Centro de Investigación en Matemáticas

Maestría en Probabilidad y Estadística

ANÁLISIS FUNCIONAL DEL PERFIL DE LAS OLAS

Tesis que presenta

Cristina Gorrostieta Hurtado

Para obtener el grado de

Maestría en Ciencias con especialidad en

Probabilidad y Estadística

Director de tesis:

_____

Dr. Joaquín Ortega Sánchez

julio 2007, Guanajuato, Gto

AGRADECIMIENTOS

- *A CONACYT por su colaboración mediante el proyecto "Análisis Estadístico de Olas Marinas", con registro 67796.*

- *George H. Smith de la Universidad de Exeter Inglaterra por los datos proporcionados.*

- *A Total E&P UK por los datos de olas de la plataforma Alwyn Norte.*

Indice

Lista de figuras

Capítulo 1

Introducción

Las olas oceánicas han sido tema de estudio de matemáticos, físicos e ingenieros durante mucho tiempo, por lo que cada vez se tiene más información acerca de su comportamiento. Sin embargo debido a su complejidad quedan aún muchos aspectos por investigar sobre ellas. Uno de los mayores intereses se presenta en las olas de tormenta en mar abierto, ya que son la causa de graves accidentes en el océano. En este contexto surge el planteamiento de la siguiente pregunta: ¿Cómo es la forma de este tipo de olas y de qué manera varían en un periodo de tiempo?. Es alrededor de esta pregunta sobre la que se desarrolla el presente trabajo de tesis.

Una manera de cuantificar el grado de alteración de las olas en un periodo de tiempo es a través de la altura significativa, así, al dividir la duración de una tormenta en periodos de tiempo, la altura significativa de los periodos proporciona información sobre la evolución y la variación en el grado de intensidad de las olas durante la tormenta. En consecuencia, la altura significativa juega un papel muy importante en la forma de las olas, por lo que uno de los intereses principales en esta tesis es explorar de qué manera está relacionada la forma de las olas en un periodo de tiempo con la altura significativa de ese periodo.

El enfoque tradicional del estudio de olas es mediante el análisis de la serie que representa la altura de la ola al tiempo t en un punto fijo, $\eta(t)$, vista como un proceso aleatorio. Algunas de las propiedades comúnmente estudiadas de este proceso, se refieren a las frecuencias de la serie, su densidad espectral y estimación de distribuciones asociadas a características de las olas que la integran como: amplitud,

periodo, valle, cresta, intensidad de cruzado entre otras. Sin embargo no se tiene información de estudios realizados sobre tipos de variación en la forma de las olas y su comparación con altura significativa.

Aunque usualmente se analiza $\eta(t)$ como un proceso aleatorio, en este caso, debido a que se tiene interés en la forma de las olas, se requiere otro enfoque que proporcione ventajas al estudio sobre este aspecto. La teoría de datos funcionales facilita herramientas que son de utilidad en el estudio de forma y variabilidad de funciones, de manera que al representar el perfil de las olas como funciones, esta teoría resulta de gran utilidad en el análisis de los datos para el objetivo propuesto.

Concretamente, dada la serie $\eta(t)$ correspondiente a alturas de olas de tormenta, la tesis presenta un análisis estadístico de la forma de las olas que la definen tratadas como funciones. Durante el desarrollo de este análisis se plantean relaciones de interés entre la forma de las olas y características típicas asociadas a $\eta(t)$, como son altura significativa, el comportamiento sinusoidal y supuestos de normalidad del proceso $\eta(t)$.

Los datos utilizados en el análisis, fueron proporcionados por George H. Smith de la Universidad de Exeter Inglaterra y las mediciones se realizaron durante una tormenta en la plataforma Alwyn en el Mar del Norte.

Inicialmente, en los capítulos dos y tres, se proporcionan los preliminares que fueron de utilidad en la realización del estudio. El capítulo dos, presenta una breve introducción a la teoría de análisis funcional, que es la metodología con la que se realizó el análisis estadístico. Esencialmente esta teoría fue propuesta por Ramsay y Dalzell (1991) en [1]; y [2], [3] son los textos fundamentales sobre este tema. Por esta razón los ejemplos y temas expuestos sobre análisis funcional en este capítulo, fueron tomados de esta bibliografía.

En el capítulo tres, se da una descripción general de algunos aspectos asociados al análisis de $\eta(t)$ visto como un proceso aleatorio, específicamente, se explican los conceptos de altura significativa, densidad espectral, proceso gaussiano y olas extremas. Las ideas básicas del capítulo fueron tomadas principalmente de [4] y [5]. Finalmente, en este capítulo se muestra como se trató la serie $\eta(t)$ para llevar a cabo el análisis funcional.

Una vez que se cuenta con los datos de forma funcional, los siguientes capítulos

presentan los análisis realizados. El capítulo 4 consiste de un análisis exploratorio de la forma de las olas promedio por periodo. El siguiente capitulo, muestra las relaciones encontradas entre altura significativa y la forma de la ola promedio en un determinado periodo de tiempo. Enseguida se realiza un análisis de la forma de las olas de un periodo de tiempo con gran altura significativa. Finalmente se presentan los resultados que comparan las olas provenientes de un proceso gaussiano con el conjunto de datos observados.

En los análisis realizados se utilizó el siguiente software: en el caso de análisis funcional de los datos se implementó a través de las funciones de MATLAB proporcionadas en la página [7]; y para llevar a cabo la exploración de los datos tratados como una serie de nivel de la superficie del mar, se empleó el *toolbox* de MATLAB para olas aleatorias (WAFO), tomado de la página [8].

Capítulo 2

Preliminares de análisis funcional de datos

En muchos campos se presentan datos que son de naturaleza funcional, es decir se tiene conjuntos de poblaciones de curvas o registros que aunque se obtengan de manera discreta es más natural considerarlos por su contexto en términos funcionales. Una de las causas por las que actualmente estos datos son tan comunes son las herramientas tecnológicas disponibles, ya que existen cada vez más equipos que permiten el registro de datos de forma rápida y efectiva, por ejemplo el monitoreo mediante sensores o registros en linea en áreas como medicina, sismología, finanzas y meteorología entre otras.

El uso de una técnica que incorpore la naturaleza funcional de los datos puede tener grandes ventajas, como son el análisis de las derivadas de las funciones y la facilidad para tomar en cuenta características de las observaciones funcionales como suavidad. Sin embargo, debido a la diversidad de herramientas disponibles para el análisis de datos discretos, es común que aunque sea conveniente pensar que los datos son expresiones de una función, se les análice bajo un enfoque discreto.

La disponibilidad de la derivada es relevante en varios casos, ya que proporciona información sobre la tasa de cambio de algún proceso, por ejemplo en mediciones de crecimiento, ajustando una curva de crecimiento suave a los datos obtenemos información sobre la velocidad del crecimiento y en qué periodos este fenómeno presenta una aceleración notable. Por otro lado, una exploración en la dinámica

de un proceso puede permitir plantear ecuaciones diferenciales que lo describan, lo cual también es una ventaja del empleo de derivadas en el análisis de los datos.

El análisis funcional de datos provee un conjunto de técnicas estadísticas para el análisis de la información de curvas o funciones. Dichas técnicas pueden considerarse como la generalización de una técnica equivalente en análisis multivariado, de manera que es posible considerar métodos como componentes principales y modelos lineales en un contexto funcional, pero aprovechando la información obtenida al tomar en cuenta toda la función.

Una característica adicional del enfoque funcional es el uso de herramientas gráficas para presentar la información, que ayudan a visualizar mejor características que desde un punto de vista multivariado se pasarían por alto.

2.1 Ejemplos de datos funcionales

Para ejemplificar algunas formas en las que pueden surgir datos funcionales, se presentan algunos ejemplos, tomados de [3] y de [2].

Ejemplo 1:

El estudio del crecimiento humano es de utilidad para ayudar a establecer un patrón de crecimiento adecuado y poder detectar posibles problemas. La recopilación de datos para un análisis de crecimiento no es una tarea fácil ya que las mediciones en las personas requieren un seguimiento por varios años. La figura 2.1, muestra mediciones de altura de 10 niñas que participaron en un estudio de crecimiento llevado a cabo en 31 etapas, durante 18 años. En la figura se observa que los tiempos de medición no son igualmente espaciados y que el crecimiento se da más rápidamente en edades tempranas, con un incremento notable en la pendiente durante la pubertad.

Cada registro de 31 mediciones puede pensarse como un vector resultado de evaluar una función de crecimiento en 31 puntos o instantes de tiempo. En consecuencia, es de interés ajustar una curva suave a las observaciones que represente dicha función y que además considere el error asociado a las mediciones realizadas. Por las características del proceso de crecimiento, se requiere que la función sea

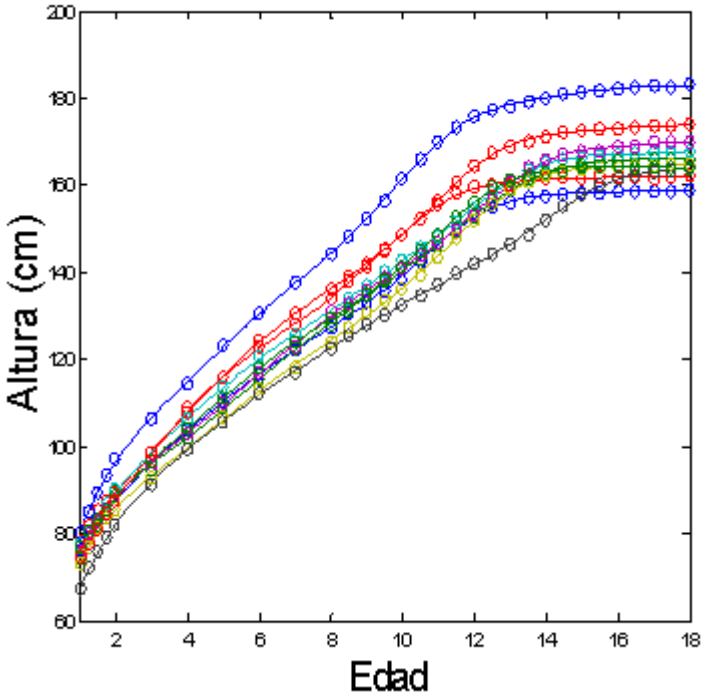


FIGURA 2.1: Mediciones de altura de 10 niñas.

monótona creciente, que sea dos veces diferenciable y que sus derivadas sean suaves para estudiar la velocidad y aceleración del crecimiento.

Utilizando una base Bspline se obtienen las curvas mostradas con los datos en la figura 2.1. Es posible considerar ahora estas curvas como datos y analizarlas desde el punto de vista funcional.

Ejemplo 2:

Consideremos la serie económica de la figura 2.2, que muestra el índice de producción mensual de bienes perecederos, es decir productos que caducan en un periodo menor a dos años, por ejemplo ropa, alimentos, cigarros, etc.

La serie corresponde a mediciones mensuales obtenidas del año de 1920 al 2000. A partir de esta serie se quiere analizar la dinámica asociada a la producción, con este propósito, podemos centrarnos en los registros de cada año pensándolos como observaciones funcionales y después comparar el comportamiento entre años.

Dada la tendencia exponencial de la serie, el análisis se realiza sobre el algoritmo del índice de producción. Una de las características importantes a determinar en la serie anual, es la tendencia por temporadas, en este caso se obtiene mayor información de este comportamiento utilizando las derivadas como se verá enseguida.

De manera general, se asocia la primera derivada (velocidad) con energía cinética y la segunda (aceleración) con energía potencial. En el caso de la producción de bienes, la energía cinética se asocia con el proceso de producción, momentos con alta energía cinética pueden interpretarse como un proceso de producción en pleno movimiento; mientras que la energía potencial se asocia con el capital disponible, bienes que están presentes y que pueden brindar a la actividad económica un cambio. Con esta interpretación de las derivadas es interesante ver como es el intercambio entre estos dos tipos de energía en un año determinado, esto se puede visualizar graficando la primera derivada contra la segunda. Por ejemplo para el año 1964 se tiene la figura 2.3, en la que se identifican dos grandes ciclos relacionados con ciclos de producción, el mayor que comienza en mayo (M) y término en agosto (A), y el segundo ciclo que comienza de octubre (O) y termina en diciembre (D).

Al comparar este tipo de gráficas entre años, es posible identificar en qué años se tuvo mayor energía en el proceso de producción así como contrastar los ciclos encontrados.

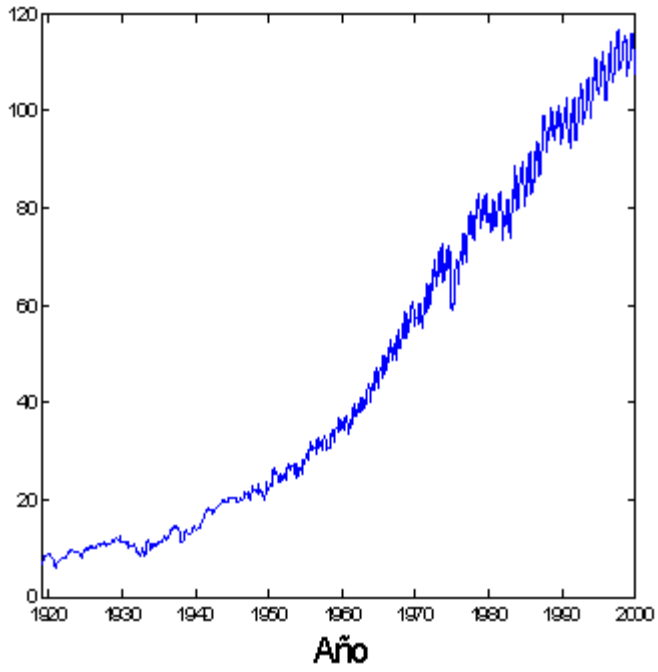


FIGURA 2.2: Índice mensual de bienes producidos.

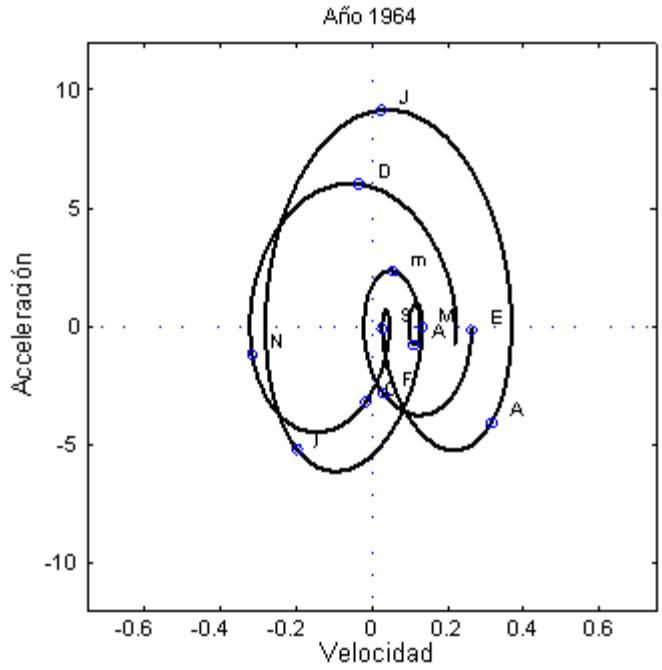


FIGURA 2.3: Gráfica de aceleración contra velocidad del índice de producción.

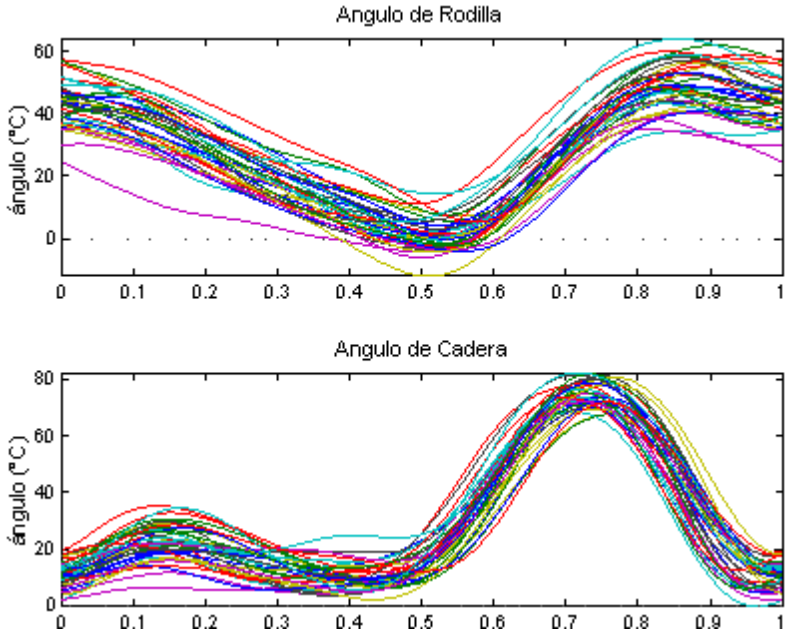


FIGURA 2.4: Angulos formados por la rodilla y cadera durante el ciclo de caminar en niños.

Ejemplo 3

Las observaciones funcionales también pueden considerarse como multivariadas, en el sentido de tener asociadas más de una curva a cada observación como en el siguiente ejemplo.

El laboratorio de análisis de movimiento en los niños, en el hospital de San Diego, recolectó los datos que se presentan en la figura 2.4 que consisten en los ángulos formados por la cadera y la rodilla de 39 niños durante el ciclo de caminar que comienza y termina cuando el talón bajo observación toca el piso. Como el tiempo es medido en términos del ciclo, cada curva es dada en términos del argumento $t \in [0, 1]$ (0 inicia ciclo, 1 termina). En consecuencia, ambos conjuntos de funciones son periódicos.

Podemos observar que el ángulo de la rodilla muestra un proceso de dos fases (los valores del ángulo de rodilla se aproximan a cero después aumentan su valor y se vuelven a acercar a cero, este comportamiento se presenta dos veces) mientras que el de la cadera es de una sola fase (los valores del ángulo de cadera decrecen

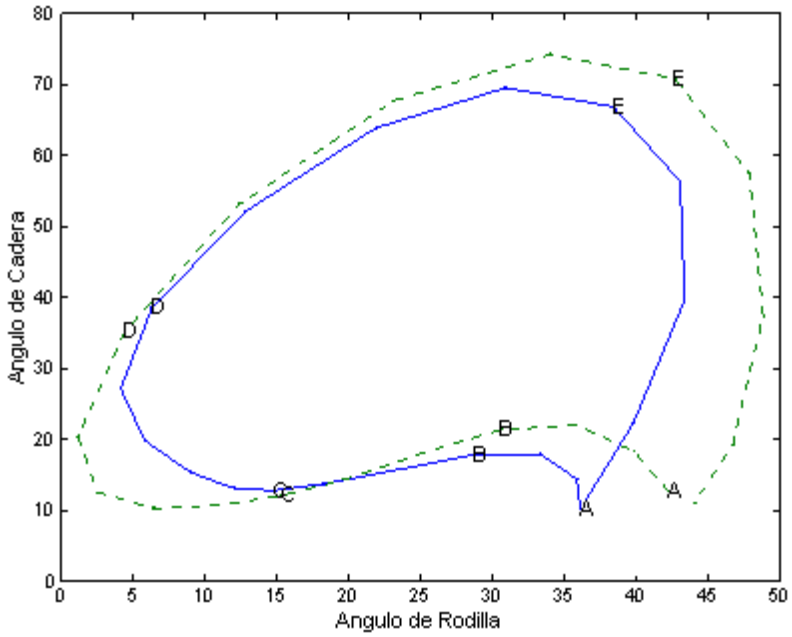


FIGURA 2.5: Angulo de rodilla y ángulo de cadera para un niño en particular comparado con el comportamiento promedio.

aproximándose a cero y después vuelven a aumentar su valor).

El carácter multivariado de estos datos se presenta en el hecho de que cada niño tiene asociadas dos curvas, una que describe el ángulo de su rodilla al caminar y otra que describe el ángulo de su cadera. De manera que es deseable analizar estos perfiles de manera conjunta

Para observar mejor la interacción de los dos ángulos se consideran gráficas como la mostrada en figura 2.5, en ella se muestra el ciclo de marcha para un solo niño, representado con la linea continua, comparado con el ciclo de marcha de la muestra promedio representado por la linea punteada. La naturaleza periódica del proceso ocasiona que se forme una curva cerrada. Las letras están colocadas en intervalos de un quinto del ciclo, de manera que el ciclo comienza y termina en la letra A.

En los ejemplos que hemos visto, la observación o dato asociado a cada individuo es una curva, o un conjunto de curvas, en el caso multivariado. No obstante, los datos que obtenemos en la práctica son vectores asociados a cada individuo y en

la siguiente sección se verá como pasar de las mediciones obtenidas a una representación funcional.

2.2 Representación funcional

El principio fundamental de análisis funcional de datos es pensar en las observaciones con que se cuenta como funciones, más que como secuencia de valores discretos. Dado que en la práctica los datos son registrados de manera discreta, estos se consideran como observaciones funcionales si es natural pensarlos como expresiones de una cierta función, que además pudiera considerarse suave en el sentido de que posea cierto número de derivadas.

En consecuencia dada una secuencia de datos discretos, el enfoque inicial es usarlos para estimar una curva que los represente y de la que sea posible calcular sus derivadas. Lo anterior con el objetivo de utilizar las curvas como observaciones y aplicar las técnicas de análisis funcional para analizarlas o hacer inferencia a partir de ellas.

Supongamos que tenemos n secuencias de observaciones discretas, cada una de ellas formada por n_i parejas (t_{ij}, y_{ij}), $j = 1, ..., n_i$,

$$\begin{gathered}((t_{11}, y_{11}), (t_{12}, y_{12}), \ldots, (t_{1n_1}, y_{1n_1}))\,, \\ ((t_{21}, y_{21}), (t_{22}, y_{22}), \ldots, (t_{2n_2}, y_{2n_2}))\,, \\ \vdots \\ ((t_{i1}, y_{11}), (t_{i2}, y_{i2}), \ldots, (t_{in_i}, y_{in_i}))\,,\end{gathered}$$

estas observaciones se suponen funcionales, si es posible pensar que las y_{ij} fueron generadas en t_{ij} por una función suave $x_i(t)$. No se requiere que los argumentos t_{ij} sean equiespaciados, ni que sean los mismos para cada observación. Por facilidad en la estimación de la curva $x(t)$, tomaremos en cuenta sólo una observación:

$$((t_1, y_1), (t_2, y_2)..., (t_N, y_N))\,.$$

Si debido a la naturaleza del proceso de medición los datos observados incluyen cierta variabilidad o un error asociado, entonces se tiene lo siguiente,

$$y_j = x(t_j) + \varepsilon_j, \tag{2.1}$$

donde los errores ε_j afectan la suavidad de la función x. Una alternativa en la estimación de x es filtrar este error, aunque también es posible ajustar una función x no tan suave, y requerir suavidad en los análisis que involucren la función estimada.

El método que se explicará para estimar la función x, es el método de bases de funciones, el cuál consiste en estimar x considerando la siguiente expansión,

$$x(t) = \sum_{k}^{K} c_k \phi_k(t) \tag{2.2}$$

en la que $\phi_k(t)$ es un conjunto de bases de funciones que se eligen previamente y c_k son los coeficientes a estimar para definir x. Usualmente estos coeficientes son estimados al minimizar la expresión

$$\sum_{j=1}^{n} [y_j - x(t_j)]^2 = \sum_{j=1}^{n} \left[y_j - \sum_{k=1}^{K} c_k \phi_k(t_j) \right]^2$$

mediante mínimos cuadrados. En caso de que se quiera establecer cierto grado de suavidad en x, obtenemos los coeficientes al minimizar,

$$\sum_{j} \{y_j - x(t_j)\}^2 + \lambda \int \{x''(t)\}^2 dt, \tag{2.3}$$

donde λ es un factor que controla la suavidad de la curva x medida mediante $\int \{x''(t)\}^2 dt$. Si λ es grande, $\lambda \to \infty$, la curva ajustada x resulta ser la regresión lineal estándar que corresponde a los datos observados. Si λ es pequeño, la curva tiende a ser más y más variable, de tal forma que cuando $\lambda \to 0$ la curva estimada x es un interpolador de los datos ($x(t_j) = y_j$ para todo j).

En algunos casos puede ser de interés el análisis de derivadas de órdenes más alto que dos, si por ejemplo se quisiera tener una representación suave de la derivada m, entonces debería considerarse el minimizar,

$$\sum_{j} \{y_j - x(t_j)\}^2 + \lambda \int \{x^{(m+2)}(t)\}^2 dt, \tag{2.4}$$

en lugar de la expresión (2.3). En [2] se presenta una descripción más técnica en la minimización y estimación de x.

