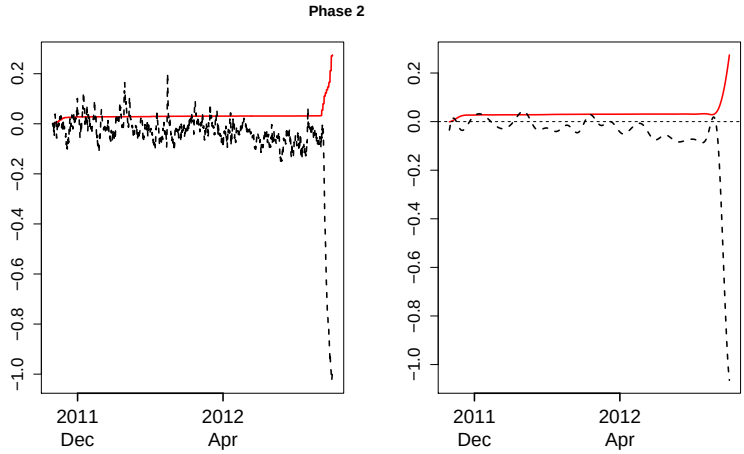


Figura 4.11: Curvas suavizadas y normalizadas sobre deformación superficial y energía sísmica acumulada con base de p-splines y parámetros óptimos para la fase 2

Función `varianza` de $x(t)$, $f_{S_X^2}(t) = S^2_{x(t)}$, $\forall t \in [0,T]$, donde:

$$S^2_{x(t)} = T^{-1}\langle x(t) - \overline{x(t)}1(t), x(t) - \overline{x(t)}1(t)\rangle = T^{-1}\int_0^T (x(t) - \overline{x(t)}1(t))^2 dt$$

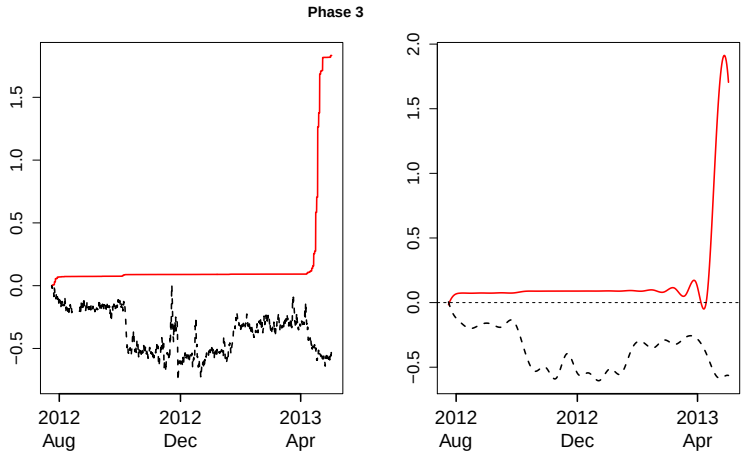


Figura 4.12: Curvas suavizadas y normalizadas sobre deformación superficial y energía sísmica acumulada con base de p-splines y parámetros óptimos para la fase 3

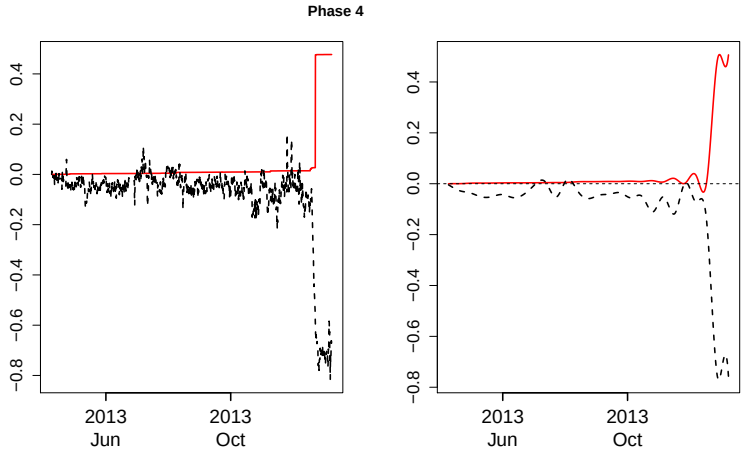


Figura 4.13: Curvas suavizadas y normalizadas sobre deformación superficial y energía sísmica acumulada con base de p-splines y parámetros óptimos para la fase 4

Función `covarianza` entre las funciones $x(t)$ e $y(t)$ es:

$f_{S_{XY}}(t) = S_{x(t)y(t)}$, donde:

$$\begin{aligned} S_{x(t)y(t)} &= T^{-1}\langle x(t) - \overline{x(t)}\mathbf{1}(t), y(t) - \overline{y(t)}\mathbf{1}(t)\rangle = \\ &= T^{-1}\int_0^T (x(t) - \overline{x(t)}\mathbf{1}(t))(y(t) - \overline{y(t)}\mathbf{1}(t))dt \end{aligned}$$

De forma que:

$$r_{x(t)y(t)} = \frac{S_{x(t)y(t)}}{S_{x(t)}S_{y(t)}}.$$

Esta medida, constante para cada par de curvas $x(t)$, $y(t)$, tomará valores en $[-1, 1]$ y nos permitirá evaluar la "relación lineal" entre las funciones.

