Universidad Central de Venezuela
Facultad de Ciencias
Postgrado en Matemática

MODELAJE ALEATORIO NO GAUSSIANO DEL MAR

Autor: José Benito Hernández Chaudary.
Tutores: Dr. José Rafael León.
Dr. Joaquín Ortega S.

Tesis Doctoral
Presentada ante la ilustre
Universidad Central de Venezuela
Para optar al título de
Doctor en Ciencias
Mención Matemática.

Caracas, 04/10/2012.

Resumen

La idea de este trabajo es realizar un análisis no lineal de las alturas de olas, en los casos de presencia de tormenta y huracanes. Sabemos que el mar en condiciones *normales* se comporta de forma Gaussiana, es decir, su distribución de probabilidad es una Gaussiana, sin embargo en condiciones extremas (tormentas y huracanes) tal distribución se pierde. Previo a este análisis, estudiaremos algunos algoritmos de segmentación, para determinar la evolución de los espectros de energía de olas marinas, para varios conjuntos de datos, que incluyen mares en condiciones normales, tormenta y huracanes.

Palabras Claves:

Análisis Espectral, Olas Aleatorias, Detección de Cambios por Contrastes Penalizado, SLEX, Períodos Estacionarios, Series Temporales, Procedimientos de Segmentación, HHT, Espectros marginales, Segmentación, Gaussianidad, Prueba de Hipótesis.

Dedicada con todo cariño y amor a dos personas muy especiales que gracias a su dedicación, esfuerzo y entrega hicieron posible que hoy día yo haya logrado realizar esta Tesis Doctoral, mis padres

José Benito e **Hilda**.

Primeramente agradezco a Dios por la salud, vida y sabiduría dados para poder realizar esta Tesis Doctoral. Gracias a mis padres José Benito e Hilda por todo el apoyo y ánimo brindado durante la realización de esta tesis. Gracias a mis tutores Dr. José Rafael León y Dr. Joaquín Ortega, por darme la oportunidad de desarrollar este trabajo, su profesionalismo, sapiencia, amistad y guía han sido de mucho valor para mi. También agradezco a todos los profesores de la Escuela de Matemática que me impartieron clases durante toda mi formación matemática, desde la licenciatura, pasando por la maestría y culminando en mis estudios doctorales, a todos ellos gracias. Gracias a mi hermano gemelo Roel por la ayuda prestada en el desarrollo de los algoritmos utilizados en esta Tesis, y también por escucharme cada vez que le contaba los intrincados enredos matemáticos nada fáciles de entender para quienes no están dedicados a las Matemática. Gracias a mi hermana Yurubí por estar pendiente siempre de mí y por ayudarme en la traducción al inglés de los artículos que resultaron de esta investigación para su publicación. También quiero agradecer la colaboración del prof. George H. Smith de la Universidad Heriot-Watt, Edinburgh, Escocia, Reino Unido quien nos proporcionó los datos de tormenta de la plataforma North Alwin, Mar del Norte. Este trabajo fue parcialmente realizado durante varias visitas al Centro de Investigaciones en Matemáticas, Guanajuato, México, el cual fue parcialmente financiado por la Universidad Central de Venezuela y el CIMAT. Agradecemos su apoyo. Este trabajo fue parcialmente financiado por el CONACYT México, subvención 52554. Igualmente agradecemos su apoyo. Finalmente, quiero agradecer a mi esposa Leydi por tener paciencia conmigo, durante la última etapa de realización de esta Tesis Doctoral.