Un factor importante a considerar en la estimación de x es la elección de las funciones base $\phi_k's$. En [6] se destacan tres puntos a considerar en la elección de una base:

▶ Que las funciones base tengan propiedades o características similares a las funciones que se desea estimar.

▶ Que se logre una buena estimación con un número de funciones base K relativamente pequeño

▶ Que las derivadas tengan un comportamiento razonable.

Algunas de las funciones base más usadas en la práctica son

- Fourier
- Spline
- Polinomial
- *Wavelets*

El sistema de base spline es una de las elecciones más comunes ya que representa una buena opción cuando se quiere aproximar funciones no periódicas, además ofrece flexibilidad y rápidez en la realización de cálculos.

Base Spline (B-splines)

Para definir el sistema de base spline, primero se describirá de manera general la estructura de una función spline. El primer paso en definir un spline es dividir el intervalo sobre el cual se quiere aproximar en L subintervalos definidos por los valores τ_l, $l = 1, ..., L-1$ llamados nodos o *breakpoints*. En cada intervalo (τ_l, τ_{l+1}), el spline es un polinomio de orden específico m. Los polinomios adyacentes se unen suavemente en el nodo que los separa, de tal forma que los valores de los polinomios son iguales en esta unión, al igual que sus derivadas hasta el orden $m-2$.

Un sistema de base spline consiste en considerar funciones $\phi_k(t)$ con las siguientes propiedades,

- Cada función base ϕ_k es una función spline definida por un orden m, y una secuencia de nodos τ.

- Cualquier combinación lineal de funciones spline es una función spline.
- Cualquier función spline definida por m y τ, puede ser expresada como una combinación lineal de estas funciones base.

Hay varias maneras de considerar estas funciones base $\phi_k(t)$, una de las más comunes es el sistema B-spline, que debe su popularidad principalmente a que presenta ventajas computacionales. En la figura 2.6, se muestran funciones base B-splines de orden $m = 2$ y $m = 4$ con siete nodos interiores. Además de las características mencionadas anteriormente, el sistema de funciones base B-spline satisface que cualquier función base es positiva sobre a lo mas m intervalos adyacentes, lo que garantiza rapidez en los cálculos aún utilizando una gran cantidad de funciones base.

Dado que los splines son construidos a partir de polinomios, el cálculo de sus derivadas es muy simple y se tiene que para un spline de orden m, el orden más alto de la derivada no trivial es $2m - 1$.

La elección más común en el orden del spline es cuatro, pues en este caso se tiene hasta la segunda derivada continua lo cual es razonable suponer en muchos casos, además visualmente una aproximación de orden 4 es suficientemente suave.

2.3 Estadísticas básicas de datos funcionales.

Las estadísticas clásicas que se calculan bajo un enfoque multivariado, pueden ser generalizadas para observaciones funcionales de la siguiente manera. Si contamos con un conjunto de N observaciones $x_1(t), x_2(t), ..., x_N(t)$, la función media es el promedio de las funciones puntualmente,

$$\overline{x}(t) = N^{-1} \sum_{i=1}^{N} x_i(t).$$

Similarmente, la función de varianza es

$$var_x(t) = (N-1)^{-1} \sum_{i=1}^{N} [x_i(t) - \overline{x}(t)]^2,$$

la desviación estándar, la raíz cuadrada de la varianza.

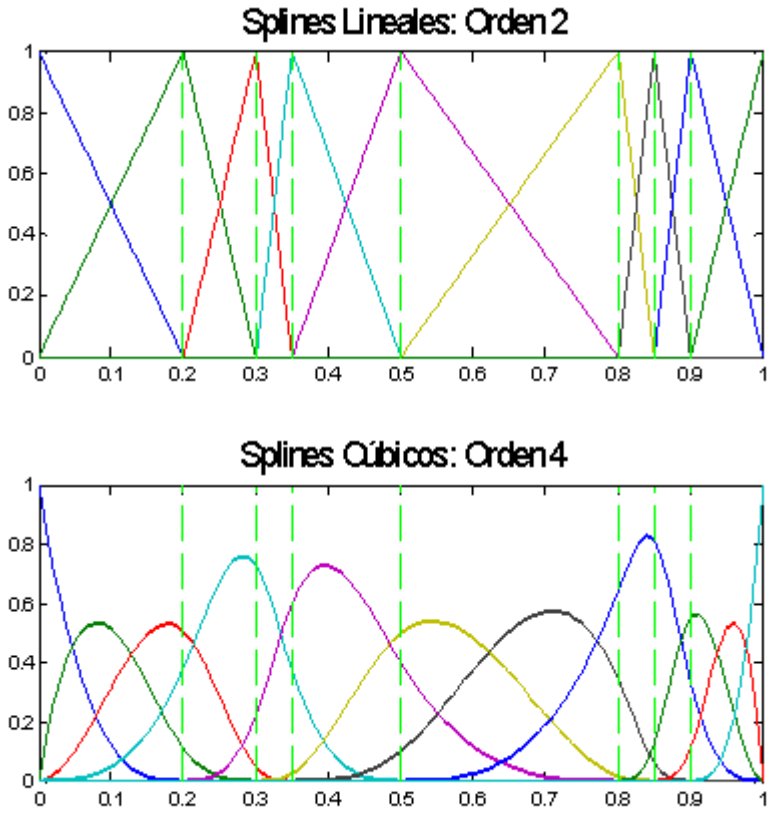


FIGURA 2.6: Funciones base B-splines.

La función de covarianza resume la dependencia entre diferentes valores de los argumentos, y se calcula para todo t_1 y t_2 como,

$$cov_X(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^{N} \{x_i(t_1) - \overline{x}(t_1)\}\{x_i(t_2) - \overline{x}(t_2)\}.$$

La función de correlación es

$$corr_X(t_1, t_2) = \frac{cov_X(t_1, t_2)}{\sqrt{var_X(t_1) var_X(t_2)}}.$$

En general, si tenemos pares de observaciones funcionales (x, z) la manera en que dependen una de la otra puede ser cuantificada por medio de la función de covarianza cruzada

$$cov_{X,Z}(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^{N} \{x_i(t_1) - \overline{x}(t_1)\}\{z_i(t_2) - \overline{z}(t_2)\},$$

o por la función de correlación cruzada

$$corr_{X,Y}(t_1, t_2) = \frac{cov_{X,Y}(t_1, t_2)}{\sqrt{var_X(t_1) var_Y(t_2)}}.$$

Derivadas y gráficas de cambio de fase

Una vez que se ha estimado la función x mediante (2.2), dependiendo de la elección de ϕ's se tiene cierta facilidad para calcular las derivadas. Se ha visto mediante ejemplos la importancia de las derivadas para extender la variedad de métodos gráficos exploratorios y también en el desarrollo de metodologías más detalladas. Información sobre el proceso que genera datos funcionales se puede encontrar examinando relaciones entre derivadas.

Graficar la derivada más alta contra la menor frecuentemente es informativo ya que podemos encontrar desviaciones de linealidad y porque la diferenciación puede exponer efectos no vistos fácilmente en las funciones originales.

Una de las herramientas más comunes es la gráfica de primera derivada contra la segunda derivada, llamada gráfica de cambio de fase. Esta gráfica nos da información de la energía asociada al proceso que describen los datos, ya que velocidad y aceleración se relacionan con energía cinética y potencial respectivamente.

Para ejemplificar consideremos la función $\sin(2\pi t)$ (figura 2.7) que describe un comportamiento armónico como la trayectoria descrita por un péndulo en movimiento, la oscilación del péndulo se debe a que hay un intercambio entre los dos tipos de energía, cinética y potencial. La energía potencial desaparece cuando el péndulo se encuentra de manera vertical, sin embargo éste volvera a subir gracias a la energía que le proporciona el movimiento, la energía cinética. El intercambio entre energía cinética y potencial puede analizarse mediante la gráfica de cambio de fase, figura 2.8, que en este caso describe un círculo por la uniformidad entre los dos tipos de energía.

La energía tiende a explicar comportamientos de ciertas variables de manera similar a las leyes físicas, por lo que podemos utilizar éste tipo de gráficas para estudiar el comportamiento del proceso que se este analizando.

En las gráficas de cambio de fase básicamente se busca identificar ciclos y hacer una asociación entre el radio de los ciclos con la energía asociada al proceso (mientras más grande el radio mayor es la energía) bajo análisis. Por ejemplo en la serie económica que se presentó antes, se utilizan este tipo de gráficas para analizar la energía asociada al sistema económico.

2.4 Componentes principales

El análisis de componentes principales es usado generalmente cuando se quiere determinar los principales modos de variación de los datos, con la intención de dar una interpretación a esta variabilidad o cuantificar las formas de variación más importantes que aproximarían a las observaciones.

De manera general la modificación principal en el desarrollo de componentes principales para el caso funcional, es cambiar las sumas que se utilizan en el contexto multivariado por integrales.

Para dar una explicación breve de componentes principales en el caso funcional, se recordará de manera general en que consiste el método de componentes principales desde el punto de vista multivariado.

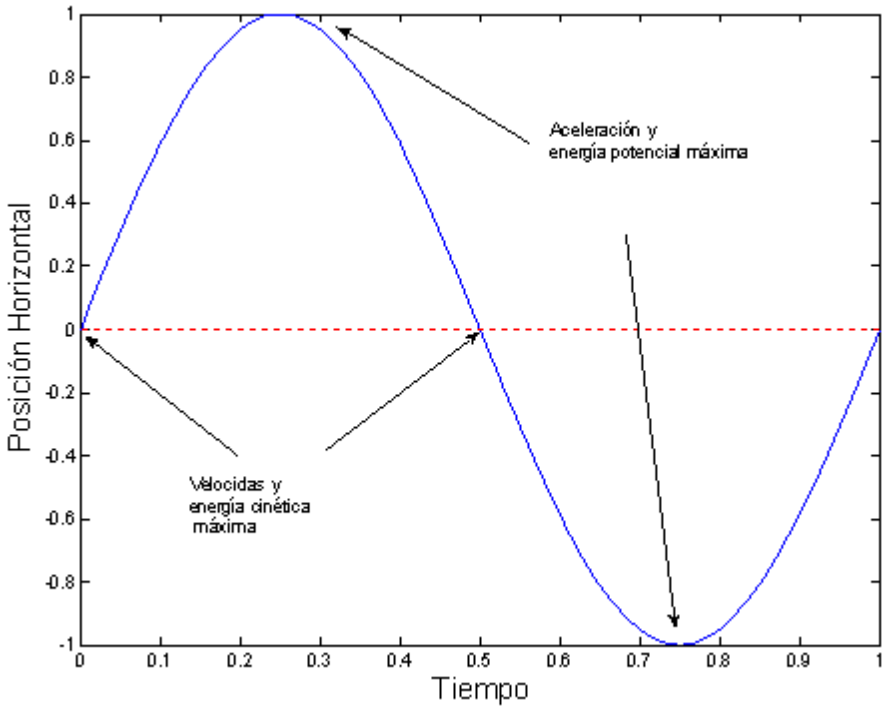


FIGURA 2.7: Función Sen(2π), que ilustra el movimiento de un péndulo.

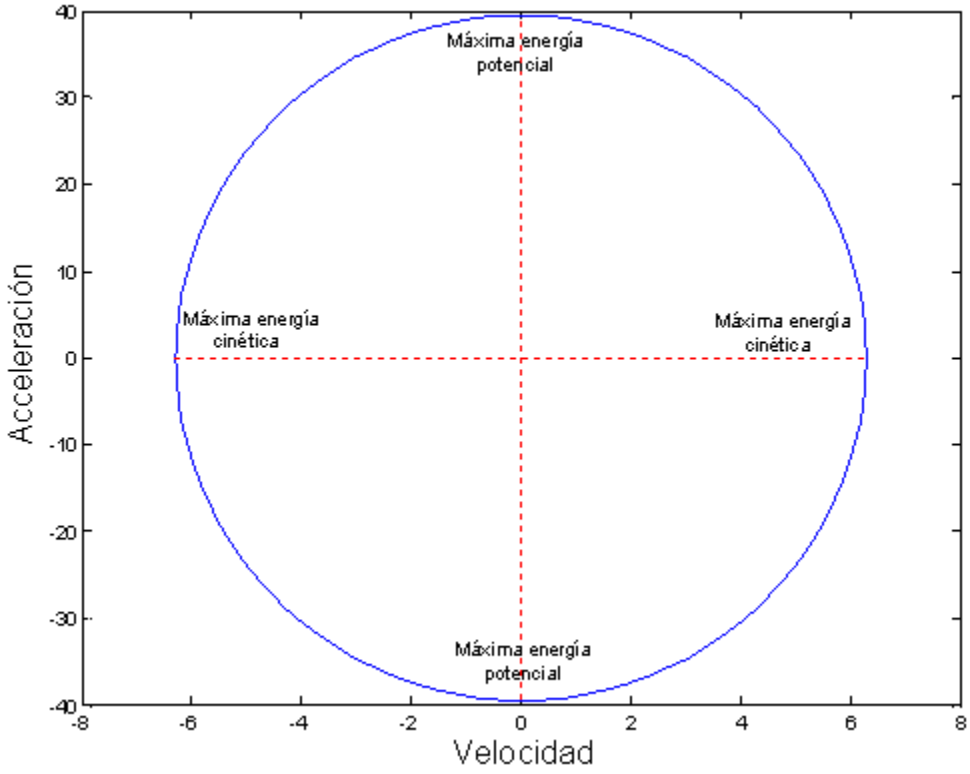


FIGURA 2.8: Comportamiento armónico de la función seno.

2.4.1 Componentes principales en análisis multivariado

Los componentes principales son definidos como combinaciones lineales de las variables originales, formando estas combinaciones de manera que se maximice su varianza, de la siguiente manera: a partir de un conjunto de N observaciones en p variables x_{ij}, $i = 1, .., N$; $j = 1, ..., p$, se quieren encontrar combinaciones lineales definidas mediante los coeficientes de p vectores $\boldsymbol{\xi}_k$, tal que, la varianza de las nuevas observaciones

$$\begin{aligned} s_{ik} &= \sum_{j}^{p} \xi_{kj} x_{ij}, \\ &= \langle \boldsymbol{\xi}_k, x_i \rangle \end{aligned}$$

se maximice en orden decreciente.

Es decir, primero se quiere determinar $\boldsymbol{\xi}_1$, maximizando la varianza de las variables s_{i1}, sujeto a la restricción $\|\boldsymbol{\xi}_1\| = 1$. Una vez calculado $\boldsymbol{\xi}_1$, se determina el vector $\boldsymbol{\xi}_2$, maximizando la variabilidad de las variables s_{i2}, con las restricciones $\|\boldsymbol{\xi}_2\| = 1$, y $\langle \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \rangle = 0$. Este procedimiento se repite sucesivamente (a lo más p veces) agregando en cada paso la restricción de que el vector a determinar $\boldsymbol{\xi}_j$, sea ortogonal a los vectores encontrados anteriormente.

Las restricciones $\left\|\boldsymbol{\xi}_j\right\| = 1$ son necesarias, ya que de no considerarlas se podría aumentar la variabilidad de las nuevas variables simplemente aumentando los elementos de los vectores ξ_k y el problema no quedaría bien definido. El objetivo de maximizar la variabilidad es identificar los más importantes modos de variación mediante las nuevas variables s_{ij} llamadas *scores* y las restricciones de ortogonalidad se imponen para que se identifiquen cosas nuevas a las encontradas previamente.

El problema de encontrar los vectores $\boldsymbol{\xi}_k$, es equivalente a encontrar los eigenvectores y eigenvalores de la matriz de covarianzas o matriz de correlación de las observaciones x_{ij} como se verá enseguida.

Consideremos las observaciones de manera que su media por variable es cero, $N^{-1} \sum_i x_{ij} = 0$, sea $\mathbf{X}$ la matriz con las observaciones x_{ij} y $\boldsymbol{\xi}$ un vector de longitud p, cuyos elementos son los pesos de una combinación lineal, entonces la varianza de las combinaciones lineales asociadas a $\boldsymbol{\xi}$, o varianza de los *scores*, es $N^{-1} (\xi' X' X \xi)$, donde $N^{-1} (X'X)$ es la matriz de covarianzas $\mathbf{V}$ de $\mathbf{X}$, por lo que el problema

anterior puede ser planteado como,

$$\max\left(\xi'\mathbf{V}\xi\right),$$
$$\text{s.a}\ \ \xi'\xi = 1$$

que se resuelve encontrando el mayor eigenvalor ρ, que satisface la ecuación

$$\mathbf{V}\boldsymbol{\xi} = \rho\boldsymbol{\xi}. \tag{2.5}$$

Hay una secuencia de eigenvalores y eigenvectores $\left(\rho_j, \boldsymbol{\xi}_j\right)$ que la satisfacen, de hecho hay a lo mas $\min\left\{p, N-1\right\}$ eigenvalores distintos de cero. Para cada j el eigenvector ξ_j satisface la ecuación (2.5) y además es ortogonal a todos los demás eigenvectores.

2.4.2 Componentes principales en análisis funcional

En el caso funcional, el análogo a las p variables es el argumento continuo s de una observación $x_i\left(s\right)$ y los vectores de peso que definen las combinaciones lineales $\boldsymbol{\xi}_k$, son remplazados por funciones de peso $\xi_k\left(s\right)$. Las combinaciones lineales o nuevas variables (*scores*) son equivalentes a la siguiente integral,

$$s_i = \int \xi_k\left(t\right) x_i\left(t\right) dt,$$

que también puede verse como producto interno en un espacio de funciones, $\left\langle \boldsymbol{\xi}_k\left(t\right), x_i\left(t\right)\right\rangle$.

Como en el análisis multivariado el cálculo de los componentes principales está asociado a resolver una ecuación similar a la de valores y vectores propios, para ver esto consideremos la siguiente función de covarianza,

$$v\left(s,t\right) = \frac{1}{N}\sum_{i=1}^{N} x_i\left(s\right) x_i\left(t\right). \tag{2.6}$$

Las funciones $\xi_k\left(s\right)$ que se desea calcular son soluciones de la ecuación

$$\int v\left(s,t\right)\xi\left(t\right) dt = \rho\xi\left(s\right),$$

definiendo,

$$V\xi = \int v\left(\cdot,t\right)\xi\left(t\right) dt,$$

donde V es llamado el operador covarianza, (2.6) se puede escribir como

$$V\xi = \rho\xi,$$

que tiene la misma forma que la ecuación (2.5), pero en este caso ξ es una eigenfunción, y ρ su eigenvalor asociado.

Una diferencia notable entre el caso funcional y el multivariado, es el número de componentes principales a considerar. Cuando se tienen vectores de dimensión p como observaciones, siguiendo el procedimiento descrito anteriormente, pueden determinarse a lo más $\min\{p, N-1\}$ vectores propios distintos de cero; en el caso funcional el argumento continuo s de la función, que puede tomar un número infinito de valores, es equivalente al número de variables. Sin embargo si las funciones $x_i(s)$ no son linealmente dependientes se tiene que el operador V es de rango $N-1$, y habrá solamente $N-1$ eigenvalores distintos de cero.

En términos de interpretación, las funciones $\xi_k(s)$ pueden considerarse como funciones de peso que dan mayor importancia a intervalos donde las observaciones presentan mayor variabilidad, dependiendo de la forma de estas funciones, es posible darles en una interpretación en el sentido de que resalten una característica relevante en los datos. A partir de las eigenfunciones se determinan los *scores* los cuales nos dan información sobre la importancia de la característica asociada a la eigenfunción en cada una de las observaciones.

Otra motivación para la determinación de los componentes principales es que este problema es equivalente a encontrar un conjunto K de funciones base tales que la expansión de la curva observada en términos de estas funciones base la aproxime tan cercanamente como sea posible. La expansión será de la forma,

$$\widehat{x}_i = \sum_{k=1}^{K} s_{ik}\xi_k(t).$$

es decir, al encontrar los componentes principales encontramos una base que minimiza el siguiente criterio,

$$PCASSE = \sum_{i=1}^{N} \|x_i - \widehat{x}_i\|^2$$

Rotación de las componentes

Se vió que los elementos $\xi_k(t)$ pueden tomarse como un conjunto de funciones base que aproximan bien a la curva registrada, sin embargo pueden existir otras funciones que también aproximen suficientemente bien a la curva. Si consideramos $(\xi_1, \xi_2, ..., \xi_k)'$ como un vector evaluado, entonces un conjunto de funciones que aproxima igual de bien a las curvas registradas son,

$$\varphi = \mathbf{T}\xi,$$

donde $\mathbf{T}$ es cualquier matriz ortonormal de orden K. Desde un punto de vista geométrico el vector φ es una rotación de las funciones ξ.

La ventaja de considerar la rotación, es que es posible que encontremos un conjunto de funciones rotadas que sean más fáciles de interpretar que las funciones ξ_k.

Una alternativa es considerar la rotación VARIMAX usada comunmente en análisis multivariado.

2.5 Modelos funcionales

Los modelos lineales funcionales se basan en los conceptos del modelo lineal clásico,

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.7}$$

donde $\mathbf{y}$ es el vector de observaciones, $\boldsymbol{\beta}$ es un vector de parámetros, $\mathbf{Z}$ una matriz que define una transformación lineal del espacio de parámetros al espacio de las observaciones y $\boldsymbol{\varepsilon}$ es un vector de errores con media cero.

En el enfoque funcional se pueden tener los siguientes casos:

1. Los parámetros y la variable respuesta son funciones.

2. La variable independiente o covariable es funcional y la variable dependiente escalar.

3. Tanto los parámetros como las variables independientes y la variable respuesta son funcionales.

Se explicará mediante un ejemplo el caso 2, en el cual se tienen n observaciones funcionales $x_i(t)$ a las que están asociadas n observaciones escalares y_i, y se quiere investigar la relación lineal entre ellas.

El ejemplo para ilustrar este caso fué tomado de [2]. En 35 estaciones meteorológicas de Canada se registró diariamente durante un año la temperatura, es decir se cuenta con 35 vectores de dimensión 365 que nos proporcionan información sobre como varía la temperatura en ciertos lugares de Canada a través del año. A estas observaciones se les ajustó una función mediante el métodos de bases de funciones, por lo que finalmente se tienen 35 observaciones funcionales (figura 2.9). Además, en cada estación meteorológica se tiene también información de la precipitación promedio anual. Se quiere investigar si a partir de los perfiles de temperatura es posible obtener de manera aproximada la cantidad de precipitación anual, para ello se propone ajustar un modelo lineal que relacione la cantidad de lluvia promedio anual en las ciudades con las variaciones de temperatura a través del año en la ciudad.

En este caso el modelo funcional toma la siguiente forma

$$y_i = \alpha + \int_0^T x_i(t)\,\beta(t)\,dt + \varepsilon_i, \tag{2.8}$$

en el que y_i denota el logaritmo de la precipitación anual, $x_i(t)$ la curva que describe el comportamiento de la temperatura anual en la estación meteorológica i (mostradas en la figura 2.9) y $\beta(t)$ es la función de regresión.

Una alternativa al modelo (2.8) es ajustar un modelo que relacione el perfil de temperatura a lo largo del año con el logaritmo de la precipitación anual, considerando el vector del los registros de temperatura diarios como variable dependiente,

$$y_i = \alpha + \sum_{j}^{365} x_{ij}\beta_j + \varepsilon.$$

Al hacer lo anterior simplemente consideraríamos una discretización de las funciones que describen el comportamiento de la temperatura. Sin embargo con esta discretización se tendrían 365 más la constante α, variables independientes y solamente 35 respuestas, entonces se podrían obtener muchos conjuntos $\{\beta_j\}$ de posibles soluciones. La figura 2.10, muestra una de estas soluciones, las letras sobre el eje

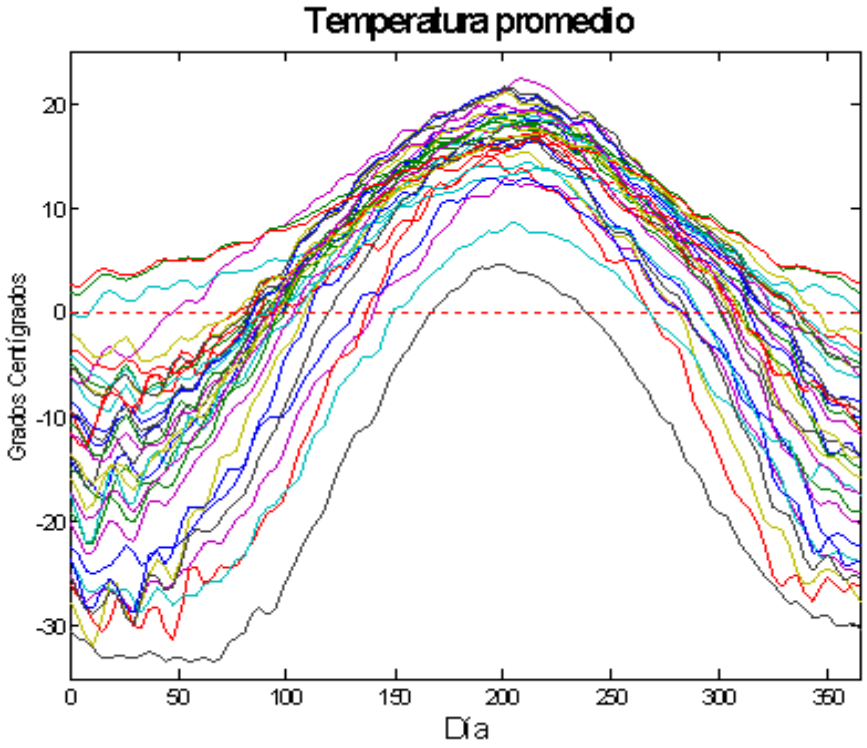


FIGURA 2.9: Temperaturas promedio registradas diariamente en 35 lugares de Canada.