Por otro lado, Sangalli en [68] define, a partir de dos datos funcionales x e y, $x, y : \tau = [0, T] \longrightarrow \mathbb{R}$ con $x, y \in L^2(\tau; \mathbb{R})$, una medida de similitud entre sus curvas asociadas como:

$$\rho(x, y) = \frac{\int_\tau x'(s)y'(s)ds}{\sqrt{\int_\tau (x'(s))^2 ds}\sqrt{\int_\tau (y'(s))^2 ds}}. \tag{4.5.2}$$

Esta medida puede ser interpretada como la versión continua del coeficiente de Pearson para las derivadas de las curvas x e y.

4.6. Análisis global y por fases

4.6.1. Análisis global

La medida de correlación funcional $r_{x(t)y(t)}$ planteada en la sección anterior ha sido implementada en *R* y posteriormente aplicada sobre las curvas de desplazamiento y energía sísmica mostradas en la figura 4.9. El objetivo es encontrar el periodo de tiempo, que llamaremos tiempo de desfase, en el que la medida de correlación funcional entre las curvas se hace máxima. Para ello la curva de deformación superficial se ha desplazado de 0 a 150 unidades de tiempo, y en cada desplazamiento se ha estimado la medida anterior. Los resultados obtenidos se recogen en el "correlograma funcional" de la figura 4.14.

Esta figura muestra que la máxima correlación entre las curvas ($r = 0{,}86$) aparece cuando la curva de deformación es desplazada 44 unidades, lo que se corresponde aproximadamente con una semana. Este valor muestra, como ya han estudiado otros autores ([54, 69]) la relación existente entre la curva de deformación y la energía

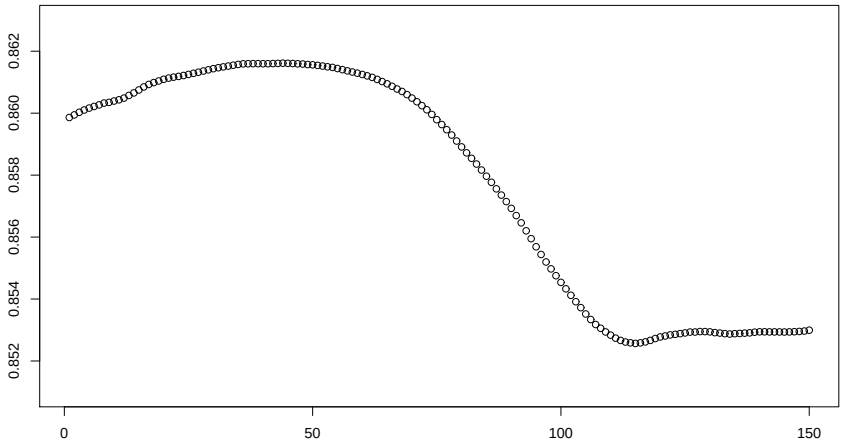


Figura 4.14:

sísmica. Sin embargo, el largo periodo de tiempo considerado en este análisis nos hace pensar en la mezcla de fenómenos sísmicos que han podido suceder en este periodo, y la influencia que este hecho podría tener sobre los resultados anteriores. Por ello, en la siguiente sección, se realizará un análisis detallado en subfases en las que los fenómenos sísmicos relevantes estarán aislados.

4.6.2. Análisis por fases

En este estudio, se ha dividido el periodo de tiempo considerado (2011-2014) en cuatro fases de forma que cada fase contenga un único periodo sísmico relevante (ver sección 4.4). A continuación, para cada fase se han estimado las curvas óptimas en la base de p-splines (Figs. 4.10, 4.11, 4.12 y 4.13).

Para cada una de las fases establecidas, se han obtenido tanto la serie de índices de correlación entre las curvas de deformación y sísmica (desplazados t unidades de tiempo, $t = 0, \ldots, n$), como los días de retardo. Los coeficientes máximos de correlación en las distintas fases se recogen en la tabla 4.1. Se puede ver cómo en las fases 1, 2 y 4 se han obtenido correlaciones por encima de 0.936, llegando en la fase 4 a 0.986; mientras que en la fase 3 se ha situado en 0.346, correspondiéndose con un fenómeno de inflación-deflación. En cuanto al retardo, va desde 31.7 días en la fase 1 a casi 3 días en la fase 4.

Cuadro 4.1: Máxima correlación funcional y tiempo de retardo por fases

FASE	CORRELACIÓN	TIEMPO DE DESFASE
Fase 1	0.936	31.7
Fase 2	0.971	0
Fase 3	0.346	1.5
Fase 4	0.986	2.67

Cuadro 4.2: Máxima correlación funcional y tiempo de retardo en las subfases de la fase 3

FASE	CORRELACIÓN	TIEMPO DE DESFASE
3 (a)	0.84	0
3 (b)	0.32	5.67
3 (c)	0.84	0

El fenómeno de inflación-deflación existente en la tercera fase nos ha llevado a considerar las subfases a, b y c, obtenidas mediante "Análisis del punto de cambio". De nuevo en estos tres subperiodos ha sido aplicado el mismo procedimiento, obteniéndose los datos mostrados en la tabla 4.2.

Al aislar el fenómeno de inflación-deflación en la subfase b, vemos que las subfases a y c muestran correlaciones de $r = 0{,}84$, mientras la fase 3 sólo se ha alcanzado una correlación de $r = 0{,}346$. Con respecto a los tiempos de desfase entre la defomación y la sísmica, es relevante el hecho de que en la primera fase, tras un largo periodo de inactividad sísmica, existe un periodo de desfase de aproximadamente un mes, mientras en las fases posteriores este tiempo es casi inexistente.