Índice general

Las tres cuartas partes de la superficie del planeta tierra están cubiertas por los océanos, los cuales almacenan el 97.26 % del total de los recursos hídricos existentes. Estas masas de agua almacenada están sometidas al sistema general de circulación generado por la acción de los rayos solares, la rotación de la tierra y las ca-racterísticas físico-químicas del agua salada. La circulación general se manifiesta principalmente en forma de corrientes. También influyen sobre el comportamiento de las masas de agua las acciones locales que están reguladas por el relieve del fondo, la cercanía a los continentes y las condiciones meteorológicas. Entre estas acciones sobresalen los sismos, las mareas y los oleajes debidos al viento. Eventos naturales como los sismos inducen la formación de olas conocidas como tsunamis o maremotos, los cuales han producido efectos catastróficos en diversas zonas costeras del mundo. Las magnitudes de los oleajes están asociados con las tormentas que se originan por la velocidad y la dirección de los vientos. En las zonas de latitud media las características de los vientos son influenciadas por las fuerzas Centrípeta y de Coriolis y ocasionan la formación de ciclones o huracanes durante algunos meses del año. Por su parte las mareas dependen de la relación Sol-Luna-Tierra. Para que se genere una ola se requiere que exista una fuente de energía que, al transmitir al agua en reposo una cantidad determinada de energía, produce un movimiento oscilatorio de las partículas del líquido sin que haya un transporte importante de masa. Este movimiento oscilatorio es similar al que se induce por vibración a una cuerda que esté fija por sus dos extremos.

Por otra parte, los huracanes, ciclones tropicales o tifones representan eventos meteorológicamente poderosos los cuales son capaces de generar estados extremos en el mar. En las regiones tropicales y semitropicales por lo general ellos representan una limitante en el diseño de estructuras costeras y de mar adentro, así como para la navegación. A pesar de las condiciones meteorológicas extremas asociadas con los huracanes, ellos ocurren relativamente poco y por lo general con un patrón de movimiento impredecible, por lo que obtener una base de datos suficientemente grande de mediciones ha sido un gran reto. Como resultado de programas de boyas, observaciones satelitales y patrones numéricos usando modelos espectrales, sobre un periodo de más de 30 años, existen bases de datos detalladas hoy día. Mejorar los pronósticos de olas generadas por huracanes o tormentas es un paso esencial para minimizar el daño causado por las tormentas tropicales en los asentamientos costeros y actividades económicas en las zonas cercanas a las costas. La Hidráulica Marítima tiene como objetivo el análisis y la cuantificación de los fenómenos que se producen en las aguas marítimas que tienen influencia sobre proyectos específicos de navegación, construcción de puertos, facilidades turísticas o protección de playas y zonas costeras, de allí la importancia de poder conseguir o construir modelos de las olas del mar.

La idea fundamental es modelar la superficie del mar como una *superficie aleatoria* que evoluciona en el tiempo, es decir, como un proceso aleatorio X que depende de la posición en el espacio $\mathbf{x}$, del tiempo t y, por

supuesto, de un parámetro aleatorio ω que pertenece a un espacio de probabilidad $(\Omega, \mathfrak{F}, P) : X(t, \mathbf{x}, \omega)$.

Una primera aproximación para construir modelos de olas aleatorias para registros de gran longitud es suponer que ellas son procesos estacionarios a trozos, es decir, que existen instantes donde el *estado* de las olas cambia, pero entre estos puntos de cambios las olas son modeladas por un proceso estacionario. Una ventaja de esta aproximación es que se puede usar el análisis espectral clásico en cada intervalo estacionario con la interpretación usual del espectro como una distribución de energía en el rango de frecuencias. Para implementar esta aproximación es necesario tener maneras de detectar los cambios de estado en el proceso, y por consiguiente en las características espectrales de la estructura de covarianza de un proceso estacionario. Por lo tanto, es razonable estudiar métodos basados en la detección de cambios en los espectros. También podemos valernos de técnicas de series temporales y utilizar métodos que busquen cambios en las alturas significativas o alguna otra característica espectral.