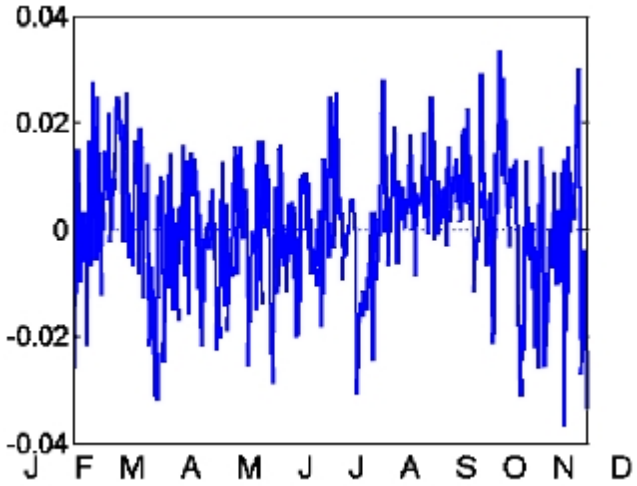


FIGURA 2.10: Aproximación a la función β, determinada al considerar un modelo discretizado diariamente.

representan las iniciales de los meses del año. Se observa que la solución presenta demasiada variabilidad, impidiendo dar una interpretación clara o sacar alguna conclusión.

Dado que al considerar una discretización con las 365 mediciones, se obtienen múltiples soluciones muy variables, puede pensarse una discretización más burda por ejemplo por meses, con lo cual se obtiene la figura 2.11 como solución al interpolar los valores del conjunto $\{\beta_j\}$. En este caso se obtiene un poco más de claridad en la solución aunque no se refleja de manera suave los cambios a través del año, y posiblemente se esté perdiendo información al considerar sólo la temperatura promedio en cada mes. Como el espacio de funciones que satisfacen (2.8) es infinito dimensional, independientemente del número de observaciones con que se cuente, al realizar la minimización usual de la suma de cuadrados de residuales

$$\sum_{i=1}^{N}(y_i - \alpha - \langle x_i, \beta\rangle)^2,$$

es posible que no se tenga un estimador consistente o fácil de entender. Para obtener un estimador más adecuado, se deben imponer ciertas restricciones al parámetro funcional β. Una opción es utilizar regularización, que consiste en minimizar la

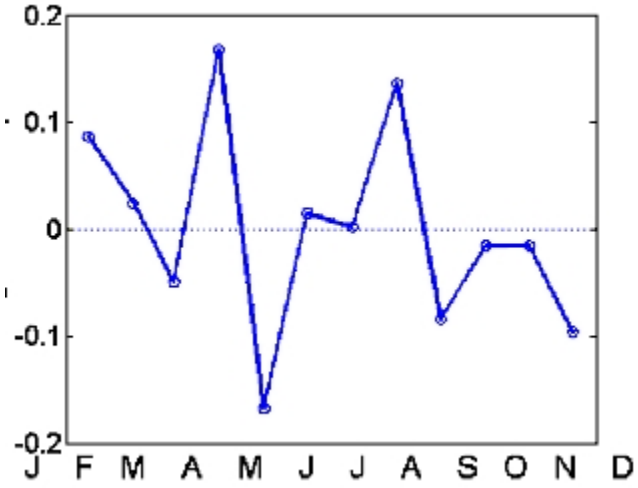


FIGURA 2.11: Aproximación de β, considerando un modelo discretizado por meses.

siguiente expresión

$$\text{PENSSE}_{\lambda}(\alpha,\beta)=\sum_{i=1}^{N}\left[y_i-\left(\alpha+\int_0^T x_i\left(t\right)\beta\left(t\right)dt\right)\right]^2+\lambda\int\left[L\beta(s)\right]^2ds, \qquad (2.9)$$

donde L es un operador diferencial que se elige de acuerdo al contexto del problema, generalmente se toma la segunda derivada para evadir excesiva fluctuación local en la función de regresión estimada.

Para el ejemplo de los perfiles de temperatura, se considera el operador lineal

$$L\beta=\left(\frac{2\pi}{365}\right)^2\beta'+\beta'''$$

porque se quiere que la función $\beta\left(t\right)$ sea periódica. El operador anterior es útil, ya que al aplicarlo a una función con un comportamiento periódico como seno, la anula.

Para minimizar (2.9), escogemos un valor de λ determinado subjetivamente o por un método como validación cruzada. La solución obtenida para el ejemplo utilizando el valor de λ que resulta de usar validación cruzada se presenta en la figura 2.12, con intervalos de confianza de 95% determinados puntualmente. Se observa que en días 100 a 250 las bandas de confianza incluyen al cero, lo que

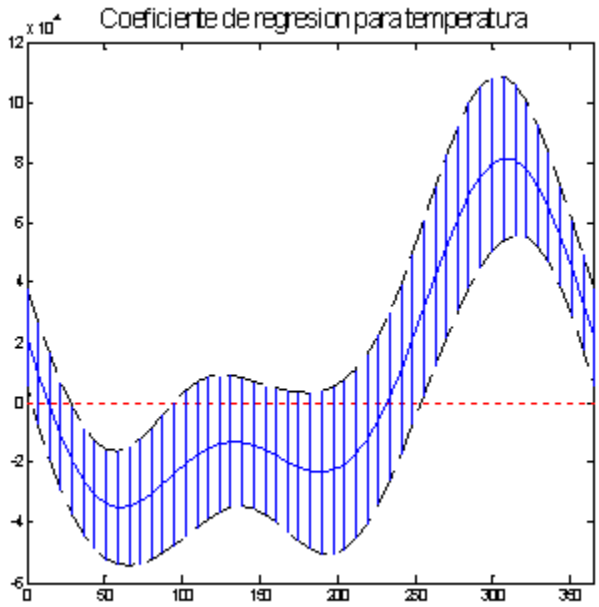


FIGURA 2.12: La función β estimada mediante regularización.

sugiere que en esos días la influencia de la temperatura sobre la precipitación no es importante, en contraste con algunos días al término del año, donde se tiene un mayor peso.

Una vez ajustado el modelo, podemos usar técnicas usuales para su evaluación, en la figura 2.13, se grafican los valores predichos por el modelo vs. los valores observados; con este modelo se obtiene un coeficiente de correlación de 0.75.

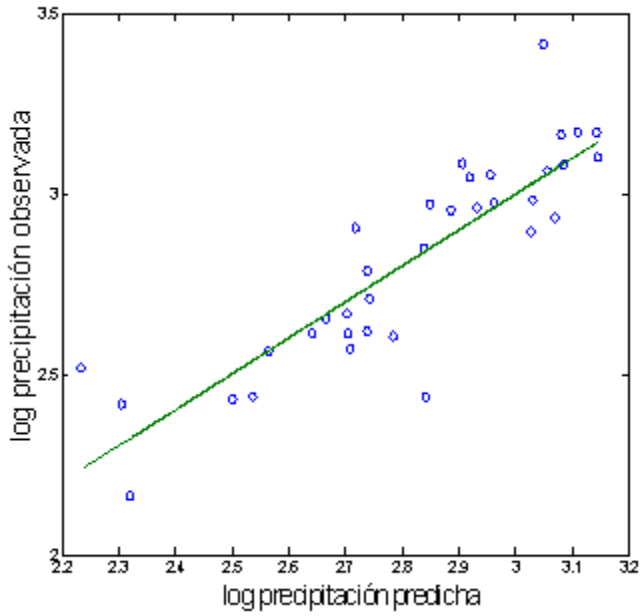


FIGURA 2.13: Valores observados contra valores predichos por el modelo.

Capítulo 3

Modelos aleatorios de olas

Las olas son extremadamente variables tanto en forma como en tamaño, su naturaleza exacta no es todavía en plenitud conocida pues la gran variedad de factores que influyen en su generación y comportamiento no permite dar una explicación sencilla de su generación.

El estudio de la forma y energía de las olas es de gran importancia principalmente en ingeniería marítima, por ejemplo para operaciones en puertos y en proyectos de diseño de estructuras, plataformas, barcos y soportes marinos, ya que estos deben de ser construidos para soportar la fuerza y velocidad inducida por ellas.

Para el análisis estadístico de las olas se requiere un buen proceso de recolección de datos. Generalmente la información y las mediciones de las olas se obtienen a partir de boyas, mediante sensores en plataformas estacionarias o bien de observaciones satelitales. Los datos resultado de mediciones por sensores en plataformas estacionarias y de boyas corresponden a mediciones de la altura del mar en ese punto. Por lo que finalmente se cuenta con información sobre la altura de las olas a lo largo del tiempo en un punto fijo.

Los datos que se utilizan en este trabajo son registros de alturas de olas tomadas en la plataforma de Alwyn norte, situada en el Mar del Norte, en aguas con profundidad de 130m. Se utilizaron tres sensores montados en la plataforma a una altura de entre 25 y 30 metros sobre el agua y el registro se realizó de manera continua con una frecuencia de muestreo de 5Hz (cada 0.2 seg).

Las mediciones se tomaron entre la medianoche del día 23 de diciembre y las

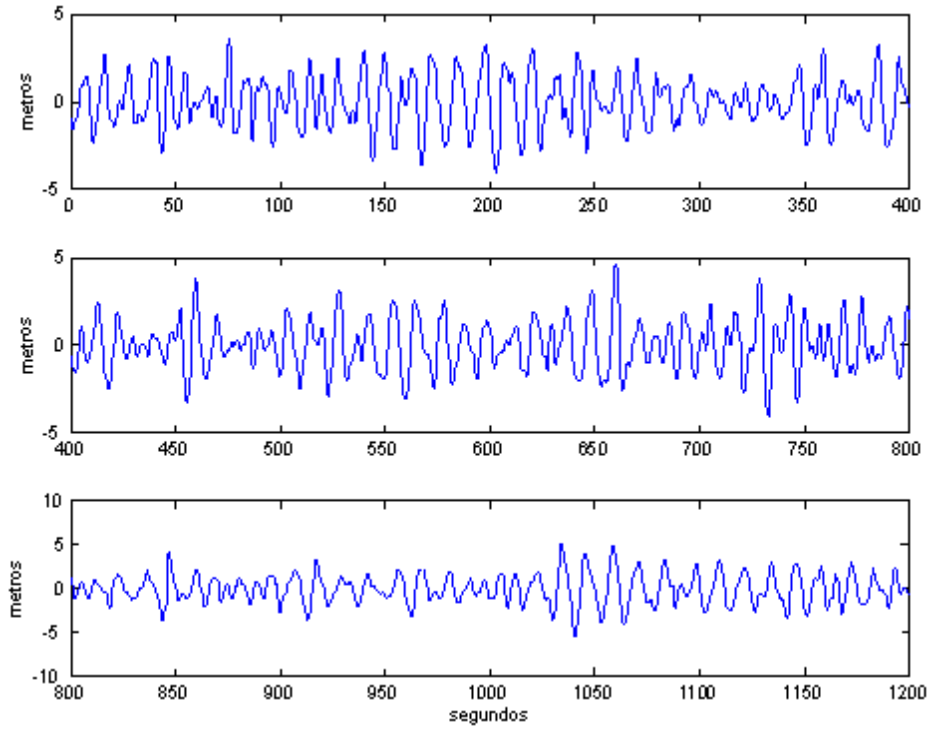


FIGURA 3.1: Un periodo de 20 min registrado el 16/11/97 de 07:33 a 07:53, en el Mar del Norte.

9 de la mañana del día 26 de diciembre de 1999 y consisten en 244 archivos que contienen los datos obtenidos durante 20 minutos. La figura 3.1 muestra un ejemplo de los datos para el primer periodo de 20 minutos.

Para su análisis usualmente cada conjunto de datos de 20 minutos se ve como una serie de tiempo, modelada como un proceso aleatorio $\eta(t)$, del cual generalmente se evalua su densidad espectral y características como altura significativa y periodo dominante. Enseguida se describirá brevemente en que consisten estos conceptos que son ampliamente usados en el estudio de las olas y se determinarán algunos de ellos para el conjunto de datos con que se cuenta.

3.1 Densidad espectral

La densidad espectral describe la distribución de la energía en las olas asociada a diferentes frecuencias en un periodo de tiempo, es decir, exhibe una separación en la energía total de acuerdo a las diferentes frecuencias presentes en el espectro.

El espectro para el proceso $\eta(t)$ puede verse como una medida que resume la descomposición del proceso en suma de olas elementales. Para una mejor descripción de la representación espectral consideremos la función de autocovarianza

$$R(t) = Cov(\eta(t), \eta(s+t)).$$

Como la altura de las olas se mide con respecto al nivel medio del mar, $E(\eta(s)) = 0$, entonces $R(t) = E(\eta(t)\eta(t+s))^2$. Dado que $R(t)$ es real y positiva definida, tiene asociada una trasformada de Fourier en términos de cosenos la cual es la función de distribución espectral $S(\omega)$

$$R(t) = \int_0^\infty \cos(\omega t)\, dS(\omega),$$

la derivada de $S(\omega)$, si existe, se le conoce como la densidad espectral o espectro $s(\omega)$. Si la función de autocovarianza es integrable, podemos determinar la distribución espectral invirtiendo la expresión anterior,

$$S(\omega) = \frac{2}{\pi} \int_0^\infty \cos(\omega t)\, R(t)\, dt.$$

Para obtener información resumida de la densidad espectral, se consideran los momentos espectrales. El momento de orden n es

$$m_n = \int_0^\infty \omega^n dS(\omega),$$

los momentos espectrales nos dan información sobre características o propiedades del proceso, por ejemplo, momentos de orden mayor a 2 están asociados a la regularidad de las trayectorias. También se tiene que $R(0) = Var(\eta(t)) = \int_0^\infty S(\omega)\, d\omega = m_0$ y como en general la energía es proporcional a $Var(\eta(t)) = E(\eta(t))^2$, se tiene que la integral del espectro representa la energía total.

Con los datos proporcionados se obtienen las densidades espectrales que se muestran en la figura 3.2, donde cada ventana muestra una cuarta parte de los 244 periodos de 20 min. Algunos de los momentos se presentan en la figura 3.3.

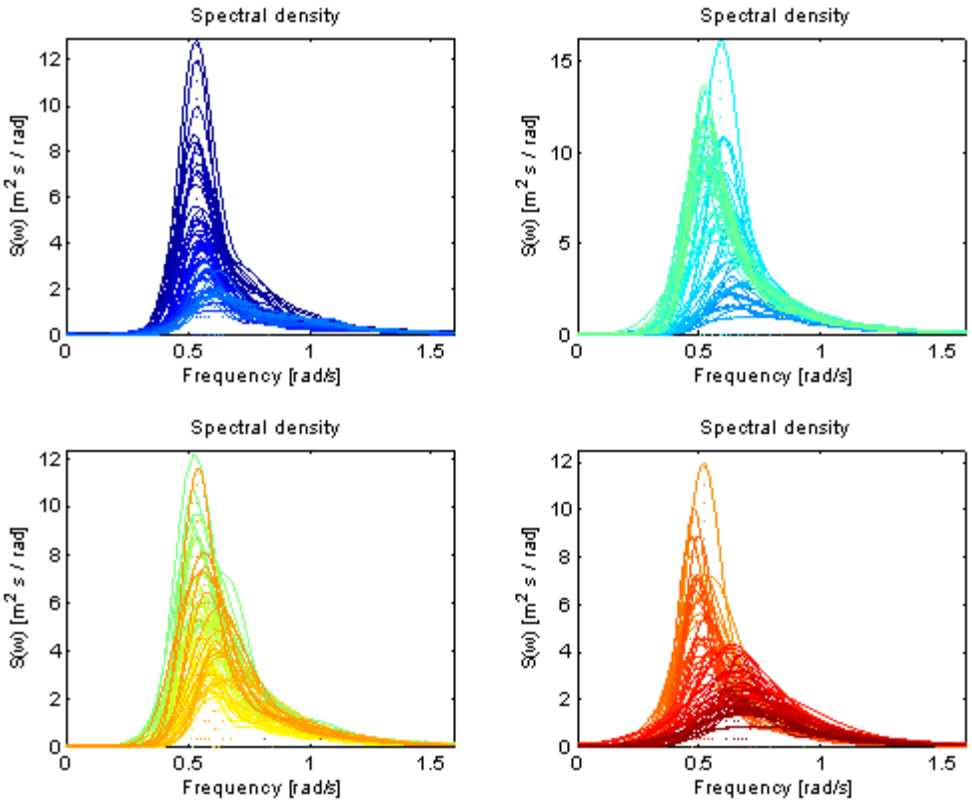


FIGURA 3.2: Densidades espectrales calculadas por periodos de 20 min.

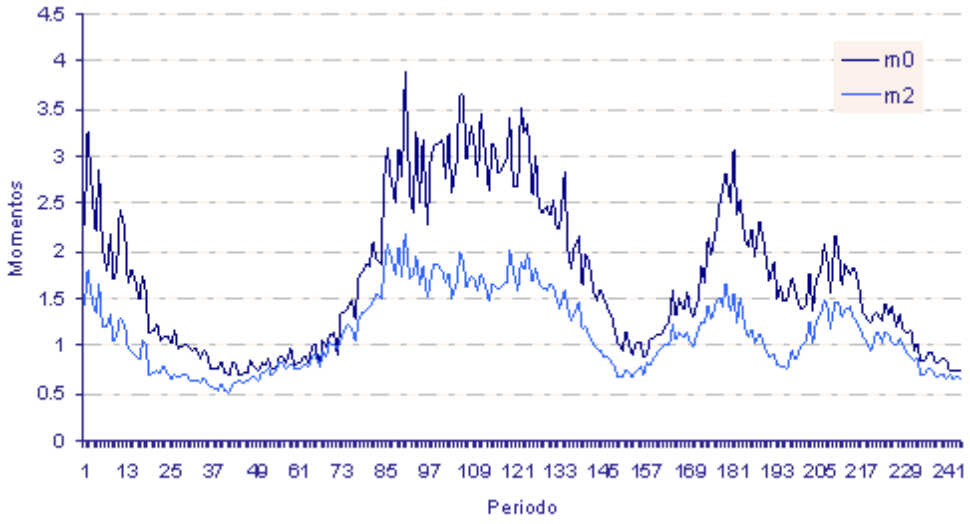


FIGURA 3.3: Momentos por periodo de los datos proporcionados.

3.2 Altura Significativa

Otra característica importante a determinar en el estudio de alturas de olas es la altura significativa. La altura significativa es una medida del estado general del mar e indica la altura de las olas más altas que es posible encontrarse durante un periodo razonable de tiempo.

Fue definida inicialmente como el promedio de 1/3 de las olas más altas observadas en una serie en un periodo de tiempo determinado. Con el desarrollo de tecnología en los instrumentos de medición del nivel de mar, hubo otros intentos de dar una definición más precisa del concepto. Actualmente la definición más aceptada es considerla como cuatro veces la desviación típica de la elevación del mar respecto a la media (cero),

$$Hs = 4\sqrt{var\left(\eta\left(t\right)\right)},$$

y en terminos del espectro la altura significativa se define como,

$$Hs = 4\sqrt{m_0}.$$

La manera más sencilla de estimarla a partir de los datos observados es a través

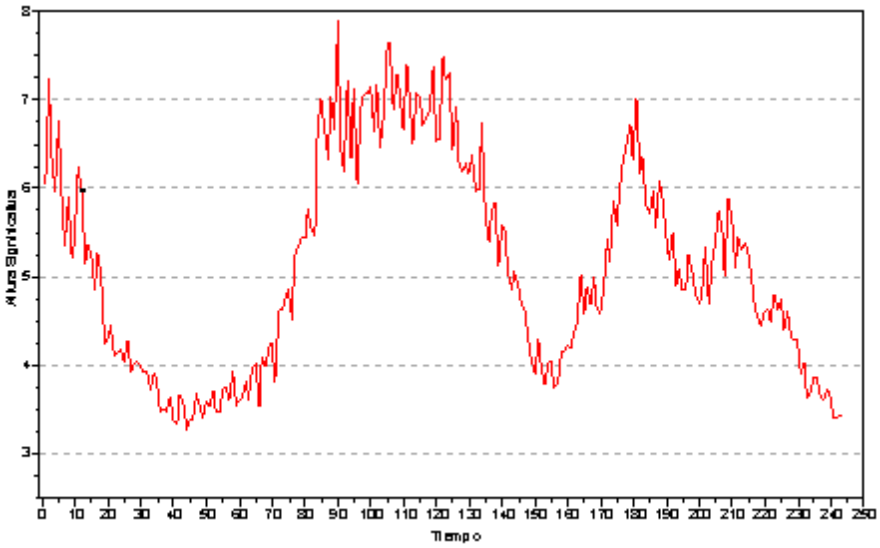


FIGURA 3.4: Altura significativa de los datos proporcionados.

de la desviación estándar muestral. Para estos datos fué calculada por periodos de 20 min y gráficamente se presenta en la figura 8. El registro comienza con una altura significativa alta (cerca de 7m), decrece a 3.5m durante un tiempo, sube a una altura significativa mayor a 6 y permanece en este estado durante un tiempo cercano a 20 hr, se calma por un periodo corto con altura significativa cercana a 4, sube nuevamente y finalmente decrece con una altura significativa de 3.5m

3.3 Proceso Gaussiano

Para el estudio del proceso aleatorio asociado a los datos frecuentemente se considera el supuesto de Gaussianidad, lo cual quiere decir que la distribución de la altura de la ola en un punto dado y en un instante de tiempo tiene distribución normal

$$P\left(\eta\left(t\right)\leq M\right)=\int_{-\infty}^{M}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{x^{2}}{2\sigma^{2}}\right\}dx$$

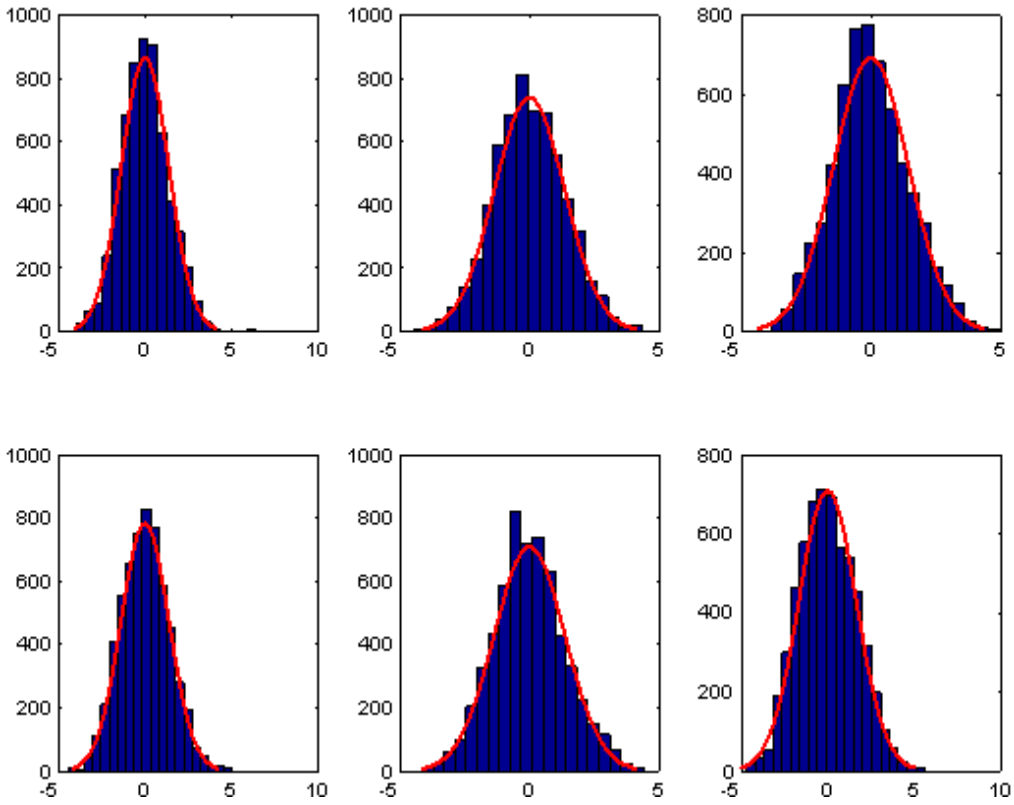


FIGURA 3.5: Histogramas para algunos de los periodos registrados de 20 min.

y que en cualesquiera instantes de tiempo $t_1, t_2, ..., t_n$ la distribución del vector $(\eta(t_1), \eta(t_2), ..., \eta(t_n))$ también es normal,

$$f_{t_1,t_2,...,t_n}(u_1, u_2, ..., u_n) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\mathbf{u}'\Sigma^{-1}\mathbf{u}\right\},$$

donde $\mathbf{u} = (u_1, u_2, .., u_n)$, $\Sigma = Cov(\eta(t_j), \eta(t_i))$.