4.7. Predicción

Se ha establecido, en las secciones anteriores, una alta correlación entre la curva de sismicidad y la curva de deformación. Además, después de un largo periodo sin actividad sísmica y volcánica, el lapso de tiempo entre el proceso de deformación y el comienzo de la actividad sísmica toma aproximadamente un mes. En esta sección se propondrá un método para predecir en tiempo real el comienzo de la actividad

sísmica.

A la serie de desplazamiento se agregarán secuencias de datos correspondientes a dos días desde aproximadamente el comienzo de la deformación el 28 de junio de 2011 ($t = 2100$). Se consideran las curvas splines con los parámetros óptimos y se fija un nivel de derivada mínima. Este nivel se calcula como el mínimo de la pendiente de los dos días anteriores. Al calcular el nivel, los primeros 30 datos no se consideran debido a las oscilaciones que suceden al comienzo de las curvas p−splines.

Las figuras 4.15, 4.16 y 4.17 muestran las curvas p−splines en las diferentes secuencias de datos.

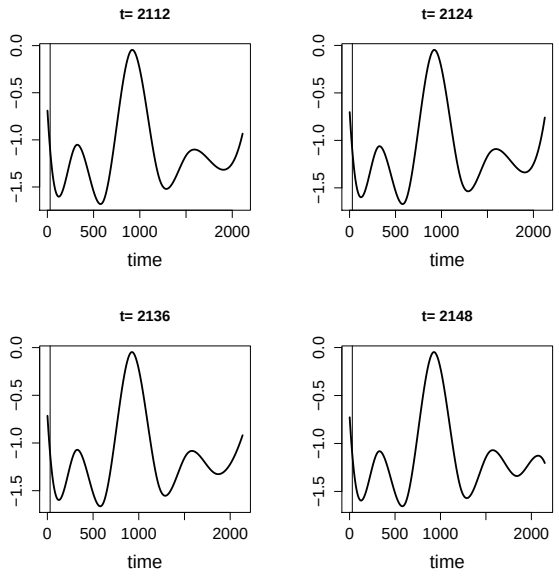


Figura 4.15: Serie de desplazamiento normalizada en la base de p-splines, considerando las secuencia de 2112, 2124, 2136 y 2148 datos

Las figuras 4.18, 4.19 y 4.20 muestran las curvas de derivadas en la base de splines y la línea que marca el nivel de pendiente mínima. Estas cifras muestran que después de 108 datos, que corresponden a un periodo de aproximadamente 18 días, se excede la pendiente mínima y, en ese caso, se activa un sistema de alerta. Esto ocurre aproximadamente 13 días antes del comienzo de la sismicidad.

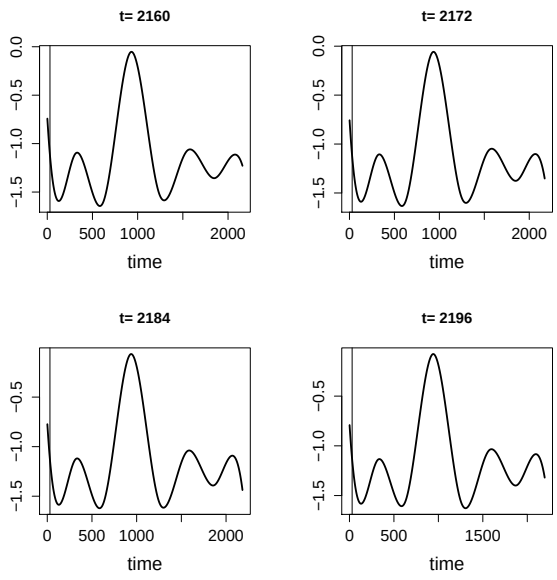


Figura 4.16: Serie de desplazamiento normalizada en la base de p-splines, considerando las secuencia de 2160, 2172, 2184 y 2196 datos

4.8. Resultados

Mediante el uso del índice de similitud propuesto por Sangalli ([68]), se ha obtenido una medida de correlación funcional. Esta medida ha sido empleada para establecer la relación entre la curva de deformación (medida por el desplazamiento entre la estación FRON y la estación de referencia LPAL) y la curva de energía sísmica acumulada.

Se ha realizado un análisis global de estos datos para validar la relación anterior y, posteriormente, mediante un análisis en varias fases y el uso de la medida de correlación definida, se ha establecido una alta relación en todas las fases ($r = 0{,}936$), excepto en la que contiene un fenómeno de inflación−deflación. Las diferentes fases han sido definidas teniendo en cuenta que un sólo evento sismo-volcánico relevante tenga lugar en cada fase. Además, cada fase finaliza el día del terremoto de mayor magnitud en el período.