Sin embargo en caso de fuertes tormentas aún en aguas profundas esta hipótesis puede no satisfacerse. Histogramas para los datos que se tienen se muestran en la figura 3.5 para 35 de los periodos de 20 minutos, en cada histograma se ajustó una densidad normal, en algunos casos el ajuste es adecuado, en otros casos los histogramas sugieren densidades más picudas o con colas mas pesadas.

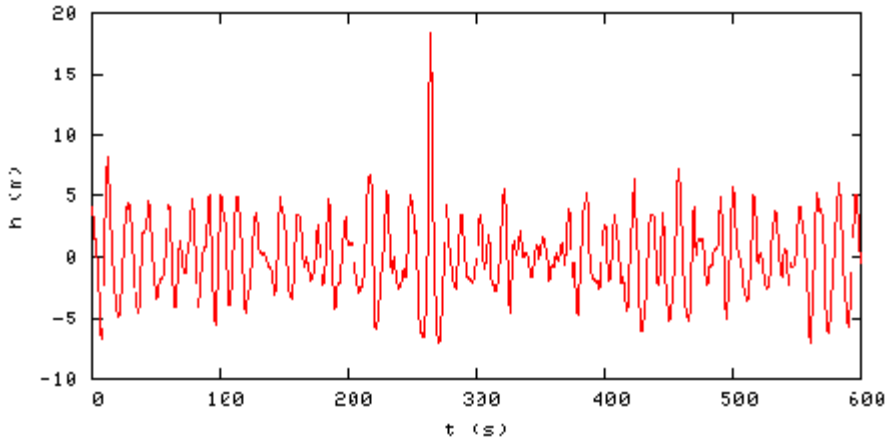


FIGURA 3.6: Ejemplo de *Freak wave* en el Mar del Norte.

3.4 *Freak waves*

Uno de los intereses en el estudio de las olas es la identificación de olas dañinas, este tipo de olas se denomina usualmente como *freak waves* o *rogue waves* y han sido la causa de pérdidas de barcos y vidas humanas. Se considera que ocurren raramente y que pueden generarse en cualquier parte del océano, aunque hay regiones más probables que otras. Su forma se caracteriza por tener una cresta o valle muy grande y tener grandes pendientes locales.

Como ejemplo de *freak waves* se presenta la figura 3.6, donde se observan registros de elevación del primero de enero de 1995 en la plataforma de Draupner del Mar del Norte, entre ellos se identifica una *freak wave* con una elevación de 18.5m, en un periodo con altura significativa de aproximadamente 12m.

Actualmente no se tiene mucha información sobre el mecanismo de generación u ocurrencia de este tipo de olas y no se tiene un citerio general para su identificación ni un acuerdo común para diferenciarlas de las olas extremas.

Harver(2000a) menciona que es es razonable definir *freak waves* como algo que esta más allá del conocimiento disponible para un diseño de rutina, lo cual implica que el criterio cambia con el tiempo y a medida de que se tenga mayor conocimiento sobre olas cada vez más extremas, es decir, cuando una ola *freak wave* sea completamente entendida, no habrá razón para seguir considerándola como tal y se tomará en

cuenta sólo como una ola extrema que es considerada en los diseños de estructuras y plataformas.

3.5 Datos de olas desde el punto de vista de análisis funcional

Otra manera de analizar los datos que definen el proceso $\eta(t)$ es separar esta serie en olas individuales, donde una ola queda definida por los cruces sucesivos hacia arriba de la media (cero).

Visualizando los datos de esta manera se obtiene un conjunto de olas para cada periodo de 20 min, por ejemplo las olas que integran uno de los periodos, estandarizadas en el intervalo [0,1], se muestran en la gráfica izquierda de la figura 3.7. Por medio de una transformación se obtuvo la gráfica de la derecha que presenta las mismas olas acomodadas de manera que el cruce central es 0.5 para todas; está gráfica muestra las olas en su forma típica y nos permite hacer una mejor comparación entre ellas.

Considerando los datos que consisten en conjuntos de olas estandarizadas en [0,1] y que cruzan en el punto medio 0.5 para cada periodo, el objetivo principal de este trabajo es realizar un análisis de la forma y variabilidad de las olas y ver su relación con otras características de interés como altura significativa.

Se quiere estudiar también como difieren las olas observadas de las olas de un proceso gaussiano y examinar la identificación empirica de *freak waves*.

Para realizar el análisis de forma y variabilidad, se emplearon las herramientas de análisis funcional de datos, pensando en los datos de olas como expresiones de una función que determina la altura en cualquier tiempo y que está siendo evaluada en los puntos muestreados.

Dado que lo que se quiere es estudiar las formas y las variaciones por periodos, el estudio de una función que nos muestre las formas de variación en los conjuntos de olas y el análisis a través de derivadas son aspectos importantes. Este último punto, por ejemplo, nos da información sobre pendientes y puntos de cambio que tienen una relación directa con la forma. El tratar los perfiles de las olas mediante análisis funcional proporciona facilidad para considerar estos dos aspectos.

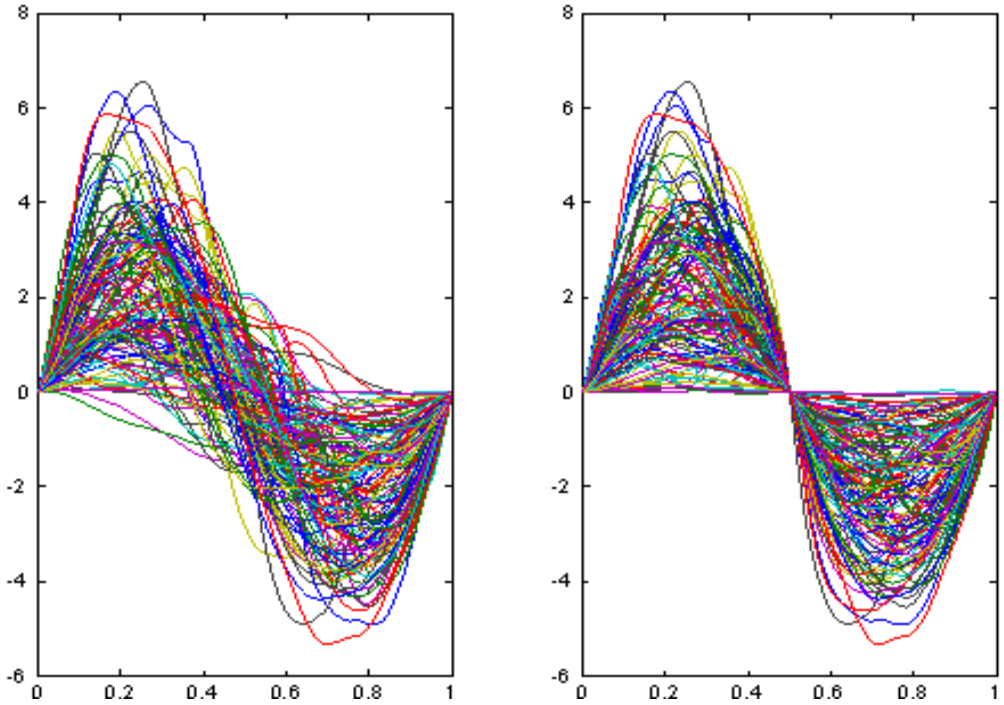


FIGURA 3.7: Olas que integran uno de los periodos de 20 min.

Como primer paso se utilizaron los datos que definen cada ola para tener una representación funcional de ella. Es decir se piensa en cada ola como una función que depende del tiempo (estandarizado a [0,1]) e indica la altura de la ola con referencia en cero (la media). Para determinar dicha función se utilizó el método de bases de funciones, tomando como base las funciones B-Splines. Como el instrumento de medición es muy preciso se busca que la función ajustada interpole practicamente a los datos observados.

Inicialmente se consideró la base de orden cuatro, con nodos en 300 puntos distribuidos en [0,1] de forma equispaciada y con K=302 funciones base. De esta manera se obtuvo practicamente una interpolación a los datos. Cuando se analizaron las derivadas de las funciones se utilizó una base de orden 6 y con valor positivo al parámetro λ que controla el suavizamiento.

Debido a la naturaleza periódica de los datos puede pensarse que utilizar la base Fourier es tal vez una mejor opción, sin embargo como se busca representar tan precisamente los datos para tomar en cuenta también pequeñas variaciones y no es el objetivo representar el caracter sinusoidal de las olas se consideró la base B-Spline.

Capítulo 4

Análisis inicial

El interés principal de este capítulo es hacer una exploración de los perfiles de olas en cada periodo y compararlos entre si, analizando como varian al cambiar la altura significativa del periodo.

Para llevar a cabo esta exploración, se emplearon las técnicas básicas del cálculo de la media, desviación estándar, exploración de derivadas y componentes principales. En algunos casos lo que se obtiene es una confirmación de resultados que eran de esperarse por la noción intuitiva de altura significativa, en otros casos se obtienen resultados gráficos no tan evidentes.

4.1 Media y varianza

A partir de la representación funcional de cada una de las olas que integran los periodos, se calcularon las estadísticas, media y desviación estándar, por intervalo de 20 minutos. Las figuras 4.1 y 4.2 muestran estas funciones asociadas con la altura significativa correspondiente al periodo. De esta manera se obtiene una ola promedio representativa por periodo, al igual que una función que indica la variabilidad en cada periodo y es posible establecer una comparación entre ellas.

Lo que resalta en la figura 4.1 es que a mayor altura en la cresta y mayor profundidad en el valle de la ola, mayor altura significativa. Respecto a la desviación estándar, periodos con altura significativa alta, en general, tienen mayor desviación estándar, aunque las desviaciones más grandes se presentan principalmente en la

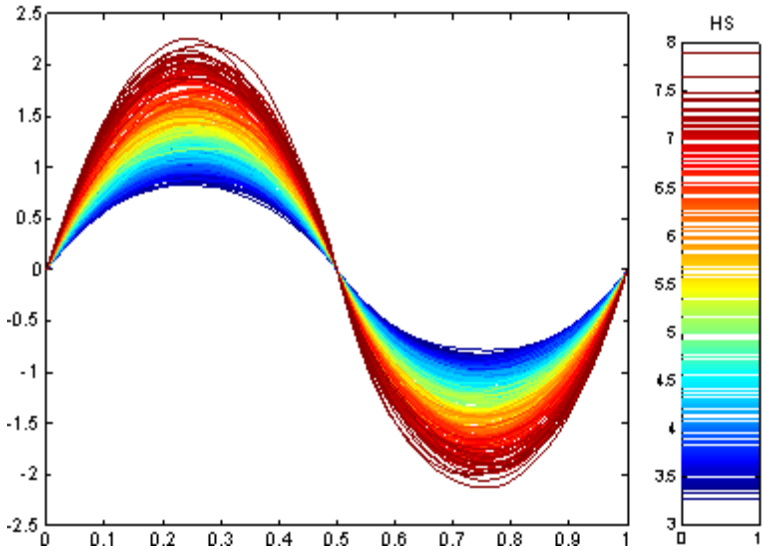


FIGURA 4.1: Olas promedio por periodo de 20 min.

cresta y valle de las olas. Debido a la asimetría de la figura 4.2, también se infiere que hay mayor variabilidad antes de que la ola cruce el nivel cero (en 0.5).

4.2 Derivadas

Uno de los aspectos importantes a analizar en la forma de las olas, es el comportamiento de las derivadas; para las olas promedio mostradas en la figura 4.1 se calcularon la primera y segunda derivada, las cuales se presentan en las figuras 4.3, 4.4.

En la gráfica de la primera derivada se observa, en general, que los intervalos con altura significativa mayor tienen pendientes más pronunciadas; inician con una pendiente alta en relación a los otros periodos, después del máximo correspondiente a la cresta de la ola (primer cruce de la derivada con el cero), las pendientes son más negativas y un comportamiento similar se tiene en la segunda mitad del intervalo [0,1]. La distinción más clara entre estas curvas se enfatiza en la pendiente de cruce, punto 0.5; es en este instante donde se presenta una diferencia mayor en la posición de una curva a la que corresponde una altura significativa no tan alta y una curva

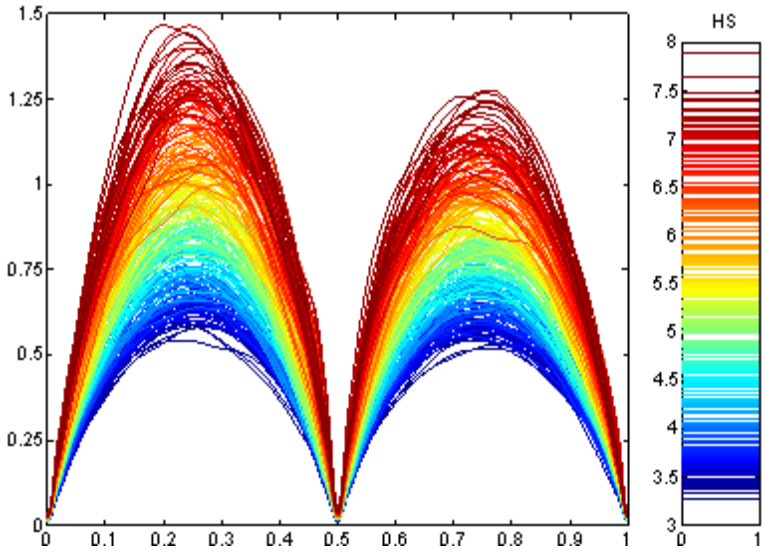


FIGURA 4.2: Desviación estándar de los periodos de 20 min.

con una altura significativa muy alta. Este comportamiento se observa mejor en la gráfica de la segunda derivada, donde se muestra que las curvas presentan un cambio más drástico en 0.5 a mayor altura significativa.

Los momentos donde se tiene mayor velocidad y aceleración, de acuerdo a las gráficas anteriores parecen estar relacionados claramente con la altura significativa del periódo, la figura 4.5, muestra tanto velocidad como aceleración para la ola promedio por periódo. Dada la relación de energía potencial con aceleración y energía cinética con velocidad, se observa en la figura que a mayor altura significativa mayor energía en el periodo puesto que las curvas con altura significativa alta están más alejadas, globalmente, de cero. La figura 4.6 indica como es la evolución en estas curvas respecto al tiempo, la ola comienza donde se presenta la etiqueta inicio y avanza en el sentido de las agujas del reloj. En la figura 4.5 se aprecia una mayor expansión de las curvas en la primera mitad, antes del cero respecto al eje de velocidad, lo que sugiere que las olas presentan mayor energía entre la cresta y el valle, de acuerdo a la figura 4.6.

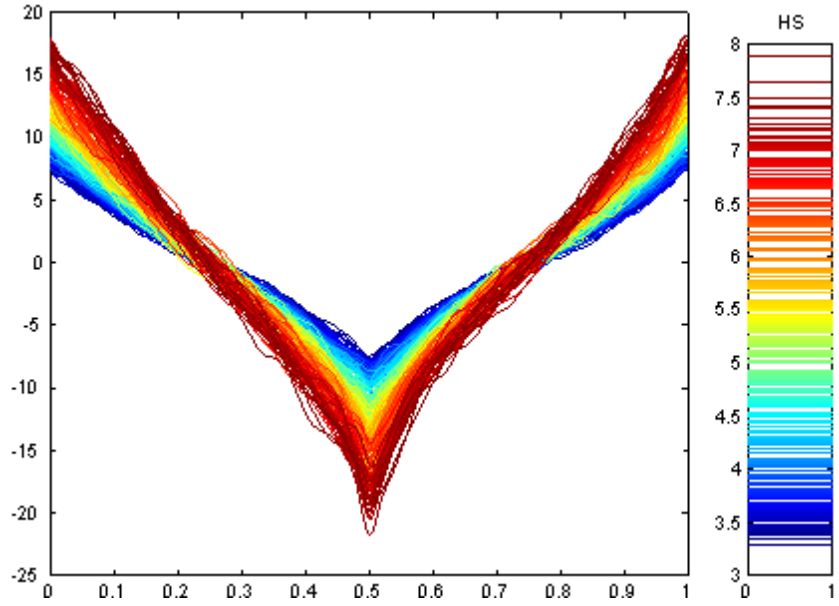


FIGURA 4.3: Primera derivada de olas promedio por periodo de 20 min.

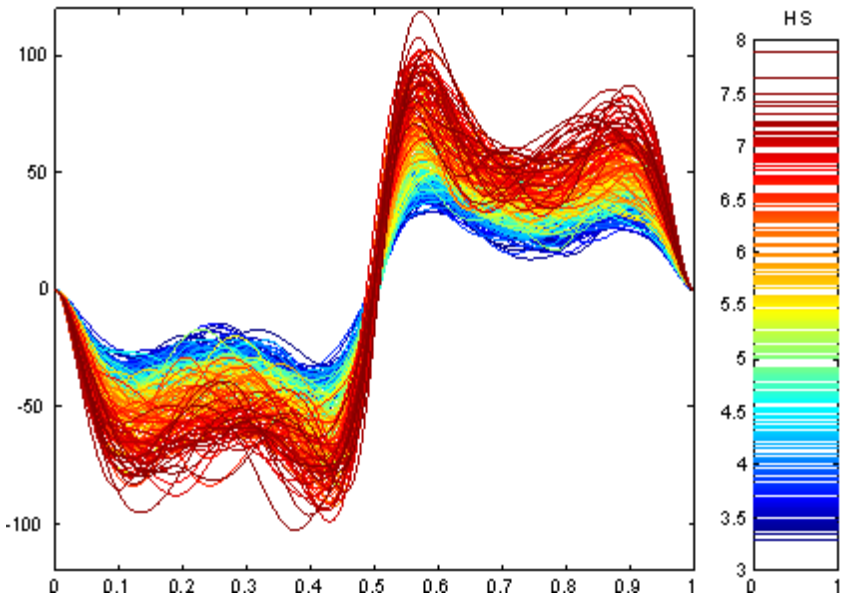


FIGURA 4.4: Segunda derivada de olas promedio por periodo de 20 min.

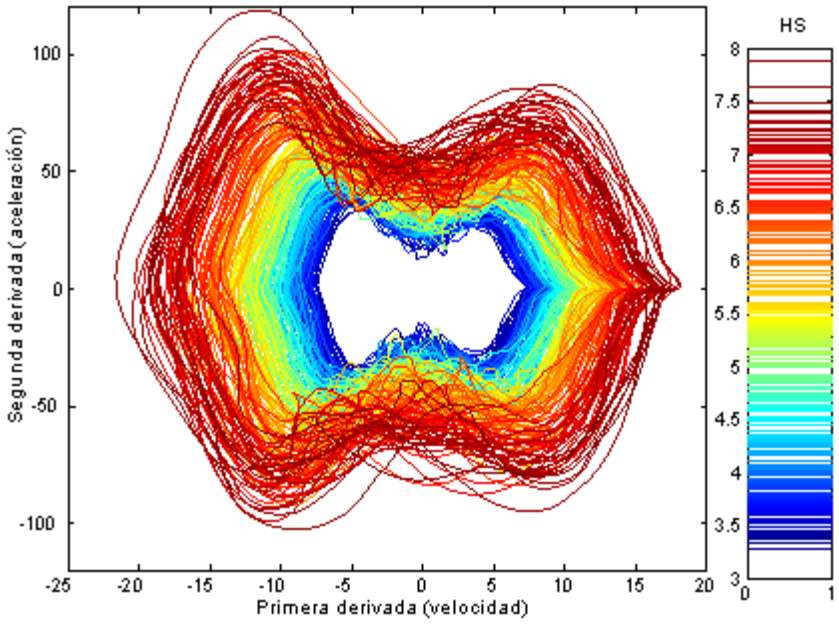


FIGURA 4.5: Gráficas de cambio de fase para cada periodo de 20 min.

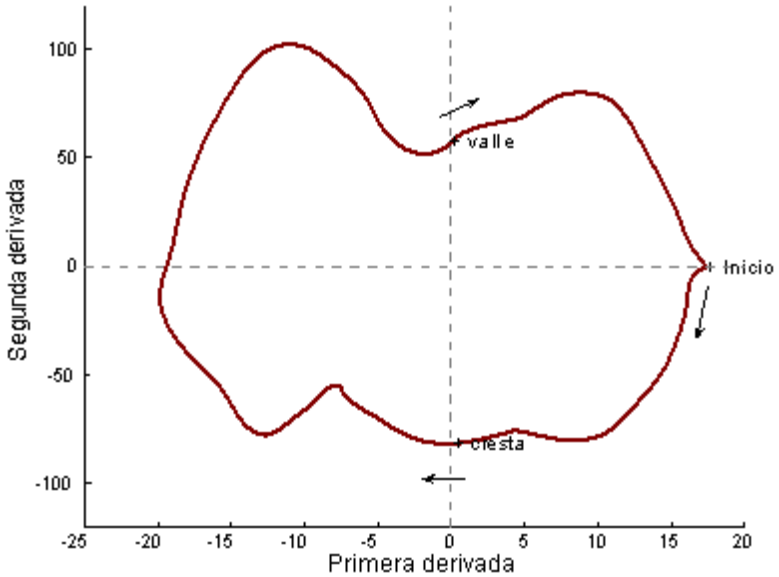


FIGURA 4.6: Gráfica de cambio de fase para un periodo. Las flechas indican la dirección en la que avanza el tiempo en [0,1].

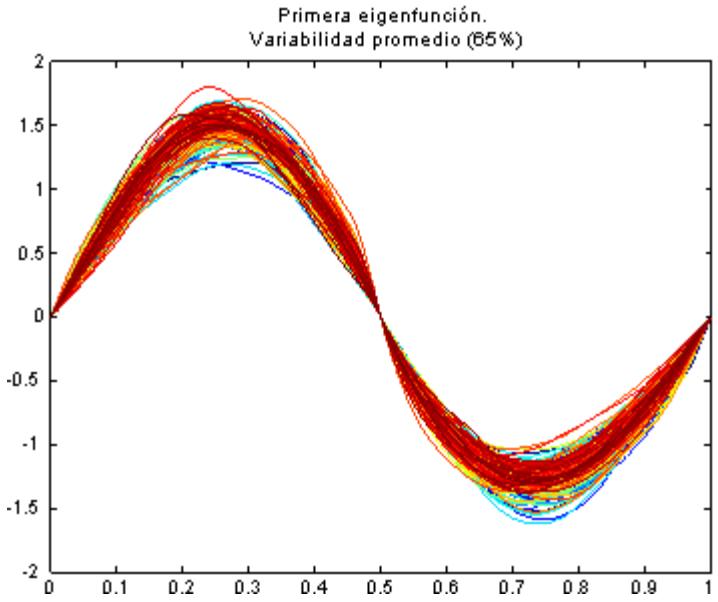


FIGURA 4.7: Primera eigenfunción para cada periodo de 20 min.

4.3 Componentes principales

Una pregunta natural en el análisis de las olas que conforman un periodo es su forma de variación. La información sobre la distribución de la variabilidad está asociada a las eigenfunciones que resultan de un análisis por componentes principales. Con el objetivo de explorar estos principales modos de variación se calcularon las tres primeras eigenfunciones en cada periodo mostradas en las figuras 4.7, 4.8 y 4.9, con ellas se obtiene, en promedio, cerca del 90% de la variabilidad total.

La primera eigenfunción, que representa el modo dominante de variación, explica en promedio hasta 65% de la variabilidad y su perfil es similar al de una curva sinusoidal, salvo que presenta ligera asimetría respecto al punto 0.5; es decir, las olas varian un poco más en el intervalo $[0, 0.05]$ que en el $[0.5, 1]$, y la variabilidad en $[0, 0.5]$ da mayor importancia a las variaciones de la cresta, mientras que en $[0.5, 1]$ las curvas estan distribuidas uniformemente, en el sentido de que son más redondas en todo el intervalo para la mayoría de los periodos. Debido a la forma de estos perfiles, *scores* altos asociados a la primera eigenfunción corresponden a olas altas en $[0, 0.05]$ y bajas en $[0.5, 1]$.

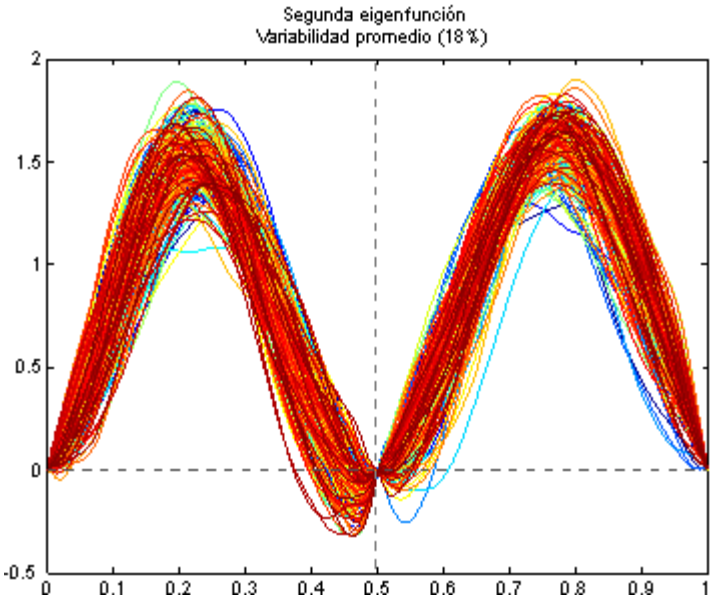


FIGURA 4.8: Segunda eigenfunción para cada periodo de 20 min.

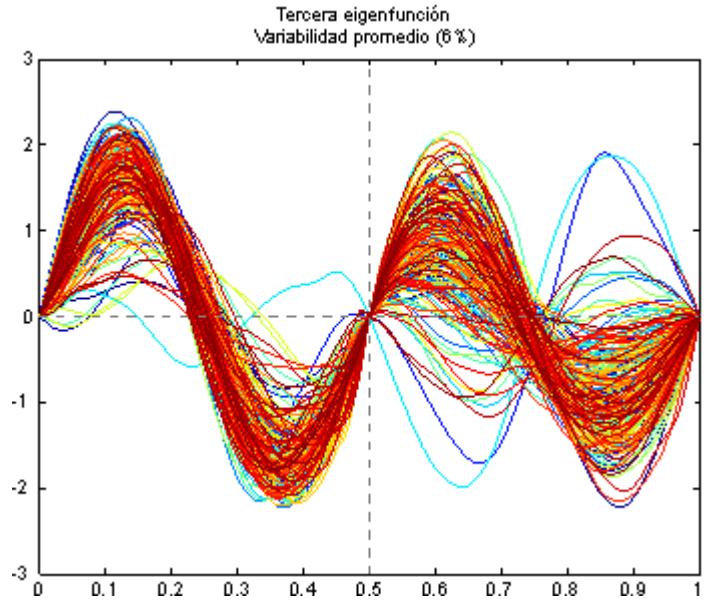


FIGURA 4.9: Tercera eigenfunción para cada periodo de 20 min.

La segunda eigenfunción representa en promedio 18% de la variación y da un mayor peso a la altura inicial y a la altura en el valle hacia el final. *Scores* altos asociados a la segunda eigenfunción corresponden a olas altas en casi todo el intervalo $[0, 0.4]$ pero bajas en aproximadamente $[0.4, 0.5]$, y altas en casi todo el intervalo $[0.5, 1]$, es decir olas altas al inicio pero con una pendiente de cruce en 0.5 no muy grande.

La variabilidad explicada por la tercera eigenfunción es 6% en promedio y presenta formas más variables entre periodos como se observa en la figura 4.9, de forma muy general marca un contraste entre las alturas al inicio y al final de los intervalos $[0, 0, 5]$ y $[0.5, 1]$.

Como pudo verse en las figuras de las eigenfunciones, no hay una relación aparente entre la altura significativa del periodo y la forma de la eigenfunción, es decir las eigenfunciones de los periodos conservan el mismo perfil sin importar la altura significativa. Sin embargo al analizar la proporción de variabilidad explicada por cada eigenfunción, ordenada de acuerdo a la altura significativa del periodo (figura 4.10), se presenta cierta tendencia: los puntos correspondientes a la segunda eigenfunción se observan un poco por debajo de la media conforme aumenta la altura significativa y al realizar una prueba de correlación entre la variable altura significativa y la variabilidad explicada en el periodo por el segundo componente se obtiene un nivel de significancia aceptable en contra de la hipótesis de no correlación; de hecho el porcentaje de variabilidad capturado con el primer componente también esta correlacionado significativamente con la altura significativa del periodo, como se muestra enseguida.

	Corr	P-value
HS. (% variabilidad alcanzado con la primera eigenfunción)	0.25	<0.0001
HS. (% variabilidad alcanzado con la segunda eigenfunción)	-0.36	<0.0001

Con base en lo anterior podemos decir que la forma de la primera eigenfunción captura mayor porcentaje de variabilidad en periodos con altura significativa alta, lo cual implica que las olas en estos periodos varian más de esta forma que en los que no tienen altura significativa tan grande. Y el porcentaje de variabilidad explicado

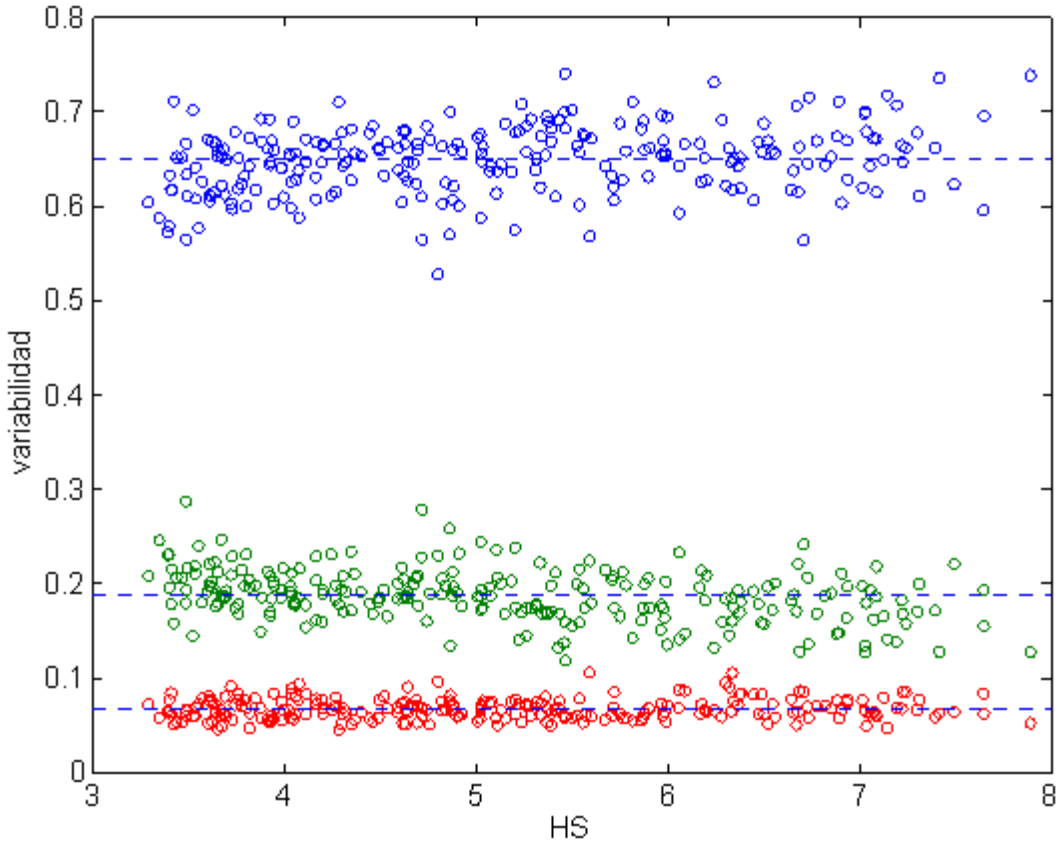


FIGURA 4.10: Proporción de variabilidad explicada por cada componente, ordenadas respecto a la altura significativa del periodo.

por el segundo componente es menor en periodos con mayor altura significativa, las olas no varian tanto de la manera que muestra en la figura 4.8 en relación a periodos con altura significativa baja.

Capítulo 5

Relación de Hs con características asociadas a la forma de la ola

Dada la importancia de la altura significativa como indicador del estado general del mar, se busca estudiar la forma de las olas promedio en relación con la altura significativa en un determinado periodo; con este propósito se ajustaron modelos que explican razonablemente la altura en función directa o indirecta de la forma de la ola promedio.

5.1 Hs con perfil de la ola promedio (periodo de 20 min). Modelo funcional

La manera más sencilla de buscar relacionar la forma de la ola con la altura significativa, es explorar la asociación de los perfiles de olas promedios con su altura significativa. Para hacer esto se propone un modelo funcional, en el que se trata como variable independiente a la función que representa la ola promedio en un periodo y como variable respuesta a la altura significativa correspondiente

$$Hs_i = \alpha + \int x_i(t)\beta(t)dt + \varepsilon_i \tag{5.1}$$

En el ajuste del modelo anterior, se consideró un parámetro de suavizamiento igual a 0.000001 para penalizar la suavidad del coeficiente de regresión $\beta(t)$. Este

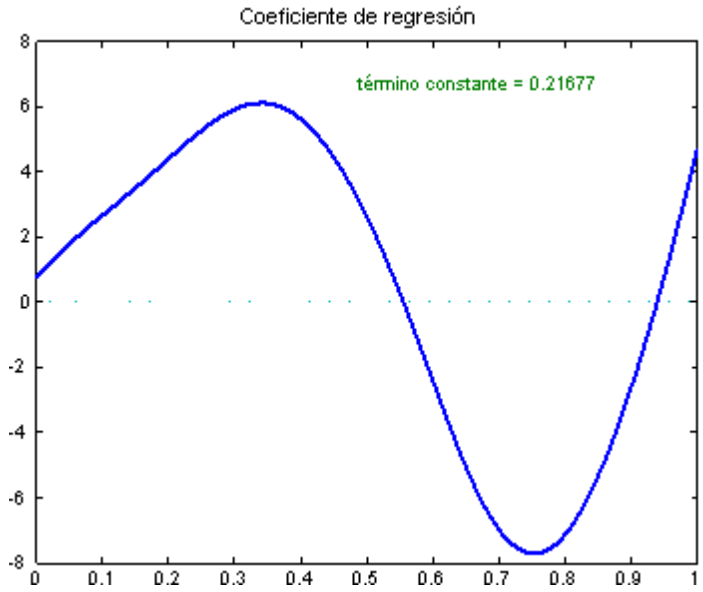


FIGURA 5.1: Función de regresión $\beta(t)$ con $\lambda = 0.000001$.

valor fue seleccionado observando el comportamiento de $\beta(t)$ ante diferentes opciones y eligiendo aquél con el que se obtienen resultados razonables en cuanto a suavidad. La función de regresión obtenida se muestra en la figura 5.1 y en la figura 5.2 con bandas de confianza de 95%. Se observa que esta función da mayor importancia a la profundidad del valle en la ola que a la cresta. Las bandas de confianza que incluyen al cero indican que la función β en esa región no es significativa. Entonces, observando la figura 5.2 puede decirse que las regiones de influencia en la forma promedio de la ola para explicar la altura significativa, son: la cresta, la pendiente antes de cruzar el nivel cero y el valle. Una ola promedio con cresta, valle y pendiente de caída altos, se espera que este asociada a un periodo de 20 min con altura significativa grande. Dado que el parámetro α en el modelo es positivo, este término puede interpretarse como una base que se agrega a la ponderación

$$\int_0^1 x_i(t)\beta(t)dt$$

La figura 5.3, presenta la altura significativa predicha contra la observada. En general estas dos cantidades se aproximan bien, como también lo refleja el coeficiente

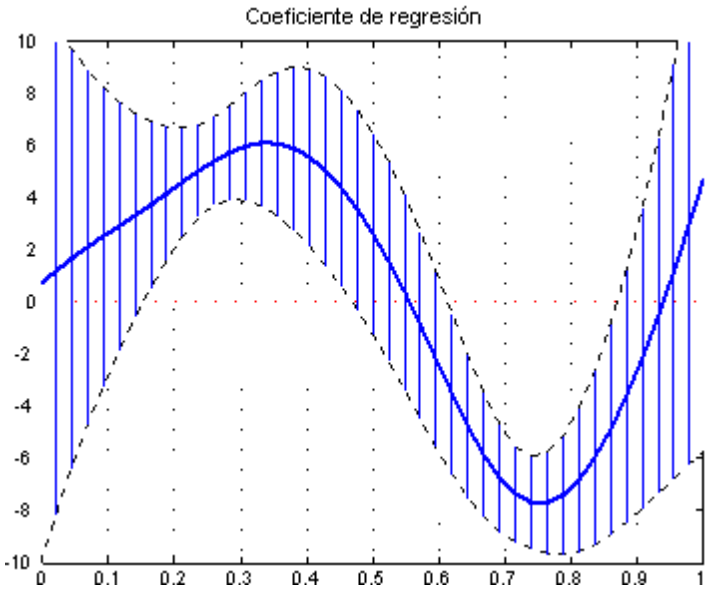


FIGURA 5.2: Función de regresión $\beta(t)$ con bandas de confianza 95%.

R^2, concluyendo que el modelo presenta un buen ajuste, aunque se tiene el siguiente inconveniente. En la figura 5.4 se muestran los residuales correspondientes a cada periodo, ordenados de menor a mayor altura significativa; se observa que a mayor altura significativa hay mayor dispersión en los residuales y por lo tanto mayor incertidumbre en el modelo y la predicción.

5.2 Hs con primer score del componente principal asociado al conjunto de olas promedio por periodo de 20 min.

Considerando como observaciones funcionales las olas promedio en cada periodo de 20 minutos, (figura 4.1) se hizo un análisis por componentes principales, obteniendo que la primera eigenfunción captura hasta el 98% de la variabilidad total. En la figura 5.5, se muestra la media de las olas promedio más y menos un múltiplo de la primera eigenfunción. Esta gráfica confirma que la variabilidad entre las olas

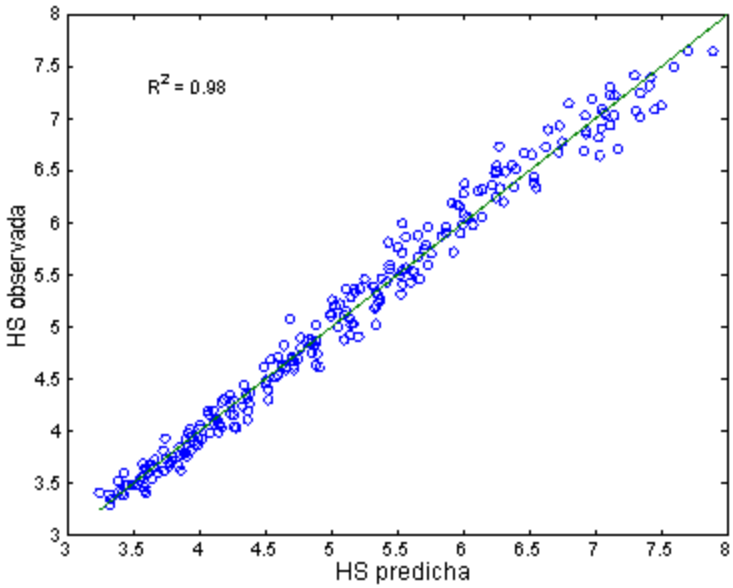


FIGURA 5.3: Evaluación del ajuste del modelo (5.1).

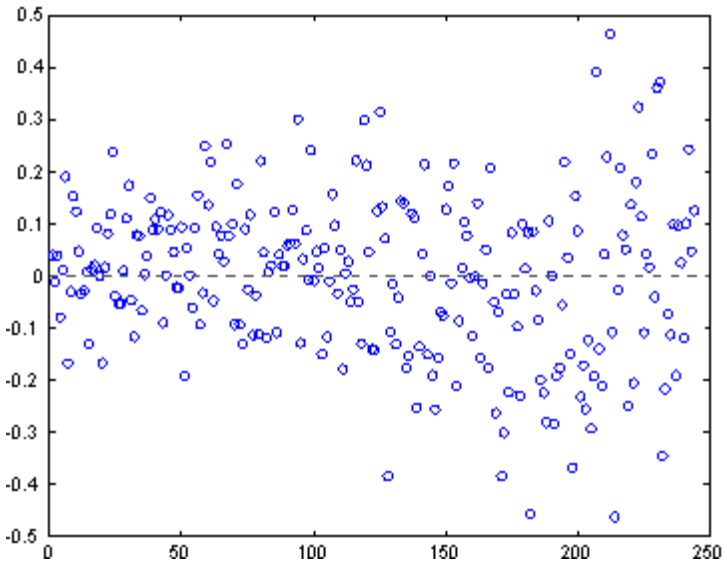


FIGURA 5.4: Residuales ordenados por periodo de menor a mayor altura significativa.

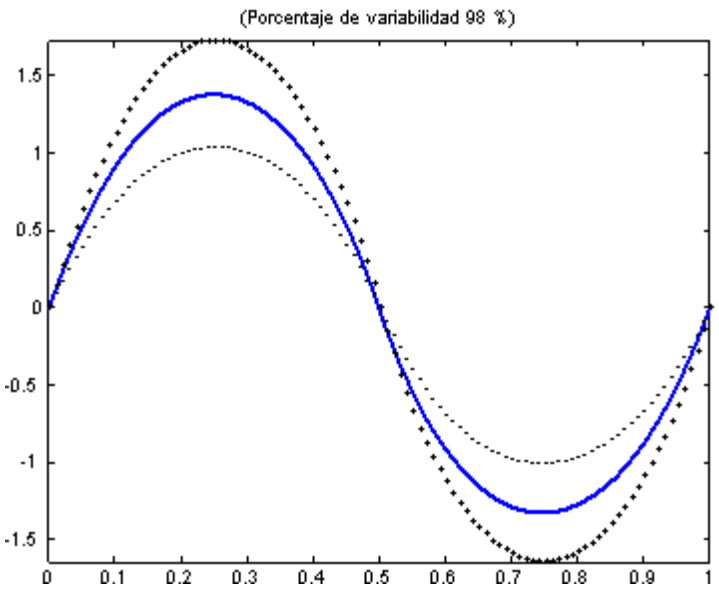


FIGURA 5.5: Media de olas promedio y curvas resultado de sumar y restar un múltiplo de la primera eigenfunción.

promedio se presenta uniformemente en la cresta y valle de forma simétrica, como también se observa en la figura 4.1. Dada la forma de la eigenfunción, *scores* altos estarán relacionados con olas promedio altas y con valle más profundo, por lo que sería natural pensar que este *score* está asociado a la altura significativa del periodo.

La figura 8 presenta el *score* de cada ola promedio, la grafica muestra un comportamiento similar al de la altura significativa de la figura 8, así que efectivamente se tiene una correlación entre estas cantidades. Para modelarla inicialmente se propuso el siguiente modelo

$$Hs = 5.12 + 4.82\text{Scr1} + \varepsilon, \tag{5.2}$$

el cual resulta significativo, y además con él se obtiene un muy buen ajuste ($R^2 = 0.98279$). Los residuales para cada periodo ordenados de menor a mayor altura significativa, se muestran en la figura 5.7 y como en el modelo funcional, presentan mayor dispersión al aumentar la altura significativa. Con el propósito de estabilizar la variabilidad de los residuales, se hizo una trasformación en la respuesta, considerando ahora $\log Hs$. Se identificó una tendencia cuadrática (figura 5.8), por lo

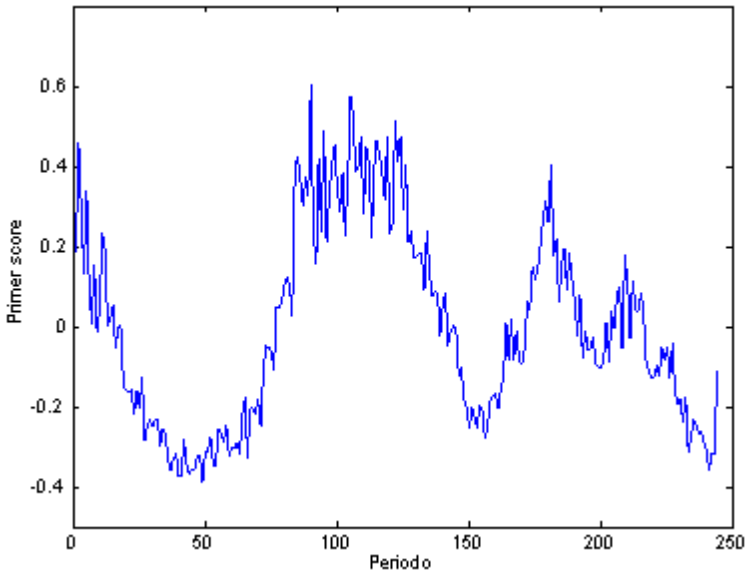


FIGURA 5.6: *Score* asociado al primer componente principal.

que se ajustó el siguiente modelo:

$$\log Hs = 1.63 + 0.98\text{Scr1} - 0.50\,(\text{Scr1})^2 \tag{5.3}$$

al que corresponde el análisis que se presenta a continuación:

Termino	**Estimación**	**Error std.**	**Razón t**	**Prob >|t|**
Intercepto	1.637	0.0025	650.36	0.0000
Scr 1	0.98	0.0078	125.39	< .0001
(Scr 1)^2	-0.51	0.0290	-17.30	< .0001

	df	**Suma de cuadrados**	**Cuadrados medios**	**Razón F**
Modelo	2	13.058	6.53	8373.815
Error	241	0.188	0.00078	Prob>F
C.Tot	243	13.246		<.0001

Al igual que en el modelo anterior se obtiene un muy buen ajuste, con un coeficiente $R^2 = 0.983$; pero en este caso los residuales no incrementan su dispersión

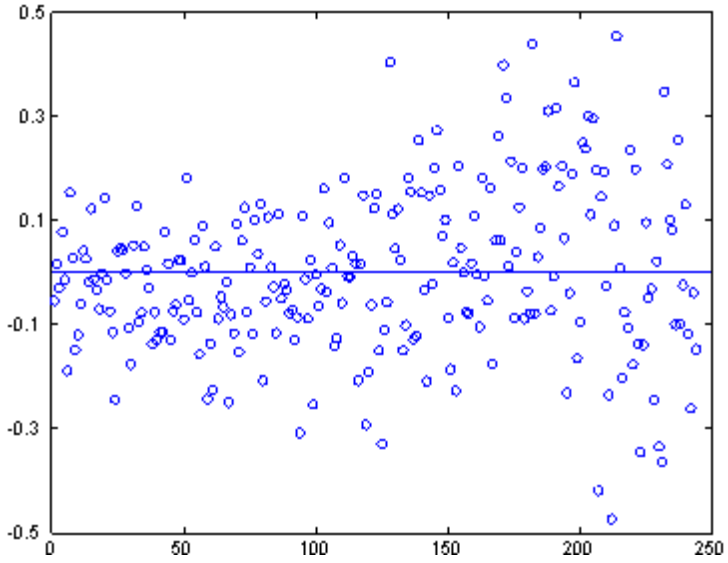


FIGURA 5.7: Residuales del modelo *score1* contra altura significativa.

cuando aumenta la altura significativa, como se observa en la figura 5.9, además su distribución se aproxima a la normal de acuerdo a la figura 5.10.

5.3 Hs con diferencias de la función seno

De las gráficas observadas correspondientes a datos de olas, se observa que básicamente las olas reflejan una aparente forma sinusoidal, por lo que puede pensarse en el perfil sinusoidal como una estructura a partir de la cual las olas se modifican, dado que las olas registradas en tormentas raramente siguen exactamente formas sinusoidales.