Un análisis detallado de la tercera fase, que utiliza el análisis del punto de cambio, muestra que solo la subfase que contiene el fenómeno anómalo de inflación-deflación produce un valor de correlación funcional bajo. En las otras subfases de esta fase

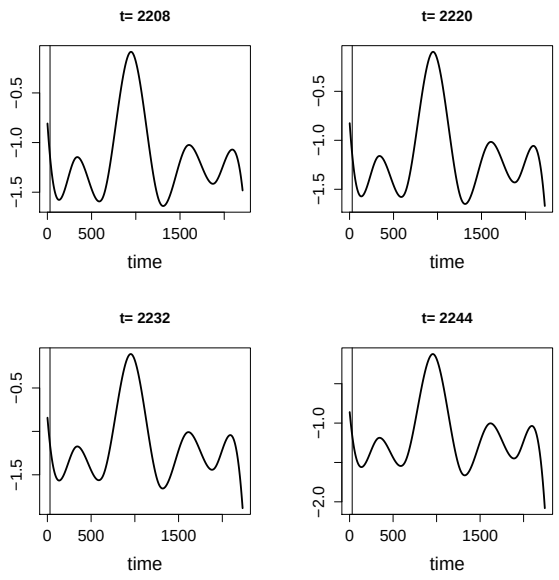


Figura 4.17: Serie de desplazamiento normalizada en la base de p-splines, considerando las secuencia de 2208, 2220, 2232 y 2244 datos

los valores de correlación son similares a los de los otros períodos. Por lo tanto, este tipo de fenómeno podría detectarse utilizando la medida de correlación propuesta.

El tiempo de sincronización entre la curva de deformación y la curva de energía sísmica acumulada se ha analizado en las diferentes fases. Se ha observado que el largo período de inactividad volcánica antes de la primera fase produce un tiempo de reacción de un mes entre el proceso de deformación y los datos sísmicos. Sin embargo, en las fases subsiguientes, los cambios producidos en la reología y dinámica del magma en las fases iniciales, produjeron un efecto en la curva de deformación de la superficie. Posiblemente debido a este efecto, la curva de deformación y la curva de energía sísmica acumulada están sincronizadas en sus movimientos inversos y, por lo tanto, el tiempo de reacción es casi inexistente.

Después de un largo período sin actividad sísmica y volcánica en la isla de El Hierro, el lapso de tiempo entre el proceso de deformación y el comienzo de la actividad sísmica tiene una duración de aproximadamente un mes.

Se ha propuesto un método para predecir en una situación similar el comienzo de la actividad sísmica en tiempo real. Este sistema de alerta, basado en los cambios producidos en las curvas derivadas cuando hay un descenso rápido en la curva de

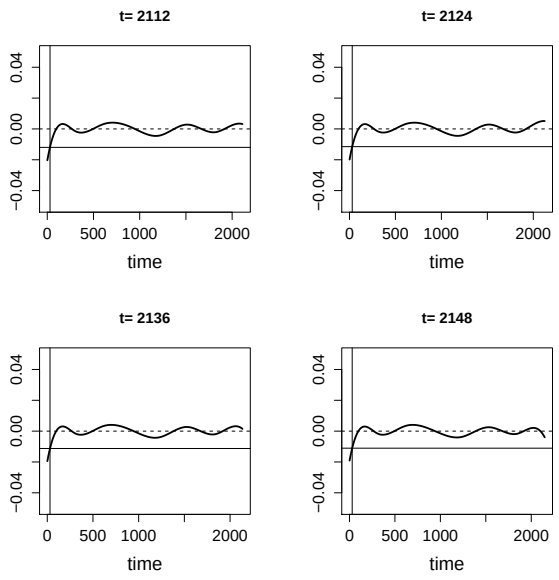


Figura 4.18: Curvas derivadas de la serie de desplazamiento normalizada en la base de p-splines, considerando las secuencia de 2112, 2124, 2136 y 2148 datos, donde la línea horizontal marca la pendiente mínima

deformación, podría activarse según este sistema aproximadamente 13 días antes del comienzo de la sismicidad.

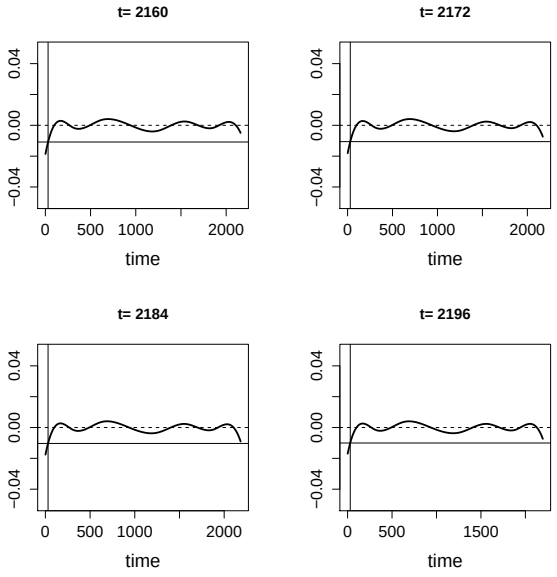


Figura 4.19: Curvas derivadas de la serie de desplazamiento normalizada en la base de p-splines, considerando las secuencia de 2160, 2172, 2184 y 2196 datos, donde la línea horizontal marca la pendiente mínima