Para analizar como difieren las olas promedio por periodo de una ola formada por senos y cosenos, se ajustó a cada ola promedio, mediante el método de bases de funciones, una curva formada por una combinación lineal de las funciones mostradas en la figura 5.11. Como el ajuste se hace mediante mínimos cuadrados, la ola ajustada representa la mejor aproximación a la ola promedio obtenida con las funciones base que modelarían adecuadamente el comportamiento de una ola sinusoidal.

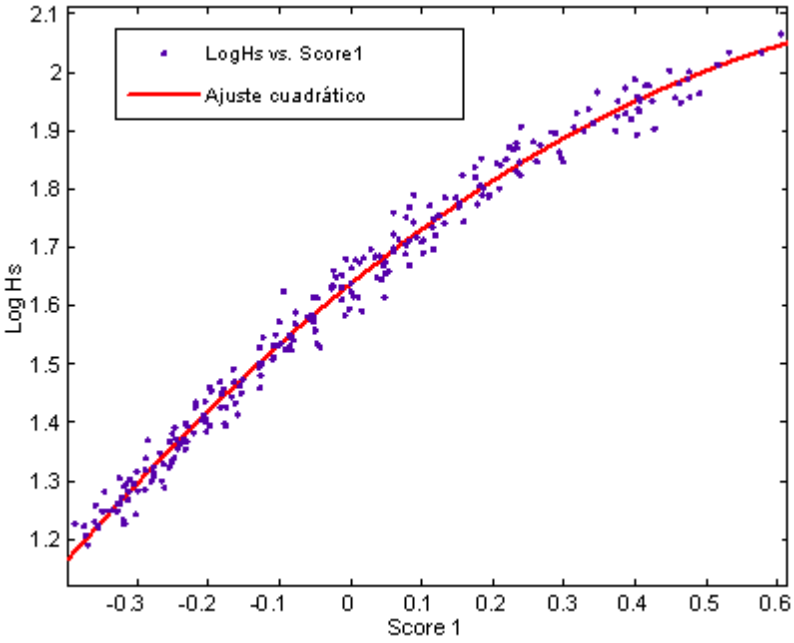


FIGURA 5.8: *Score* de olas promedio contra logarítmo de altura significativa. Modelo (5.3).

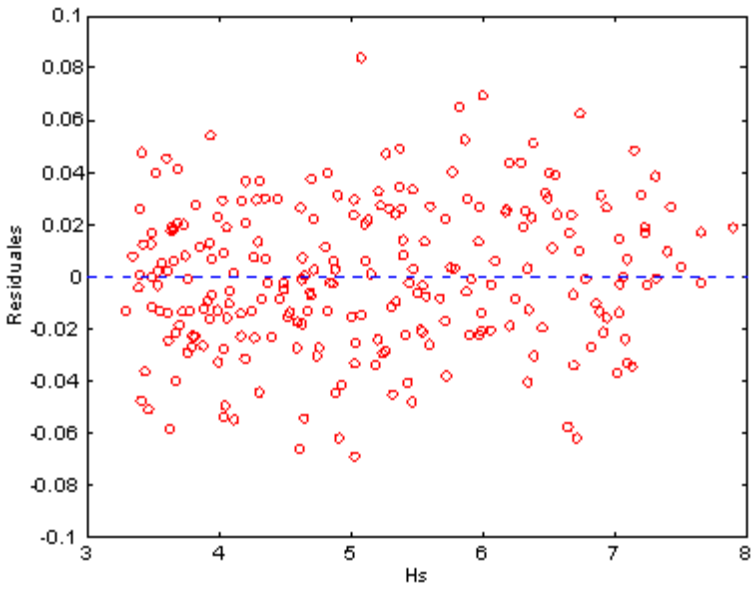


FIGURA 5.9: Altura significativa contra residuales correspondientes al modelo (5.3).

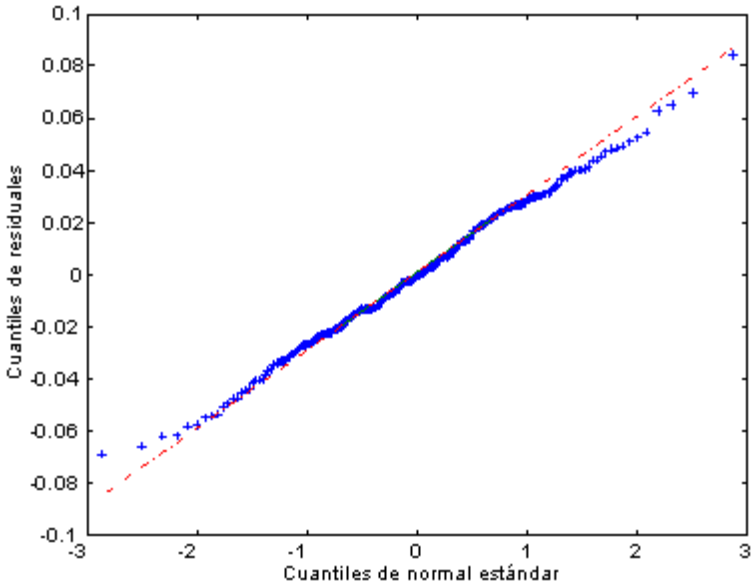


FIGURA 5.10: Gráfica cuantil-cuantil para los residuales del modelo (5.3).

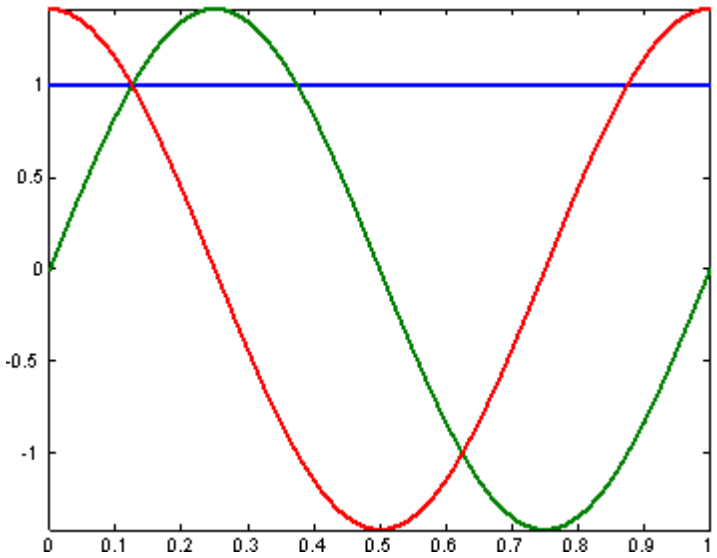


FIGURA 5.11: Primeras tres funciones del sistema de bases de funciones Fourier: 1, $sen\,(\omega t)$, $\cos(\omega t)$.

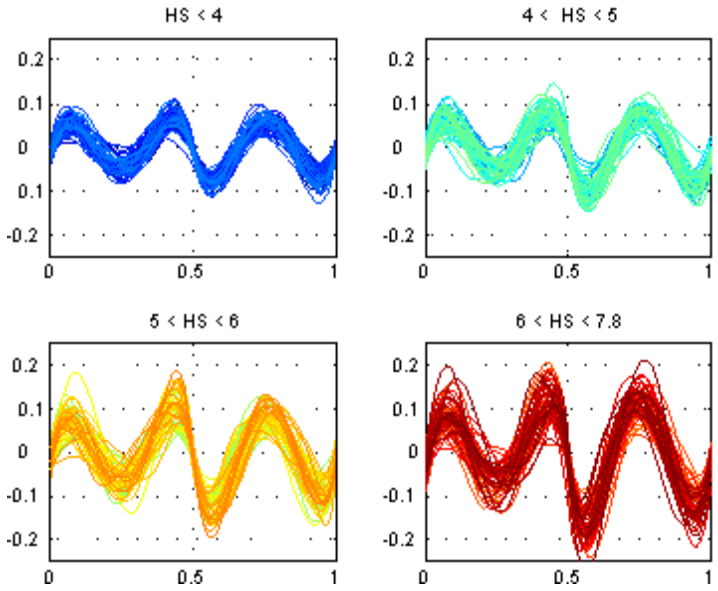


FIGURA 5.12: Diferencias entre la aproximación mediante senos y cosenos a la ola promedio por periodo y la ola promedio.

Las curvas resultado del ajuste se compararon con las curvas promedio observadas. Para hacer esta comparación se determinaron las funciones diferencias que se muestran en la figura 5.12 y las derivadas de estas funciones diferencia, figura 5.13.

Las funciones diferencia muestran que las curvas registradas y las observadas difieren principalmente en las pendientes con la que la ola comienza y termina, las pendientes antes y después del punto de cruce 0.5, en la cresta y en el valle. Las pendientes en las curvas promedio registradas son más pronunciadas y las curvas ajustadas sobreestiman un poco la cresta y valle; otro aspecto a destacar en la gráfica de la figura 5.12 son los cambios de las curvas diferencias, ya que al aumentar la altura significativa se observa que las diferencias en general son más grandes, con especial énfasis después del punto 0.5. Lo anterior refleja que es más difícil modelar mediante funciones básicas (seno y coseno), en periodos de mayor altura significativa

Las derivadas de las funciones diferencias, presentan una distinción más clara ante cambios de altura significativa como se ve en la figura 5.13. Los componentes principales de las derivadas manifiestan que el modo de variabilidad principal se

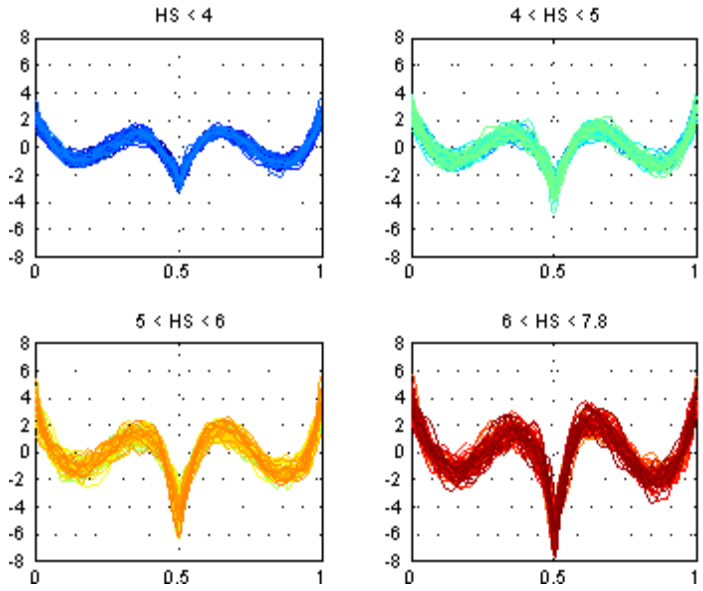


FIGURA 5.13: Derivadas de funciones diferencia 5.12

presenta en el punto 0.5, figura 5.15. La figura 5.14 da una mejor idea de la variabilidad dominante; al examinar esta figura, podemos decir que hay mayor variabilidad en las derivadas en el intervalo $[0.5, 1]$, que en el $[0, 0.5)$. Curvas a las que les corresponda un alto *score* indicarán que presentan un pico muy bajo en 0.5, un máximo muy alto en la región $[0.5, 0.7]$ y un mínimo muy bajo en $[0.75, 0.95]$, en la figura 5.12 se observa que las derivadas que muestran este tipo de comportamiento están asociadas a periodos con altura significativa alta, por lo que es posible que la variación prevalente en las derivadas se deba en su mayoría a los cambios en los periodos de altura significativa, y que el *score* asociado a la primera eigenfunción este relacionado directamente con la altura significativa de cada periodo. Para verificar lo anterior se ajustó un modelo entre este *score* y la altura significativa, inicialmente se propuso un modelo lineal, sin embargo se obtuvo un mejor ajuste en términos de aproximación de normalidad a los residuales al ajustar el modelo cuadrático (figura 5.16),

$$\log Hs = 1.63 - 0.6\text{Scr} - 0.24\,(\text{Scr})^2\,. \tag{5.4}$$

El análisis del modelo (5.4) es el siguiente:

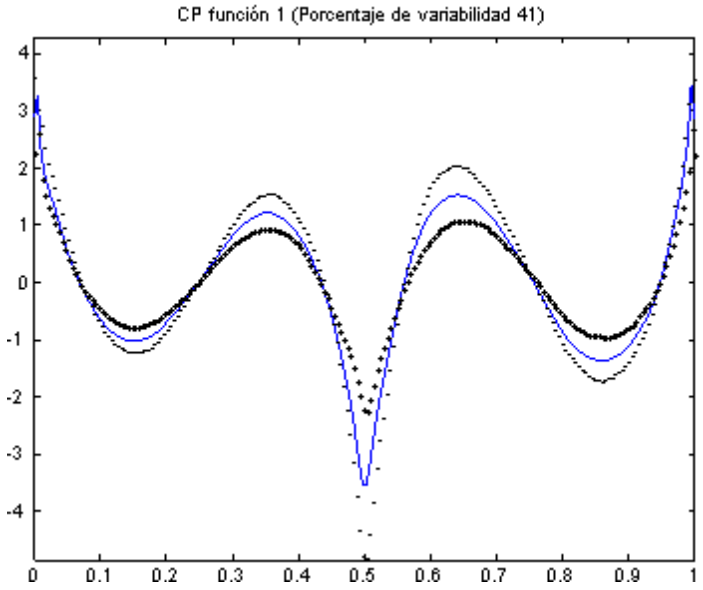


FIGURA 5.14: Media de las derivadas de las funciones diferencia y curvas resultado de sumar y restar un multiplo de la primera eigenfunción.

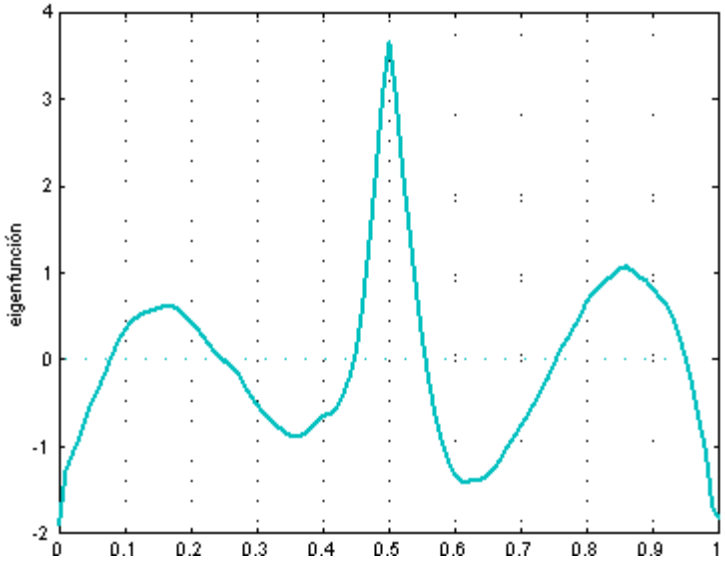


FIGURA 5.15: Primera eigenfunción de las derivadas de las diferencias.

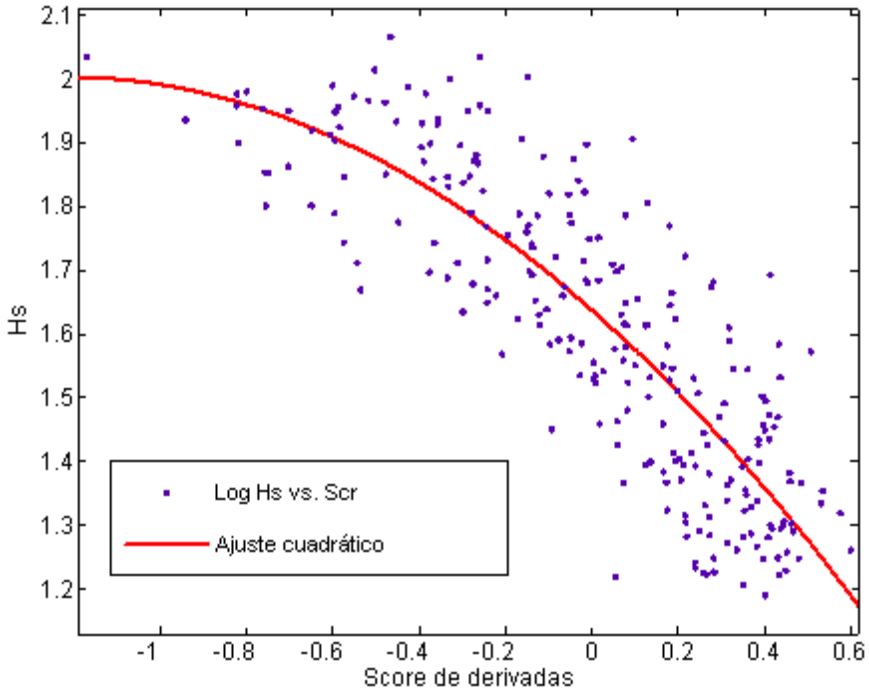


FIGURA 5.16: Regresión ajustada correspondiente a (5.4).

Termino	**Estimación**	**Error std.**	**Razón t**	**Prob >|t|**
Intercepto	1.639	0.010689	153.24	< .0001
Scr	-0.598	0.0259	-23.13	< .0001
$(\text{Scr})^2$	-0.245	0.00554	-4.43	< .0001

	df	**Suma de cuadrados**	**Cuadrados medios**	**Razón F**
Modelo	2	9.43	4.71	298.45
Error	241	3.81	0.0158	Prob>F
C.Tot	243	13.24		<.0001

Bajo este modelo se logra un ajuste razonable con un coeficiente $R^2 = 0.71$, aunque la gráfica de residuales contra altura significativa (figura 5.17) presenta, al igual que en el ajuste de un modelo lineal, una tendencia creciente en los residuales. En cuanto al comportamiento normal de los residuales se observa, en la figura 5.18, que es

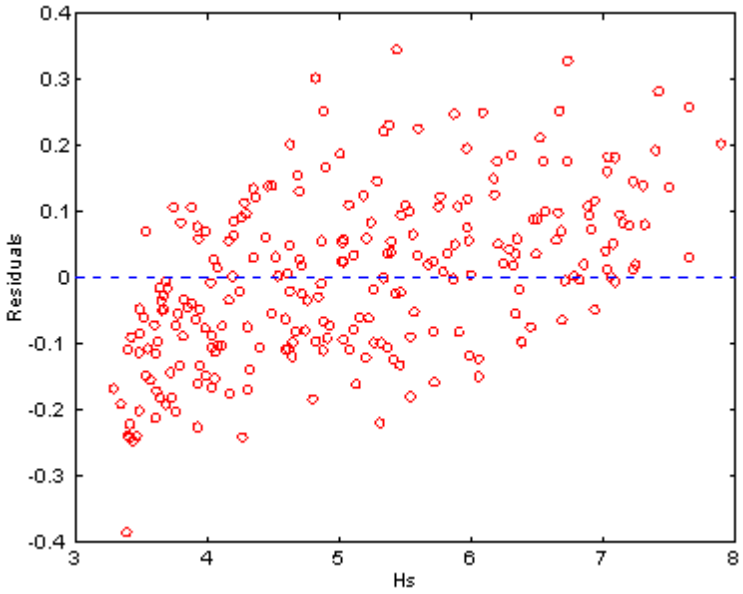


FIGURA 5.17: Gráfica de residuales del modelo (5.4).

aceptable. Debido al comportamiento observado en los residuales, es recomendable explorar otros modelos que expliquen la tendencia observada, lo cuál queda por hacer como trabajo a futuro.

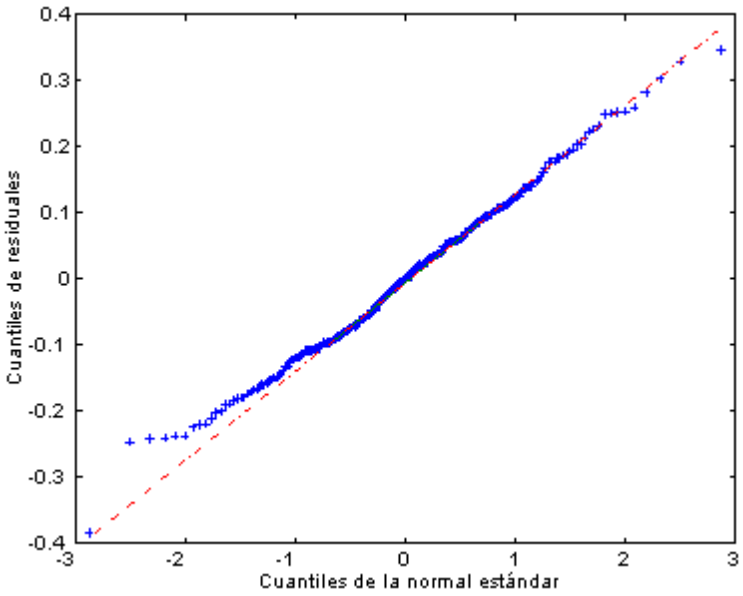


FIGURA 5.18: Gráfica quantil-quantil de los residuales del modelo (5.4).

Capítulo 6

Análisis del periodo de mayor altura significativa

Hasta ahora hemos explorado las variaciones y las características generales de las olas promedio asociadas a un periodo. No obstante, los perfiles de las olas que conforman un periodo pudieran aportar también información interesante. Uno de los periodos de mayor interés es el de mayor altura significativa, debido a que corresponde al nivel de tormenta más fuerte. Para los datos proporcionados, la mayor altura significativa estimada en uno de los periodos de 20 minutos fue de $7.89m$; las olas que integran este periodo se presentan en la figura 6.1. En este capítulo se realiza un análisis de estos perfiles, sus formas de variación, y se da una clasificación empírica de estas olas con base en las pendientes que las definen.

6.1 Componentes principales

Utilizando las olas que integran el periodo de mayor altura significativa como datos, se realizó un análisis de componentes principales. La variabilidad explicada por cada eigenfunción se muestra en la tabla siguiente, con los dos primeros componentes se logra explicar el 87% de la variabilidad presente en los datos y con los primeros cinco hasta el 97%, por lo que se considerarán cinco eigenfunciones en el análisis.

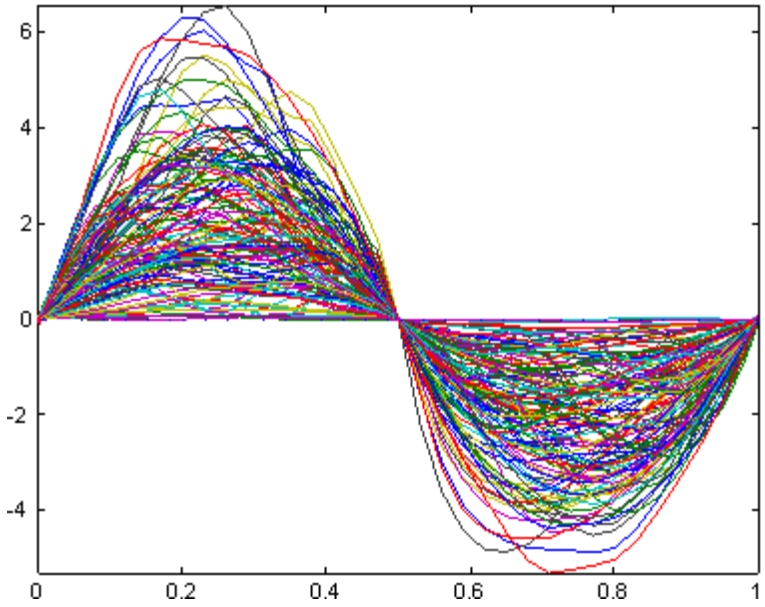


FIGURA 6.1: Olas que integran el periodo de 20 min. con altura significativa 7.89.

	CP 1	CP 2	CP 3	CP 4	CP 5	Total
Porcentaje	74%	13%	5%	3%	2%	97%

La figura 6.2 muestra la ola promedio del periodo y el efecto sobre de ella de las primeras cinco eigenfunciones resultado de la rotación VARIMAX. Con base en esta gráfica, podemos asociar a cada eigenfunción características particulares en la forma de la ola, por ejemplo la primera eigenfunción se relaciona principalmente con la altura de la cresta y la tercera con la pendiente con la cual se cruza el cero. De manera que olas a las que corresponda un *score* alto de la primera eigenfunción serán olas muy altas y a las que les corresponda un *score* alto de la tercera eigenfunción tendrán pendientes muy pronunciadas antes del cruce con el cero.

Dada la situación anterior, es posible establecer un orden entre las olas que integran el periodo, definido por la característica en la forma de la ola que resalta la eigenfunción correspondiente. La manera de cuantificar que tan importante es en la ola $x_i(t)$, la característica asociada a la función de peso $\xi_k(t)$ o eigenfunción, es

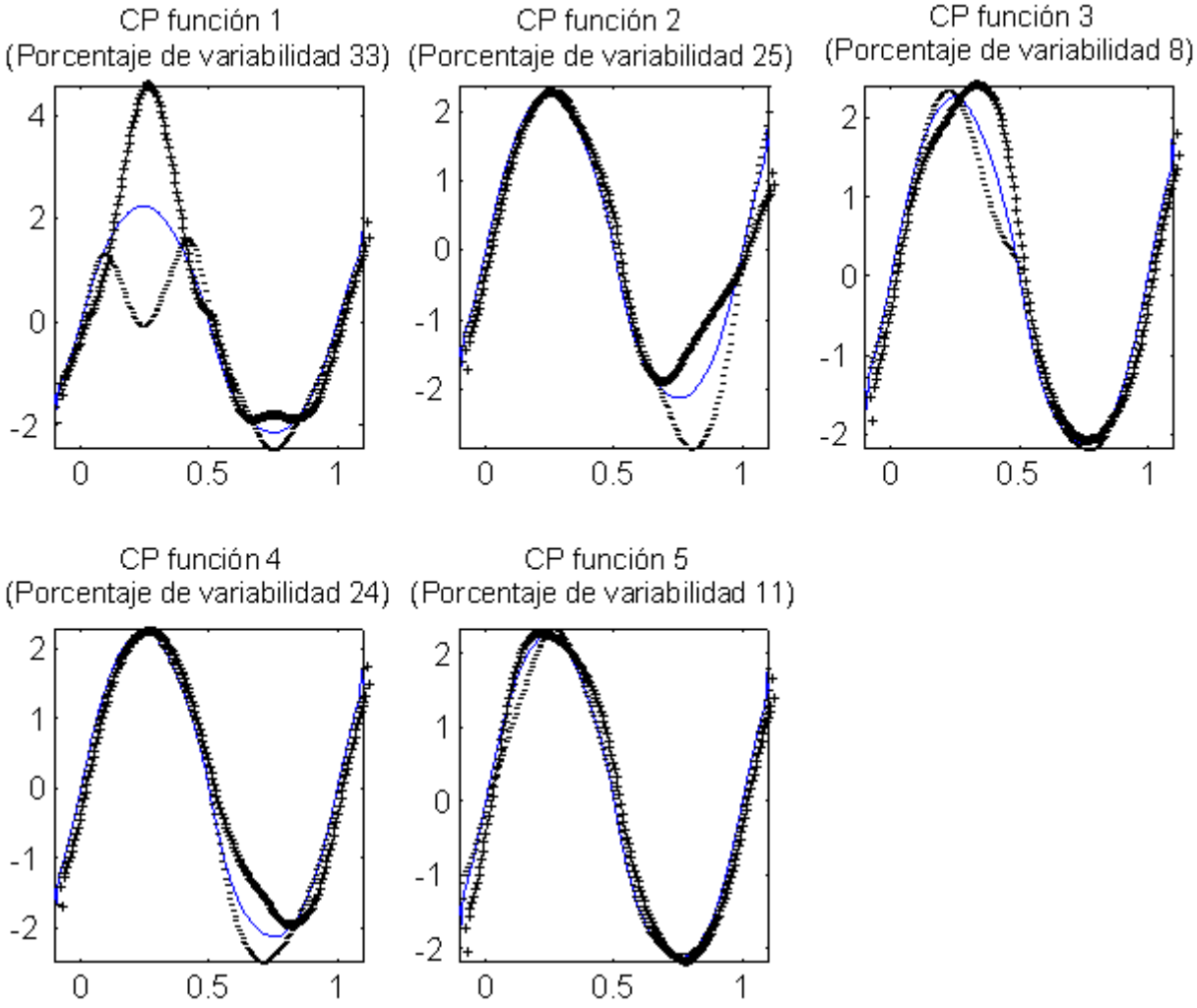


FIGURA 6.2: Ola promedio del periodo y curvas obtenidas al sumar y restar un múltiplo de la correspondiente eigenfunción rotada con VARIMAX.

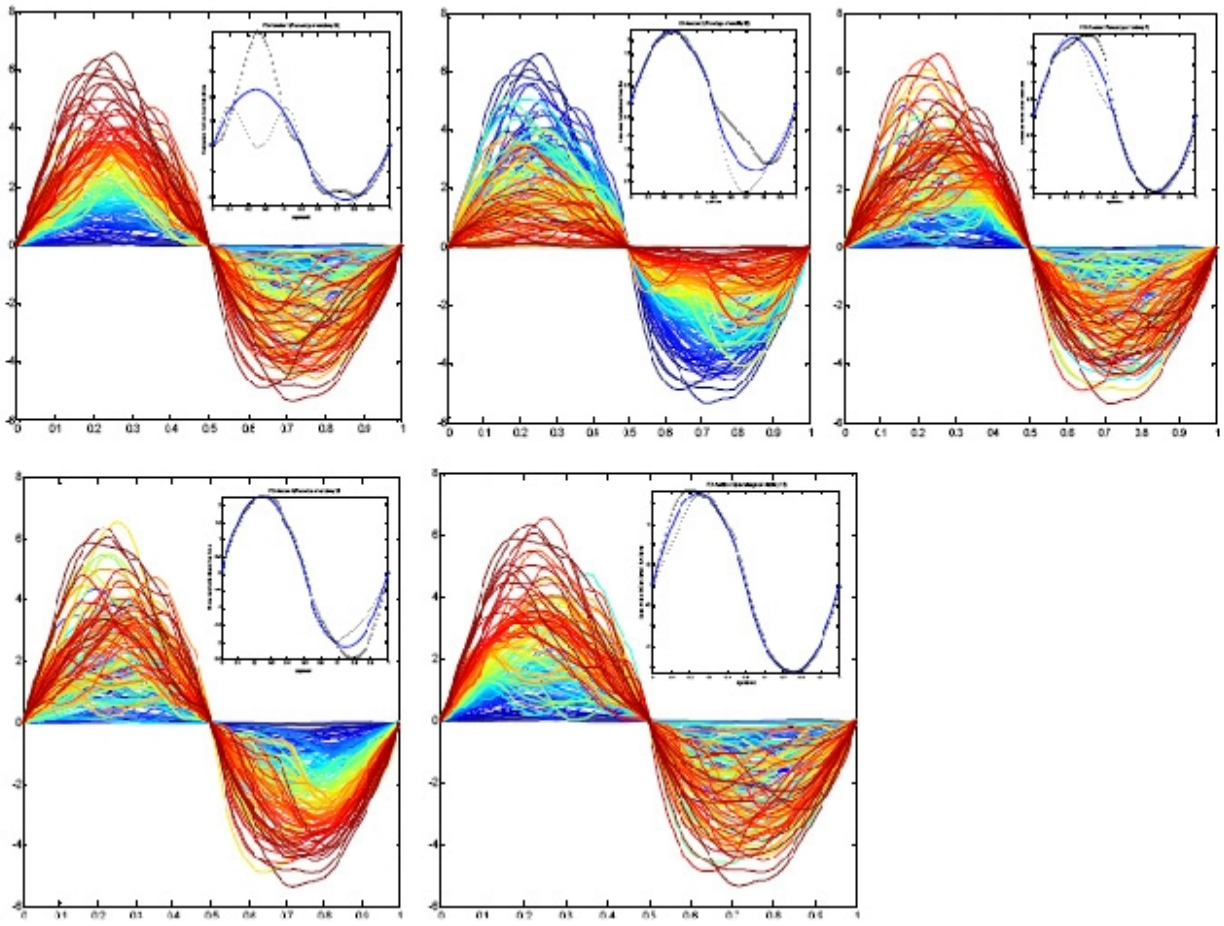


FIGURA 6.3: Olas ordenadas de acuerdo a la magnitud del *score* asociado a cada eigenfunción.

através de los *scores*

$$\int \xi_k x_i(t)\, dt.$$

Las olas ordenadas de acuerdo al *score* correspondiente a cada eigenfunción, se muestran en la figura 6.3, en este caso la escala de colores está asociada a las magnitud del *score* obtenido. Para cada componente se observa una transición uniforme en los colores en alguna parte de la ola. A grandes rasgos, se observan diferencias en la cresta, en la pendiente con que inicia el valle, en la pendiente de caida con que cruza el nivel cero, en la pendiente con que terminan la olas y en la pendiente con la que inician las olas; siendo las variaciones en la cresta y las pendientes con las que se inicia el valle, los modos más importantes de variación, como se observó en la tabla anterior.

6.2 Aproximación de olas observadas mediante eigenfunciones

Uno de los aspectos importantes del cálculo de componentes principales es que proveen de un conjunto de funciones ortonormales tales que la expansión de la curva en términos de estas funciones la aproximen tan bien como sea posible; esto nos permite aproximar cualquier ola del periodo a partir de las eigenfunciones rotadas. Por ejemplo para aproximar la ola $(x_i(t))$ que se muestra con linea punteada en la figura 6.4, se utiliza la expresión $\widehat{x}_i = \overline{x}_i(t) + \sum_{k=1}^{5} s_{ik}\xi_k(t)$, donde s_{ik}, es el score y $\xi_k(t)$ la eigenfunción k. Se observa en la figura como evoluciona la aproximación al ir aumentando términos en la suma anterior, de hecho cada aproximación modifica la característica asociada a la eigenfunción que se agrega; así en la primera aproximación se observa que la modificación de la ola promedio se dá básicamente en la altura de la ola, la segunda aproximación ajusta la pendiente con que se inicia el valle, la tercera la pendiente antes del cruce con el cero y así sucesivamente hasta que se logra una aproximación razonable.

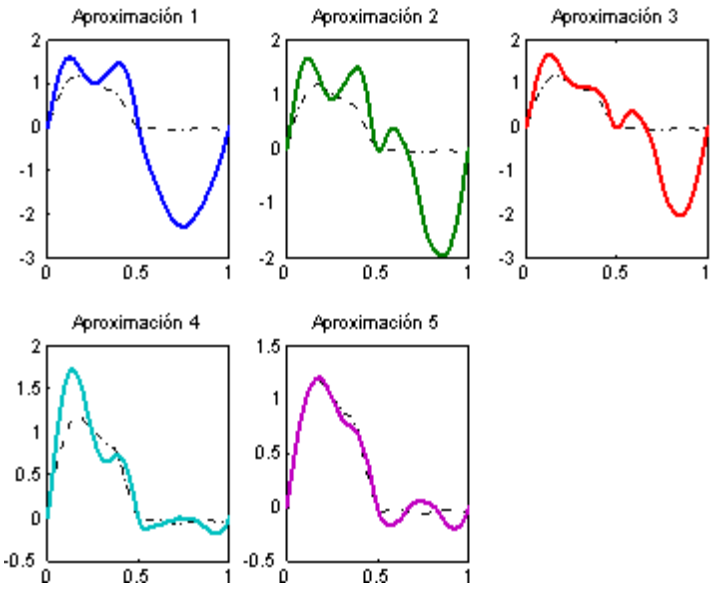


FIGURA 6.4: Aproximación de una ola mediante componentes principales.

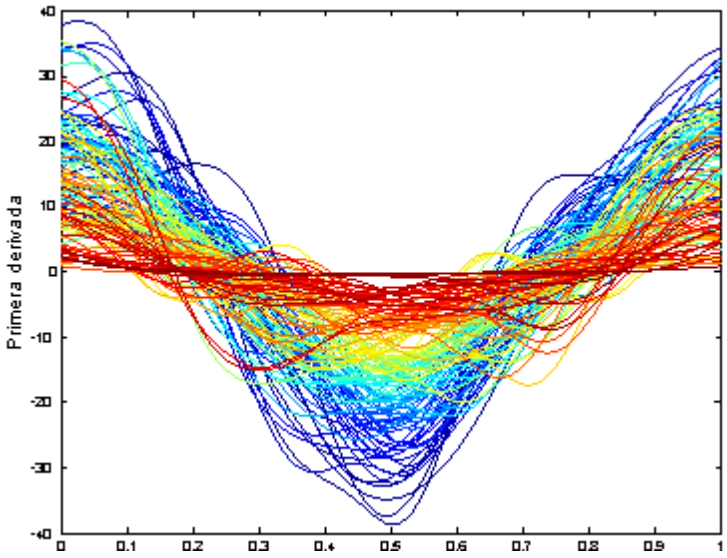


FIGURA 6.5: Derivadas del periodo con mayor altura significativa. La escala de colores está asociada a la magnitud del primer score.

6.3 Clasificación de las olas del periodo

Se considerarán ahora las derivadas de las olas en el periodo, las cuales nos dan información sobre las pendientes en cada punto. Como se vió en el capítulo 3 una pendiente muy pronunciada al caer es una de las características de *freak waves*, por lo que buscar este tipo de comportamiento en un conjunto de olas es de interés para una posible identificación de *freak waves.*

Las derivadas de las olas del periodo se muestran en la figura 6.5, para estas curvas se calcularon tres eigenfuciones obteniendo con ellas hasta el 92% de la variabilidad total. A las funciones resultantes se les aplicó la rotación VARIMAX con el objetivo de lograr mayor interpretabilidad en los resultados, la figura 6.6, muestra la ola promedio y las curvas resultado de sumar y restar un múltiplo de la eigenfunción correspondiente. Se observa que el primer componente está asociado a la variabilidad en las pendientes centrales, entre ellas la pendiente con la que la ola cae o cruza el nivel cero; el segundo componente está asociado a las pendientes con las que la ola inicia y el tercero a las pendientes con que termina.

Debido a que el primer componente representa una característica de interés, podemos utilizar el *score* de la primera eigenfunción para identificar las olas en el periodo que destaquen en esta característica. La escala de colores mostrada en la figura 6.5 corresponde a la magnitud de este *score*, las curvas azules están asociadas a niveles altos y las rojas a niveles bajos. En esta figura se observa que efectivamente, curvas con un *score* muy bajo tienen pendientes más pronunciadas alrededor del punto 0.5; este comportamiento observado directamente en las olas del periodo, se muestra en la figura 6.7, y en este caso la magnitud del *score* se refleja en el orden que produce en las pendientes centrales de las olas. La figura 6.8 muestra las cuatro primeras olas de este orden, que coinciden con las olas de mayor pendiente en la caída antes del cruce con el cero.

Empleando los tres *scores,* correspondientes a las eigenfunciones analizadas anteriormente, como una observación en tres variables, podemos determinar mediante clusters una clasificación de las olas que forman el periodo. El método utilizado para el cálculo de los clusters es de tipo jerárquico, considera una matriz de distancias entre observaciones definida por la métrica euclidiana y el método de Ward para evaluar la distancia entre conglomerados. El resultado de esta clasificación se

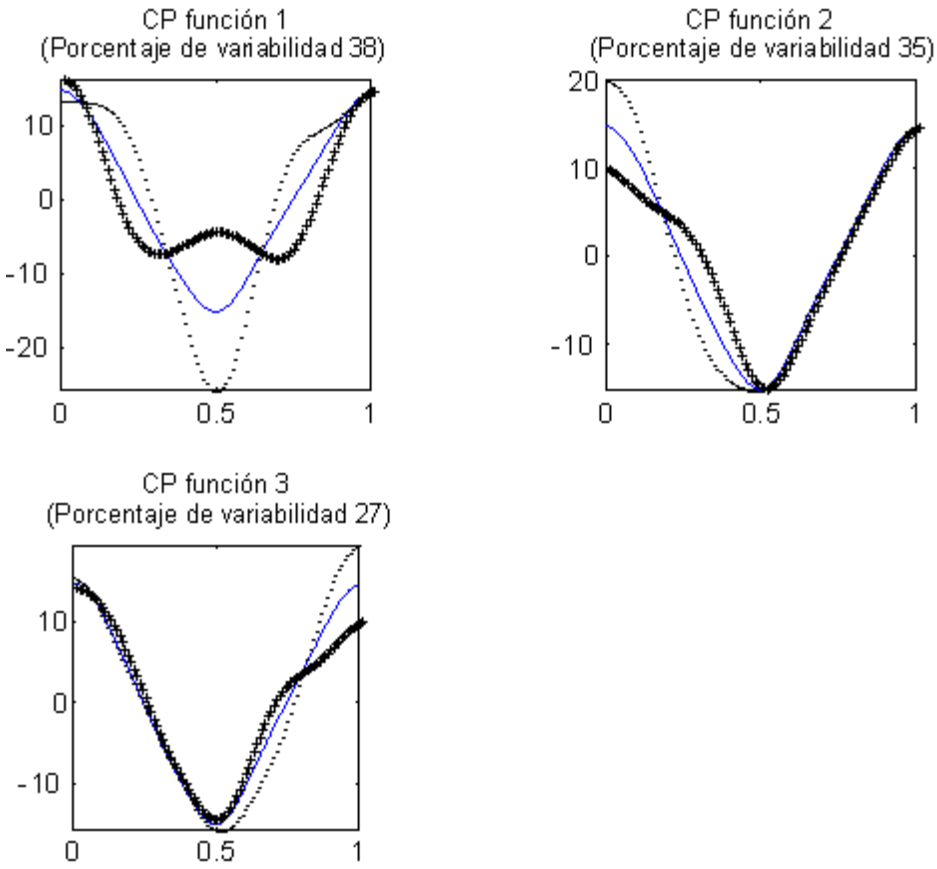


FIGURA 6.6: Derivada promedio y curvas obtenidas al sumar y restar un múltiplo de la correspondiente eigenfunción.

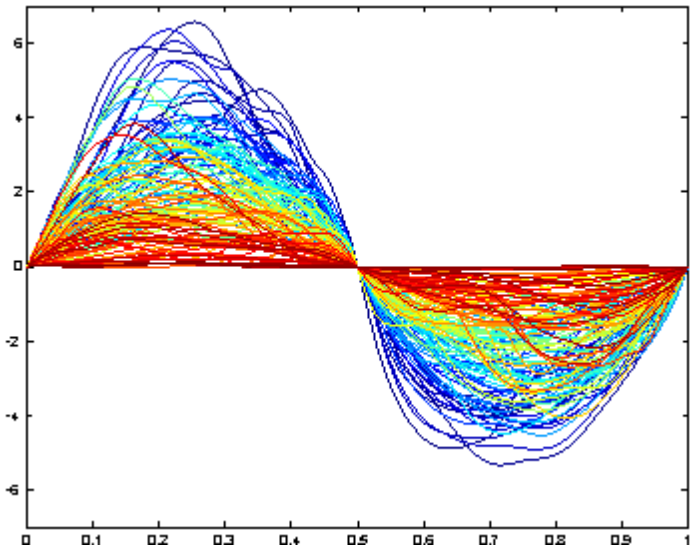


FIGURA 6.7: Olas ordenadas de acuerdo al primer *score* de las derivadas.

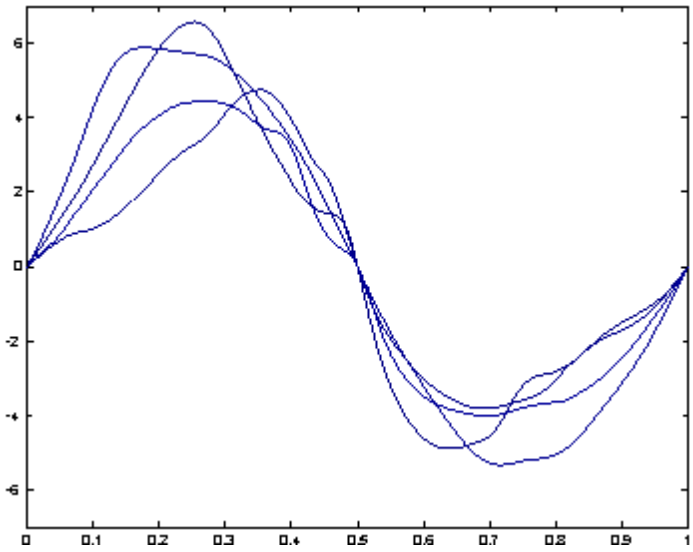


FIGURA 6.8: Olas identificadas con mayor pendiente antes de cruzar el nivel cero.

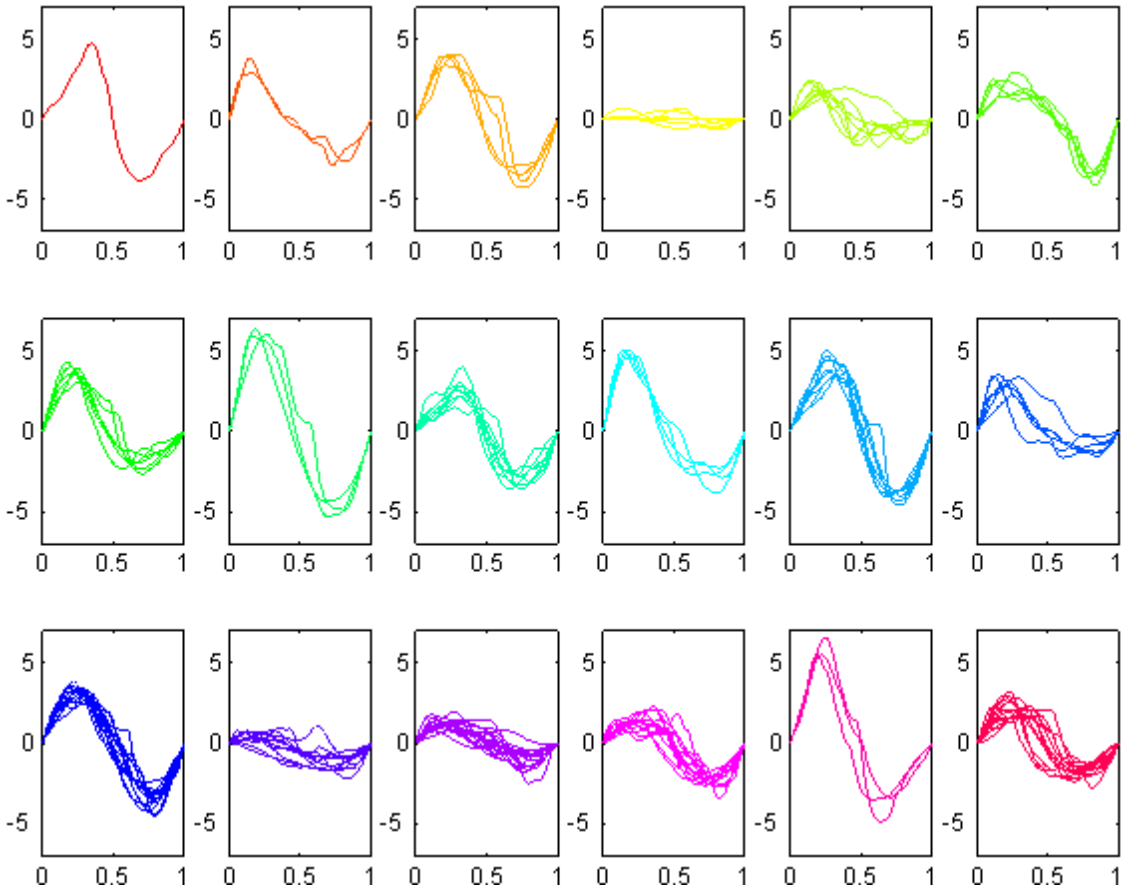


FIGURA 6.9: Clasificación de las olas obtenida a partir de los *scores* de las derivadas.

muestra en la figura 6.9.

Capítulo 7

Comparación de los datos registrados con datos simulados de un proceso gaussiano.

Un supuesto común, sobre el proceso $\eta(t)$ que representa la altura respecto a la media del nivel del mar, es el de gaussianidad. Dada la covarianza del proceso, estimable a partir de la densidad espectral, es posible generar olas que simulen un proceso gaussiano con la función de autocovarianza estimada. Siguiendo este procedimiento se simularon observaciones de $\eta(t)$ de un proceso gaussiano, correspondientes a mediciones con la misma frecuencia de muestreo que los datos (5Hz), por periodos de 1hr.

Para comparar los resultados de la simulación con las olas observadas se calculó la diferencia entre las olas promedio simuladas y las observadas por periodos de 1hr, estas diferencias se separaron por intervalos de altura significativa. En la figura 7.1 se presentan las diferencias asociadas a periodos con altura significativa menor a 4, entre 4 y 5; 5 a 6 y mayores a 6, respectivamente. Se observa una tendencia que indica un aumento en las diferencias, principalmente en la cresta y después del punto 0.5, al aumentar la altura significativa. Lo que sugiere que las olas observadas difieren de las gaussianas básicamente en estas regiones.