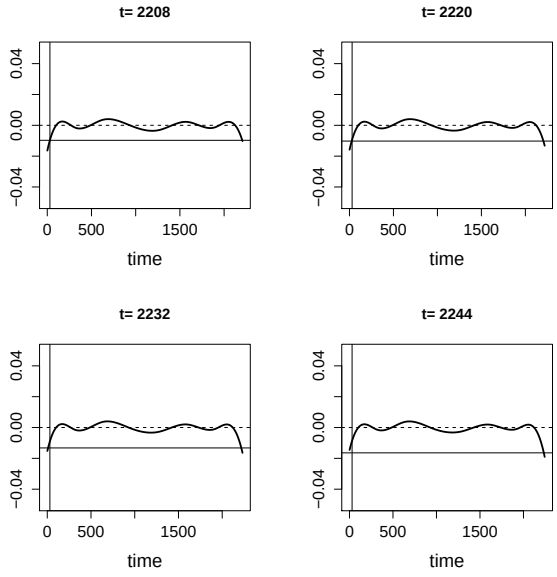


Figura 4.20: Curvas derivadas de la serie de desplazamiento normalizada en la base de p-splines, considerando las secuencia de 2208, 2220, 2232 y 2244 datos, donde la línea horizontal marca la pendiente mínima

Capítulo 5

Conclusiones, Limitaciones y Futuras líneas de trabajo

La principales conclusiones que extraemos de la Memoria presentada son:

- Dada la naturaleza de los datos procedentes de observaciones GPS, el marco de análisis que proporciona el FDA es muy útil para el tratamiento de problemas del ámbito de la Geodesia.

- El estudio de la variabilidad de las mediciones obtenidas en las estaciones de la Red Spina y la clasificación de las mismas en grupos homogéneos, ayudan a comprender el comportamiento de las placas tectónicas de la región Iberia-África.

- Mediante el análisis de correlaciones entre las series desplazadas de Deformación y Energía acumulada, para el total del periodo considerado, hemos comprobado que transcurre aproximadamente una semana entre el proceso de deformación y el comienzo de la actividad sísmica en la Isla de El Hierro. Este periodo coincide con el que han obtenido otros investigadores.

- El análisis anterior se ha reproducido descomponiendo el periodo total en cuatro fases, caracterizadas porque cada una de ellas contiene un único evento sísmico relevante. Para cada fase se ha obtenido su tiempo de retardo.

Las limitaciones de nuestro trabajo fueron:

- La falta de información sobre el borde de la zona de subducción de las placas ibérica y africana, nos ha condicionado a aplicar procedimientos de carácter exploratorio. Es razonable pensar que con una mayor información podríamos aplicar en el futuro modelos confirmatorios. A ello contribuirá, sin duda, el aumento de la densidad de la red de posicionamiento y la mejora del criterio de elección de la localización de las nuevas estaciones.

En relación a las futuras líneas de investigación, destacar las siguientes:

- Crear un protocolo de análisis de los datos procedentes de la red de estaciones, de forma que nos permita analizar la información registrada y procesada en un tiempo lo más reducido posible. La consecución de este objetivo se verá favorecido, con la ampliación que se está haciendo del número de estaciones GPS, que dotará al territorio de una red lo suficientemente densa como para determinar con buena precisión las fronteras de las placas tectónicas.

- Relacionado con el punto anterior, pretendemos proponer y validar modelos funcionales que nos ayuden a entender el comportamiento dinámico de las placas tectónicas en la región Iberia-África.

- Aplicar las técnicas FDA en otras áreas en las que, dada la naturaleza y disponibilidad de los datos, dichas técnicas sean eficientes. En particular, hemos empezado a trabajar en el estudio de la volatilidad en los mercados de valores, a través de una medida basada en las curvas de velocidad.

- Una de las cuestiones más interesantes del análisis funcional, que es determinante en los resultados que se obtendrán, es la decisión sobre la medida a usar para determinar la disimilaridad entre las curvas. Estamos trabajando en la definición de una nueva semimétrica para datos funcionales basada en la distancia de Lipschitz. Esta semimétrica se adaptaría bien a situaciones en las que, aunque las variables se midan en un continuo, es decir tendría formato funcional, sólo tomaría valores significativamente distintos de cero en instantes puntuales.