Para analizar la variación de las diferencias se determinaron componentes principales, con el primer componente se obtuvo 76% de la variabilidad. En la figura 7.2

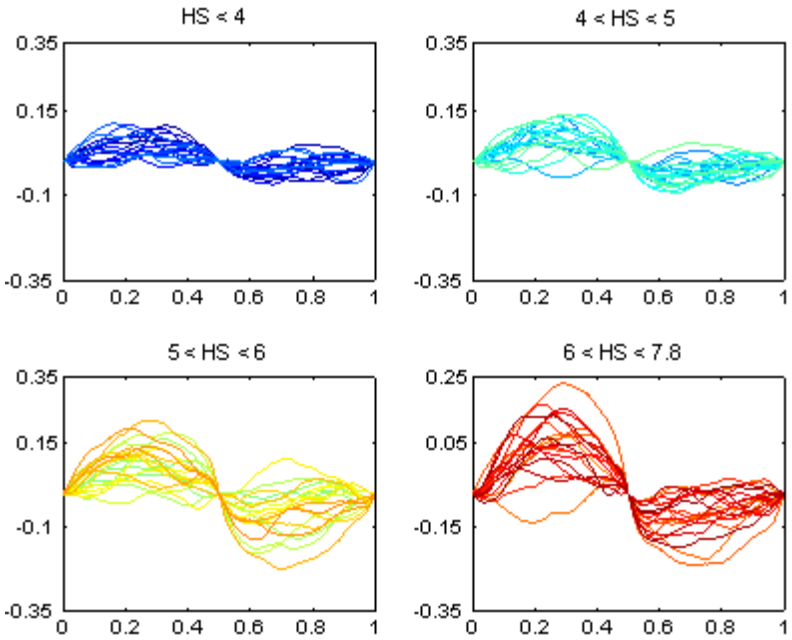


FIGURA 7.1: Diferencias: (Ola promedio − Ola promedio Gaussiana).

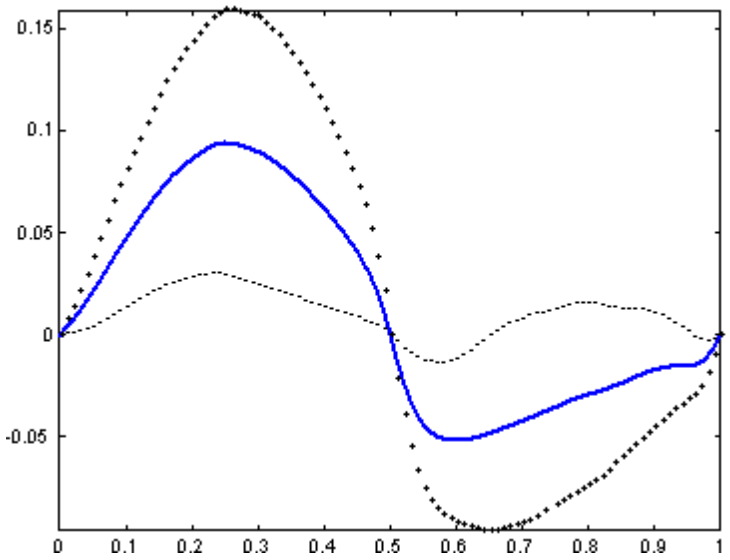


FIGURA 7.2: Promedio de diferencias mas/menos un múltiplo del primer componente principal.

se muestra el promedio de las diferencias y el promedio de las diferencias resultado de sumar y restar un múltiplo de la primera eigenfunción. La ola promedio indica que la diferencia principal se da en la altura de la ola. En general, las diferencias son mayores en el intervalo $[0, 0.5]$ que en el $[0.5, 1]$, en este último intervalo destaca la región $[0.55, 0.65]$ y hacia el final de la ola la diferencia se aproxima de manera creciente a cero. En cuanto al principal modo de variación éste se presenta en la alturas y en el valle en la parte $[0.55, 0.95]$.

Con el propósito de realizar una exploración mayor entre las diferencias de las olas simuladas y las observadas, se calculó la distancia entre ellas mediante la siguiente fórmula,

$$d\left(O_g, O_r\right) = \left(\int_0^1 \left|O_g\left(t\right) - O_r\left(t\right)\right|^2 dt\right)^{\frac{1}{2}}.$$

Graficando estas distancias contra la altura significativa del periodo correspondiente (figura 7.3), se identifica una tendencia creciente, lo cual confirma el comportamiento observado en las figuras que muestran la diferencias entre las dos olas. En la figura 7.3 también se observa que aumenta la variabilidad de las distancias al incrementar la altura significativa, para estabilizar estas variaciones se aplicó la trasformación

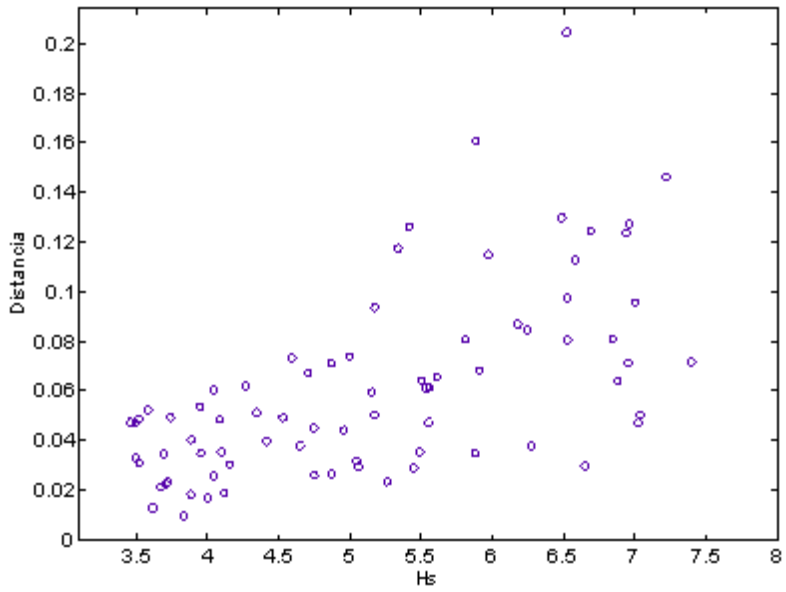


FIGURA 7.3: Altura significativa contra distancia entre olas promedio del proceso simulado y el observado.

logaritmo a las distancias. Como resultado se obtiene la gráfica de la figura 7.4, en la cual se continua observando la tendencia creciente. Un análisis de correlación entre el logaritmo de las distancias y la altura significativa indica que estas cantidades estan correlacionadas significativamente de manera positiva.

Variable	**Media**	**Error std.**	**Correlación**	**Prob** $>\|t\|$
HS	5.151922	1.177408	0.629859	<.00001
Log(Distancia entre olas)	-2.97632	0.616762		

Para modelar esta correlación se ajustó un modelo lineal que busca explicar la distancia entre las olas, es decir que tan grande es la diferencia con las olas de un procesos gaussiano, en términos de la altura significativa del periodo. El resultado del ajuste es el siguiente:

$$Log(\text{Distancia entre olas}) = -4.67 + 0.33\text{Hs} + \varepsilon \tag{7.1}$$

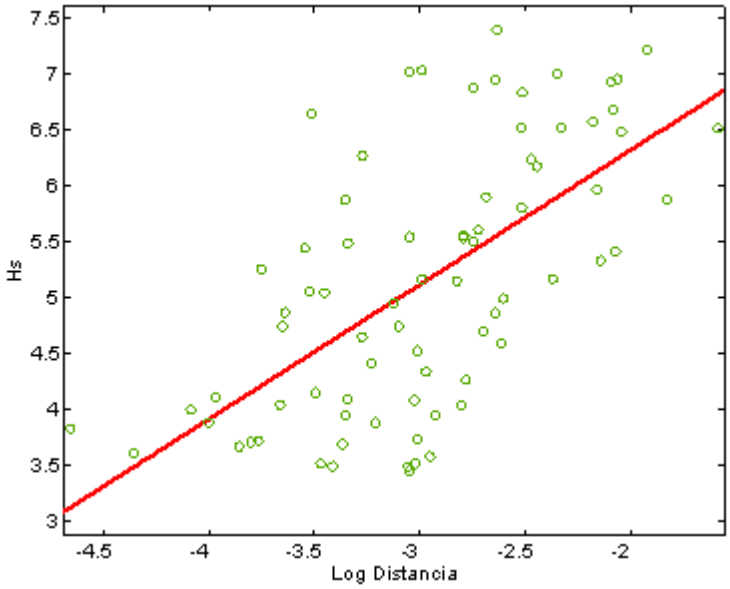


FIGURA 7.4: Distancia entre olas promedio del proceso simulado y el observado contra altura significativa.

Termino	**Estimación**	**Error std.**	**Razon t**	**Prob >|t|**
Intercepto	-4.67	0.2482	-18.84	<.0001
Scr 1	0.33	0.046981	7.02	< .0001

	df	**Suma de cuadrados**	**Cuadrados medios**	**Razon F**
Modelo	1	11.469	11.469	49.323
Error	242	17.440	0.232	Prob>F
C.Tot	243	28.9		<.0001

$$R^2 = 0.39.$$

El comportamiento de los residuales se presenta en la figura 7.5. En términos de la R^2 no se obtiene un buen ajuste por lo que el modelo no es útil para predecir, sin embargo resulta significativo, lo que refuerza el supuesto de que el nivel de altura significativa si tiene influencia en la diferencia entre las olas simuladas y

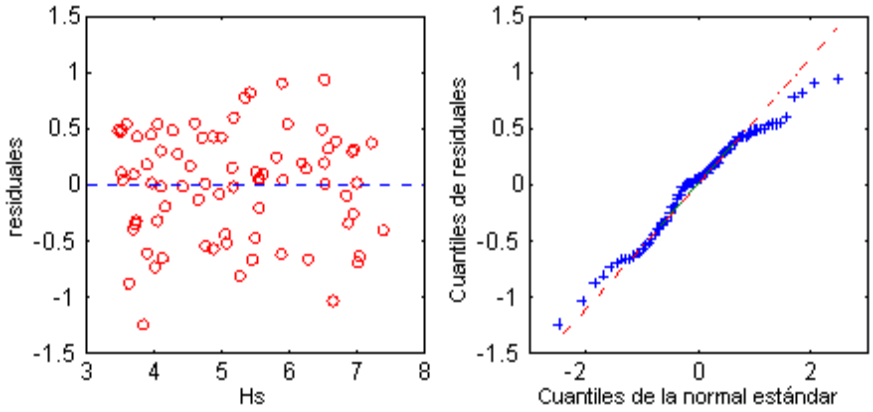


FIGURA 7.5: Comportamiento de residuales del modelo (7.1).

las observadas, con mayor altura significativa mayor es la diferencia de un proceso gaussiano y esta diferencia se dá principalmente en la cresta y en una pequeña región después del valle, como se vió antes.

Para análizar con mayor detalle como cambia la forma de las funciones en la figura 7.1, ante los cambios de altura significativa, se calcularon sus derivadas, las cuales se muestran en la figura 7.6, en esta figura se observa que el aumento en la altura significativa repercute principalmente en la pendiente que corresponde al punto 0.5. Es decir en las funciones diferencia, correspondientes a periodos de altura significativa alta, se presenta una diferencia grande entre las olas antes del punto 0.5, lo que ocasiona que la pendiente en este punto sea muy negativa.

El primer componente principal asociado a la derivadas de las funciones diferencia (figuras 7.7 y 7.8), refleja que el modo de variación dominante se presenta en el punto 0.5, dado que los cambios de altura significativa afectan principalmente este punto en las derivadas, podría pensarse que hay una asociación entre el *score* del primer componente con la altura significativa. La figura 7.9 presenta este *score* contra la altura significativa, se observa que mientras más decrece el *score* mayor es la altura significativa. Este comportamiento concuerda con lo observado, ya que un *score* muy negativo esta asociado a picos pronunciados en 0.5 en las derivadas.

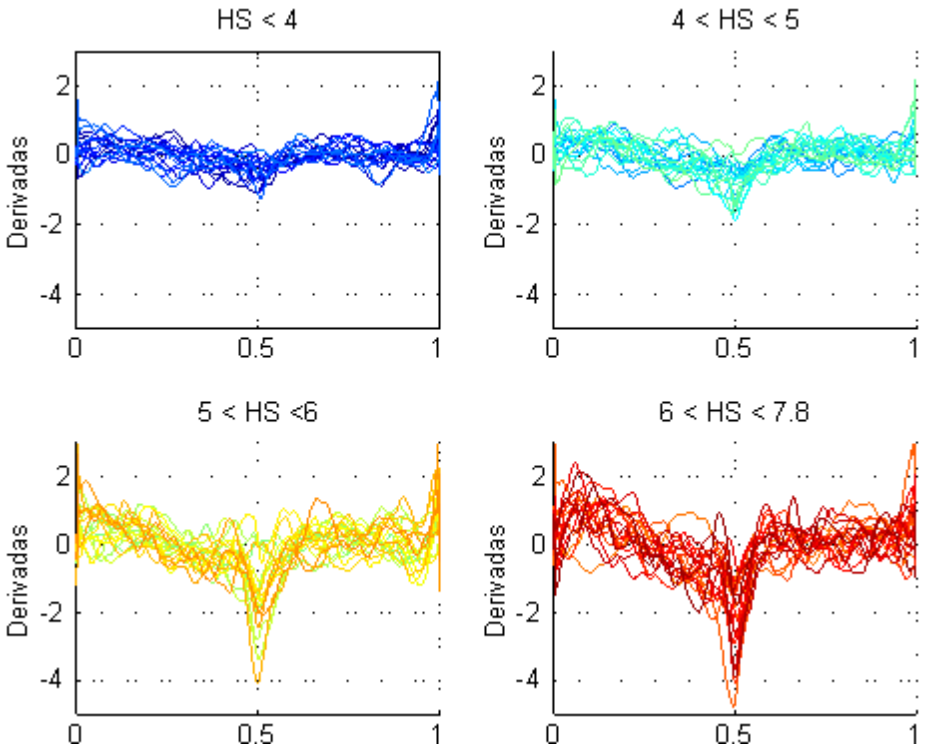


FIGURA 7.6: Derivadas de las funciones diferencia de mostradas en la figura 7.1.

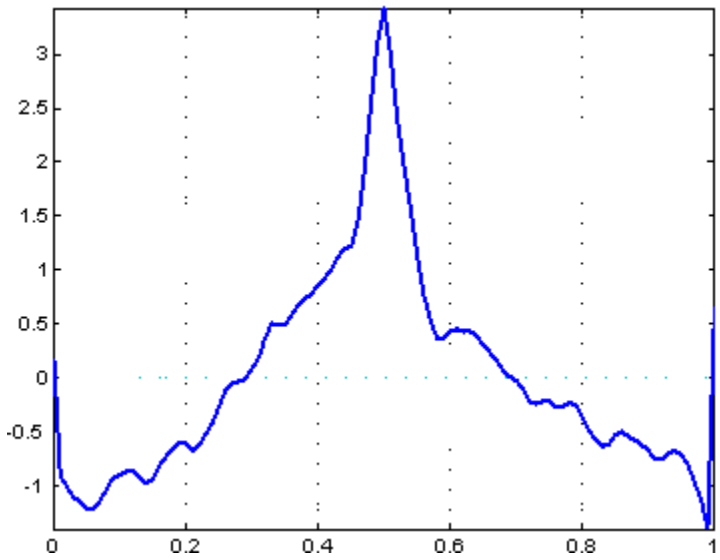


FIGURA 7.7: Primera eigenfunción de las derivadas de las funciones diferencia.

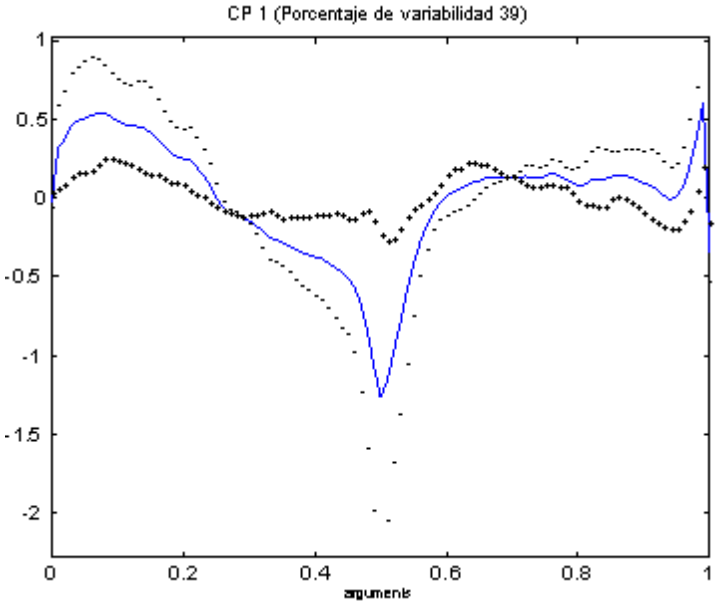


FIGURA 7.8: Promedio de derivadas de funciones diferencia y funciones resultado de sumar y restar un múltiplo de la primera eigenfunción.

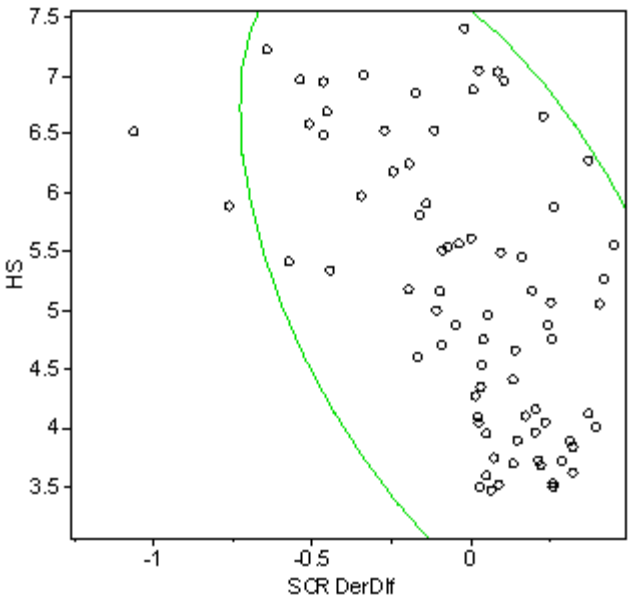


FIGURA 7.9: *Score* de la derivadas (figura 7.6) contra altura significativa.

Se presenta enseguida el análisis de correlación entre estas dos cantidades.

Variable	**Media**	**Error std.**	**Correlación**	**Prob** $>\|t\|$
Score de derivadas	2.208e-7	0.297377	-0.54651	<.00001
Hs	5.151922	1.177408		

Capítulo 8

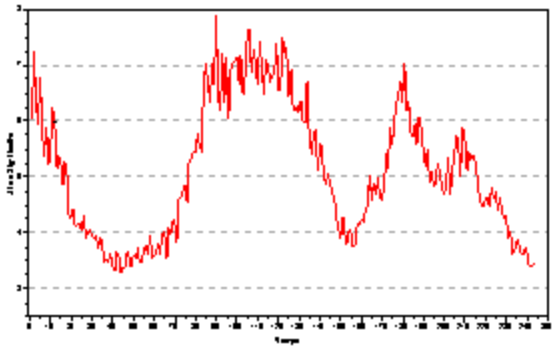
Conclusiones

La teoría de análisis funcional de datos empleada en la exploración de la forma de las olas y su relación con altura significativa, resultó una herramienta cómoda para este propósito. Al considerar los perfiles de las olas como funciones se obtuvieron una gran variedad de herramientas en el análisis de forma, variabilidad, y en la determinación de relaciones con otras variables de interés como la altura significativa.

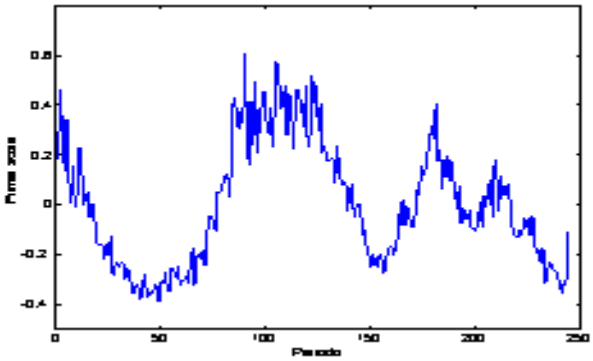
De manera general se vio que la forma de la ola promedio por periodo definida por su perfil y sus derivadas varía gradualmente ante diferentes valores de altura significativa, los periodos con altura significativa grande corresponden a olas más altas y profundas con pendientes más pronunciadas y con mayor energía. La diferencia principal en las pendientes de las olas que corresponden a distintos niveles de altura significativa se presenta en la región comprendida entre la cresta y el valle.

Mediante un modelo funcional se asoció la forma de la ola promedio de un periodo con su altura significativa, la función de regresión obtenida sugiere que la región del valle y la pendiente con que la ola promedio cruza el cero son regiones importantes para la aproximación de la altura significativa.

Se determino una relación muy clara entre la altura significativa y el primer *score* de las olas promedio asociado a la altura, como puede verse en las siguientes figuras.



Altura significativa de los datos proporcionados.



Score asociado al primer componente principal.

Al comparar la ola promedio observada con la ola regular (formada por $sen(\omega t)$, $cos(\omega t)$) mejor ajustada, se encontró que la diferencia principal entre las olas observadas y las regulares se presenta en las pendientes, es decir, las pendientes de las olas regulares son menos pronunciadas y la diferencia más importante se presenta en las pendientes con las que la ola cruza el nivel cero.

En la exploración del periodo con altura significativa mayor, se obtuvo una clasificación de las olas que integran este periodo a través de los *scores* obtenidos por componentes principales, a cada *score* fue posible asociarle una característica particular de la forma de la ola. En particular, por medio de un score pudo representarse

la severidad de la pendiente en una ola antes del cruce con el cero, esta pendiente es una característica importante en la identificación de *freak waves.* En la clasificación obtenida pudieron identificarse grupos pequeños de olas que destacan algunas de las olas más peligrosas del periodo y que probablemente correspondan a *freak waves.*

El análisis realizado de las diferencias entre las olas observadas y las provenientes de un proceso Gaussiano mostró que las regiones en la ola donde estas diferencias se enfatizan son la altura y la región en la que inicia el valle de la ola. Además estas diferencias son mayores al aumentar la altura significativa. A grandes rasgos se observó que al aumentar la altura significativa las diferencias entre las olas gaussianas y las observadas son más variables. Sin embargo aún queda por hacer una exploración más profunda sobre las formas de variación de las olas gaussianas y cómo difieren de las variaciones que presentan las olas observadas, mediante una comparación del análisis de componentes principales.

Gran parte de las observaciones y resultados presentados son generalmente conocidos, sin embargo desde el enfoque funcional fue posible deducirlos de una manera muy natural, además de que no se requerieron una gran cantidad de supuestos.

Un aspecto importante en el análisis de los datos de olas desde el punto de vista funcional, fue la alineación de las olas que integran cada uno de los periodos, respecto a esto queda por explorar otras formas de alineación así como analizar la información de las funciones que definen la alineación.

Un aspecto que resaltó en los ajustes de modelos que involucraban altura significativa, fue la variabilidad asociada a la aproximación de los modelos, en general se vió que al aumentar la altura significativa aumentaba esta variabilidad, por lo que queda como trabajo a futuro una exploración de modelos que se ajusten de una manera más adecuada a esta variación.

Bibliografía

[1] Ramsay J.O.,Dalzell C.J. *Some Tools for Functional Data Analysis*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 53, No. 3. (1991), pp. 539-572.

[2] Ramsay, J. O., Silverman, B. W. (2005). *Functional Data Analysis,* second edition. Springer Series in Statistics, New York.

[3] Ramsay, J. O., Silverman, B. W. (2002). *Applied Functional Data Analysis. Methods and Case Estudies*. Springer Series in Statistics.

[4] Aage C., Allan T., Carter D. J. T., Lindgren G y Olagnon M. *Oceans from Space. A textbook for Offshore Engineers and Naval Architects*. Ifremer.

[5] Ortega J. *Estudio de algunas propiedades del mar usando modelos aleatorios*. Boletín de la Asociación Matemática Venezolana Vol VIII No.2 (2001), pp. 111-130.

[6] http://ego.psych.mcgill.ca/misc/fda

[7] ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns

[8] http://www.maths.lth.se/matstat/wafo/download