Apéndice A

CÓDIGO R

```
#FUNCIóN ELIMINACION DE ATIPICOS


#El conjunto de Datos (Norte, Este o Altitud) contiene 1095 filas
#y 109 columnas (La primera es la Fecha y a continuación están los datos
de desplazamiento y el error por estación)


funcelimatip<-function(Datos,ini,fin,cotasigma){
Datos$Fecha<-ini:fin
numestaciones<-(ncol(Datos)-1)/2
as.matrix(Datos[ ,seq(2,ncol(Datos),2)[,nrow=nrow(Datos),ncol=numestaciones)->D0
t(D0)->Akk
data.frame(t(t(D0)-t(D0)[,1]))->Datos0
#Datos0 contiene sólo los datos de las estaciones inicializados
atipicos<-0
DatossinAtipEstacion<-matrix(0,nrow=length(Datos$Fecha),ncol=numestaciones)
for (j in seq(1:numestaciones)){
auxs<-c(Datos[, 2*j+1 ])
auxd<-c(Datos0[,j])
mean(na.omit(auxs))->sigmaest
which(auxs>cotasigma*sigmaest)->atipest
length(atipest)+atipicos->atipicos
auxd[atipest]<-NA
auxd->DatossinAtipEstacion[,j]
```

```
rm(auxd,auxs,sigmaest,atipest)}
return(list(A=atipicos,D=DatossinAtipEstacion))
}


#FILTRO KALMAN


buildFun2 <- function(theta) { dlmModPoly(order=2, dV = exp(theta[1]),
dW = c(exp(theta[2]), exp(theta[3]) )) }

#Función para aplicar el filtro Kalman de orden 2 a cada variable
#del conjunto Datos
library(“dlm”, lib.loc=“∼/R/win-library/3.2”)
filtroKalman2Datos<-function(Datos){
z<-names(Datos)[-1]
Datosfilt<-Datos[,-1]
Datosnuevos<-Datosfilt[-dim(Datosfilt)[1],]
dim(Datosfilt)[2]->variables
for (i in 1:(variables)){
aux<-Datosfilt[,i]
EstimaMLE2 <- dlmMLE(aux, parm = c(0, 0, 0), build = buildFun2)


EstimaPAR2 <- buildFun2(EstimaMLE2$par)
Filtro2 <- dlmFilter(aux, EstimaPAR2)
Filtro2$f[-1]->pro2
plot(pro2, type = ’p’, col = “red”, pch=20 ,main=paste(z[i], “con filtro
Kalman de orden 2”))
pro2->Datosnuevos[,i]
points(aux,add=TRUE,col=“blue”)
rm(aux,EstimaMLE2,EstimaPAR2,Filtro2)
}
Datos[-dim(Datos)[1],1]->tiempos
data.frame(cbind(tiempos,Datosnuevos))->filtroKalman2Datos
}


#CORRELACIÓN FUNCIONAL
```

```
corfd<-function(M,ini,fin,col1,col2,n,la){
if(col1!=col2){
T<-fin-ini
i1<-t(M[,col1])
i2<-t(M[,col2])
fdata(i1,argvals=ini:fin)->curva1
fdata(i2,argvals=ini:fin)->curva2
l<-length(seq(ini,fin,1))
fdata(rep(1,l),argvals=ini:fin)->curvaid1
#Ajustamos los datos a las curvas b-splines
bsplinesx1<-fdata2fd(curva1, nbasis=n,lambda=la)
bsplinesx2<-fdata2fd(curva2, nbasis=n,lambda=la)
bsplinesid1<-fdata2fd(curvaid1,nbasis=n,lambda=la)


# Convertimos de nuevo en fdata:
fdatax1=fdata(bsplinesx1)
fdatax2=fdata(bsplinesx2)
fdataid1=fdata(bsplinesid1)


#Calculamos las medias:
mx1<-(1/T)*inprod.fdata(fdatax1,fdataid1)
mx2<-(1/T)*inprod.fdata(fdatax2,fdataid1)


#Calculamos las varianzas:
difsx1<-M[,1]-rep(mx1,l)
fdata(difsx1,argvals=ini:fin)->curvadifsx1
bsplinesdifsx1<-fdata2fd(curvadifsx1,nbasis=n,lambda=la)
fdatadifsx1=fdata(bsplinesdifsx1)
sx1<-sqrt((1/T)*inprod.fdata(fdatadifsx1,fdatadifsx1))
difsx2<-M[,2]-rep(mx2,l)
fdata(difsx2,argvals=ini:fin)->curvadifsx2
bsplinesdifsx2<-fdata2fd(curvadifsx2,nbasis=n,lambda=la)
fdatadifsx2=fdata(bsplinesdifsx2)
sx2<-sqrt((1/T)*inprod.fdata(fdatadifsx2,fdatadifsx2))


#calculamos la covarianza:
sxy<-(1/T)*inprod.fdata(fdatadifsx1,fdatadifsx2)
```

```
r<-sxy/(sx1*sx2)else{
r<-1}
}
```

Apéndice B

ACRÓNIMOS

Capítulo 1. Análisis de Datos Funcionales

FDA	—	Functional Data Analysis.
FPC	—	Functional Principal Components.
FPCA	—	Functional Principal Components Analysis.
PLS	—	Partial Least Squares.
MPLSR	—	Multivariate Partial Least Squares Regression.
GCVC	—	Generalized Cross Validation Criterion

Capítulo 2. Datos geodésicos

NAVSTAR	—	NAVigation System with Timing and Ranging.
GPS	—	Global Positioning System.
GNNS	—	Global Navigation Satellite System.
*ITRF*2008	—	International Terrestrial Reference System 2008.
IGS	—	International GNNS Service.
MLD	—	Modelo Lineal Dinámico.

Capítulo 3. Análisis de la red SPINA desde un enfoque funcional

SPINA	—	Sur de la Península Ibérica y Norte de África.
RENEP	—	Red Nacional de Estaciones Permanentes de Portugal.
RAP	—	Red Andaluza de Posicionamiento.

IGN	—	Instituto Geográfico Nacional Español.
REGAM	—	Red Geodésica Activa de Murcia.
ERVA	—	Red de Estaciones de Referencia de Valencia.
NA	—	Not Available.
ENU	—	East, North and Up coordinates
MSE	—	Mean square error.
ENU	—	East, North and Up coordinates

Bibliografía

[1] JO Ramsay and BW Silverman. *Functional Data Analysis*. Springer, New York, NY, 2005.

[2] F Ferraty and P Vieu. *Nonparametric Functional Data analysis*. Springer-Verlag: Berlin, 2006.

[3] JP Aubin. *Applied Functional Analysis*. Willey Interscience Publication. 2 Ed., 2000.

[4] JL Torrecillas Noguerales. Análisis de datos funcionales, clasificación y selección de variables. trabajo fin de master. Master's thesis, Universidad Autónoma de Madrid, Escuela Politécnica Superior, 2010.

[5] J Dauxois and Y Romain. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12:54—-61, 1982.

[6] V Navarro Pérez. Análisis de datos funcionales, implementación y aplicaciones. (proyecto fin de carrera). Master's thesis, Universitat Politécnica de Catalunya, Facultat de Matematiques i Estadística, 2004.

[7] C De Boor. Package for calculating with b-splines. *Journal of Numerical Analysis*, 14:441–472, 1977.

[8] F O′Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Sciences*, 1:505–527, 1986.

[9] PHC Eilers and BD Marx. Flexible smoothing with b-splines and penalties (with discussion). *Statistical Sciences*, 11:89–121, 1996.

[10] MC Aguilera Morillo. Estimación penalizada con datos funcionales. trabajo investigación master estadística aplicada. Master's thesis, Universidad de Granada, 2009.

[11] J.A Rice and CO Wu. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57:253–259, 2001.

[12] P Craven and G Wahba. Smoothing noisy data with splines functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979.

[13] BB Hubbert. *The world according to wavelets: The story of a mathematical technique in the making.* AK Peters/CRC Press, 1998.

[14] P Bermolen and F Larroca. Cálculo de la ventana óptima para el estimador de nadaraya-watson. monografia presentada para la aprobación del curso automatic learning machines. dictado por el prof. gonzalo perera. Technical report, IMERL, 2005.

[15] EA Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 10:186–190, 1964.

[16] GS Watson. Smooth regression analysis. *Sankhya. Series A*, 26:101–116, 1964.

[17] D Ruppert, SJ Sheather, and MP Wand. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(462):1257–1270, 1995.

[18] JO Ramsay. When the data are functions. *Psychometrika*, 47(4):379–396, 1982.

[19] E Piza. Estudio de las métricas inducidas por un analisis en componentes principales. *Revista de Matemática: Teoría y Aplicaciones*, 3:27–36, 1996.

[20] J Peng and M˙ Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *The Annals of Applied Statistics*, 2:1056–1077, 2008.

[21] C Abraham, PA Cornillon, E Matzner-Lober, and N Molinari. Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics*, 30:581–595, 2003.

[22] M˙Functional variance processes.

[23] M˙Functional data analysis for volatility. *Journal of Econometrics*, 165(2):233–245, 2011.

[24] L Ferreira and DB Hitchcock. A comparison of hierarchical methods for clustering functional data. *Communications in Statistics-Simulation and Computation*, 38:1925–1949, 2009.

[25] Gerald Beer and Michael J Hoffman. The lipschitz metric for real-valued continuous functions. *Journal of Mathematical Analysis and Applications*, 406(1):229–236, 2013.

[26] Nik Weaver. *Lipschitz algebras*. World Scientific, 1999.

[27] J Jacques and C Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8:24, 2014.

[28] C Bouveyron and C Brunet. Model-based clustering of high-dimensional data: A review. Technical report, University of Paris, Panthéon-Sorbonne, 2012.

[29] JD Banfield and AE Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1993.

[30] Thaddeus Tarpey. Linear transformations and the k-means clustering algorithm: applications to clustering curves. *the american statistician*, 61(1):34–40, 2007.

[31] M Yamamoto. Clustering of functional data in a low-dimensional subspace. *Advances in Data Analysis and Classification*, 6(3):219–247, 2012.

[32] G James and CA Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98:397–408, 2003.

[33] PJ Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[34] M Berrocoso, ME Ramírez, A Pérez Peña, JM Enríquez de Salamanca García, A Fernández, and C Torrecillas. *Sistema de Posicionamiento Global*. Servicio de publicaciones, Universidad de Cádiz, 2003.

[35] B Rosado. *Modelización matemática de la actividad volcánica: Análisis de series temporales GNSS, algoritmos de inversión y pronóstico espacio-temporal*. PhD thesis, Universidad de Cádiz, 2019.

[36] A Leick. *Functional Data Analysis*. Ed. Wiley Intersciencies, New York, 1995.

[37] P Teunisen and A Kleusberg. *GPS for Geodesy*. Springer-Verlag, New York, 1998.

[38] R Dach, U Hugentobler, P Fridez, and M Meindl. Bernese gps software ver. 5.0. 2011.

[39] L Ostini. *Analysis and Quality Assessment of GNSS Derived Parameter Time Series. PhD thesis*. PhD thesis, University of Bern, 2012.

[40] RE Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82 (Series D):35–45, 1960.

[41] P Harris, C Brunsdon, M Charlton, S Juggins, and A Clarke. Multivariate spatial outlier detection using robust geographically weighted methods. *Math Geosci*, 46:1–31, 2014.

[42] J Sancho, C Iglesias, and J Piñeiro. Study of water quality in a spanish river based on statistical process control and functional data analysis. *Math Geosci*, 48:163–186, 2016.

[43] B Rosado, I Barbero, A Jiménez, R Páez, G Prates, A Fernández-Ros, J Gárate, and M Berrocoso. *SPINA Region (South of Iberian Peninsula, North of Africa) GNSS Geodynamic Model*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017.

[44] E Buforn, M Bezzeghoud, A Udias, and C Pro. Seismic sources on the iberia-african plate boundary and their tectonic implications. *Pure and Applied Geophysics*, 161(3):623–646, 2004.

[45] V Araña and R Ortiz. *The Canary Islands: Tectonics, magmatism and geodynamic framework. In Magmatismin extensional structural settings (The Phanerozoic African plate)*. eds A. B. Kampunzu and R. T. Lubala. Springer-Verlag, Barcelona, Spain, 1991.

[46] JC Carracedo, SJ Day, H Guillou, and FJ Pérez Torrado. Giant quaternary landslides in the evolution of la palma and el hierro, canary islands. *Journal of Volcanology and Geothermal Research*, 94:169–190, 1999.

[47] F Anguita and F Hernán. The canary islands origin: a unifying model. *Journal of Volcanology and Geothermal Research*, 103:1—-26, 2000.

[48] MJR Gee, DG Masson, AB Watts, and NC Mitchell. Offshore continuation of volcanic rift zones. el hierro. canary islands. *Journal of Volcanology and Geothermal Research*, 105:107–119, 2001.

[49] NC Mitchell, DG Masson, AB Watts, MLR Gee, and R Urgeles. The morphology of the submarine flanks of volcanic ocean islands. a comparative study of the canary and hawaiian hotspot islands. *Journal of Volcanology and Geothermal Research*, 115:83—-107, 2002.

[50] NA Stroncik, A Klügel, and TH Hansteen. The morphology of the submarine flanks of volcanic ocean islands. a comparative study of the canary and hawaiian hotspot islands: Constraints from phenocrysts and naturally quenched basaltic glasses in submarine rocks. *Contrib. Mineral. Petrol.*, 157:593–607, 2009.

[51] A Hernández-Pacheco. Sobre una posible erupción en 1793 en la isla de el hierro (canarias). *Estudios geológicos*, 38:15–26, 1982.

[52] JC Carracedo, E Rodríguez Badiola, H Guillou, J De la Nuez, and FJ Pérez-Torrado. Geology and volcanology of la palma and el hierro (canary islands). *Estudios geológicos*, 57:175—-273, 2001.

[53] G Prates, A García, A Fernández-Ros, JM Marrero, R Ortiz, and M Berrocoso. Enhancement of sub-daily positioning solutions for surface deformation surveillance at el hierro volcano (canary islands, spain). *Bull. Volcanol.*, 75(6):1–9, 2013.

[54] H Lamolda, A Felpeto, and A Bethencourt. Time lag between deformation and seismicity along monogenetic volcanic unrest periods: The case of el hierro island (canary islands). *Geophys. Res. Lett.*, 44:6771–6777, 2017.

[55] Joan Martí, Antonio Castro, Carmen Rodríguez, Fidel Costa, Sandra Carrasquilla, Rocío Pedreira, and Xavier Bolos. Correlation of magma evolution and geophysical monitoring during the 2011–2012 el hierro (canary islands) submarine eruption. *Journal of Petrology*, 54(7):1349–1373, 2013.

[56] Daniel Dzurisin. *Volcano deformation: new geodetic monitoring techniques*. Springer Science & Business Media, 2006.

[57] C López, MJ Blanco, R Abella, B Brenes, VM Cabrera Rodríguez, B Casas, I Domínguez Cerdeña, A Felpeto, M Fernández de Villalta, C Del Fresno, et al. Monitoring the volcanic unrest of el hierro (canary islands) before the onset of the 2011–2012 submarine eruption. *Geophysical Research Letters*, 39(13), 2012.

[58] M Berrocoso, A Fernández-Ros, G Prates, M Martín, R Hurtado, J Pereda, MJ García, L García Cañada, R Ortiz, and A García. Analysis of surface deformation during the eruptive process of el hierro island (canary island, spain): Detection, evolution and forecasting. In *EGU General Assembly Conference Abstracts*, volume 14, pages 43—-51, 2012.

[59] J Saastamoinen. Contribution of the theory of atmospheric refraction. *Géodésique*, 107:13–34, 1973.

[60] AE Niell. Global mapping functions for the atmosphere delay at radio wavelengths. *J. geophys. Res.*, 101(B2):3227–3246, 1996.

[61] L Mervart. *Ambiguity Resolution Techniques in Geodetic and Geodynamic Applica-tions of the Global Positioning System. (PhD thesis)*. PhD thesis, University of Bern, 1995.

[62] P Rebischung, J Ray, C Benoist, L Metivier, and Z Altamimi. Error Analysis of the IGS repro2 Station Position Time Series. In *AGU Fall Meeting Abstracts*, volume 2015, pages G23B–1065, Dec 2015.

[63] Z Altamimi, X Collilieux, and L Métivier. Itrf2008: an improved solution of the international terrestrial reference frame. *J. Geod.*, 85(8):457–473, 2011.

[64] GL Choy and JL Boatwright. Global patterns of radiated energy and apparent stress. *Journal of Geophysical Research*, 100:18205–18228, 1995.

[65] Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. Segmenting time series: A survey and novel approach. In *Data mining in Time Series Databases*, pages 1–21. World Scientific, Singapore, 2004.

[66] S Aminikhanghahi and DJ Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51:339–367, 2017.

[67] M Febrero-Bande and M Oviedo de la Fuente. Statistical computing in functional data analysis: The r package fda.usc. *Journal of Statistical Software*, 51:1–28, 2012.

[68] LM Sangalli, P Secchi, S Vantini, and A Veneziani. A case study in exploratory functional data analysis: geometrical features of the internal carotid artery. *Journal of the American Statistical Association*, 104:37—-48, 2009.

[69] I Domínguez Cerdeña, L García Cañada, MA Benito Sanz, C Del Fresno, H Lamolda, J Pereda de Pablo, and C Sánchez Sanz. On the relation between ground surface deformation and seismicity during the 2012-2014 succesive magmatic intrusions at el hierro island. *Technophysics*, 744:422–437, 2018.