En este trabajo estudiaremos cuatro métodos para detectar estos cambios. Dos de estos métodos están basados en la detección de cambios en el espectro, ellos son: Detection of Changes by Penalized Contrasts (DCPC) Lavielle (1998 [17], 1999 [18]), Lavielle y Ludeña (2000 [19]), y Smooth Localized complex EXponentials (SLEX) Ombao et. al (2002 [22]). Ambos métodos han sido implementados por sus autores en MATLAB y han sido exitosamente usados en otras áreas, particularmente para el análisis de electroencefalogramas. Los otros dos métodos están basados en cambios en las características espectrales y son: el algoritmo de Soukissian & Samalekos (2005)[32] que es un método de segmentación para alturas significativas basado en la determinación de períodos de estabilidad, crecimiento y decrecimiento utilizando técnicas de series temporales. Este método consiste en hacer regresión lineal local donde los puntos iniciales y finales son los puntos extremos (máximos y mínimos locales) de la serie temporal y luego definiendo una función de costo para determinar la mejor configuración de los intervalos. Aplicaremos este procedimiento a varias características espectrales y compararemos los resultados con un método de segmentación diferente descrito a continuación. El otro método de segmentación está basado en valores medios sobre ventanas en movimiento, usando una banda con ancho fijo para determinar los puntos de cambio en el registro de olas. Los intervalos durante el cual los valores se mantienen dentro de las bandas alrededor de la media se considerarán estables mientras que aquellos que estén por encima (o debajo) se considerarán crecientes (o decrecientes). De esta manera determinaremos los intervalos de estabilidad, crecimiento y decrecimiento de los registros. Ambos métodos fueron implementados en MATLAB.

Ahora bien, las características de las olas durante una tormenta o un huracán cambian rápidamente y también son claramente no lineales y no gaussianos. Por lo que los modelos usuales, eminentemente gaussianos, para mares estacionarios no ajustan bien o hacen difícil la predicción del comportamiento de las olas

cuando hay presencia de tormentas o huracanes. Por estas razones se hace necesario buscar o idear modelos no-gaussianos y no-lineales para poder realizar un ajuste más satisfactorio de las olas. Un método que ha sido ideado para el estudio de ondas distorsionadas no lineales es la Transformada de Hilbert-Huang (HHT) Huang y Shen (2005, [13]), Huang et. al (2004, [14]).

Para realizar el análisis y comparación de estos algoritmos utilizaremos varios conjuntos de datos: Primero haremos una comparación de los algoritmos DCPC y SLEX, para ello consideraremos tres conjuntos de datos. Los dos primeros corresponden a olas en situación normal. Estos conjuntos de datos corresponden a 3 días en septiembre de 2005, iniciando el 1^{ro} de septiembre a las 0h, para dos boyas desplegadas por *Costal Data Information Program, Integrative Oceanography Division*, operado por el *Scripps Institution of Oceanography* (http://cdip.ucsd.edu/): Estación 067 Isla San Nicolás en las costas de California con una profundidad de 360m y la Estación 106 en la Bahía Waimea, Hawaii con una profundidad de 200m. En ambos casos la frecuencia de muestreo es 1.280Hz. Usaremos ambos métodos de segmentación para estos conjuntos de datos. Mientras que el tercer conjunto corresponde a los datos del huracán Camille, los cuáles tienen una situación de alta no-estacionaridad. Este conjunto de datos es bien conocido y ha sido considerado previamente por diversos autores (véase por ejemplo, Forristall (1978 [9]) y Guedes Soares et al. (2004 [11]). El huracán Camille ocurrió el 17 de agosto de 1969, y fue uno de los huracanes más fuertes que alcanzó la costa Este de los EE.UU en el siglo XX. Éste pasó a unos 23Km de la plataforma donde estaba un dispositivo de medición de alturas de olas. Este dispositivo de medición dejó de funcionar alrededor de las 4:30pm del 17 de agosto y la serie se inició a las 6:00pm del día previo (16 de agosto), con una frecuencia de muestreo de 1Hz. En todos los casos compararemos los intervalos de estacionaridad obtenidos con ambos métodos y también estudiaremos tres características espectrales: Energía total, valor máximo y frecuencia dominante. Trazaremos un gráfico y compararemos la evolución de las tres características espectrales para ambos métodos con los tres conjuntos de datos. Este análisis se hará en el capítulo 2.

Luego, realizaremos un estudio comparativo de los algoritmos de Soukissian y el de bandas. Consideraremos la evolución de los espectros de energía de olas durante un período de un año con datos de una estación situada en Waimea Bay, Hawaii, EEUU, con las siguientes características: Latitud 21°40.364' N, Longitud 158°06.949' W, profundidad del agua (m): 198.00. La serie temporal tiene una frecuencia de muestreo de 1.280Hz. Usando los registros de alturas de olas calcularemos los espectros de energía cada 15 minutos lo que nos permitirá determinar cambios en períodos pequeños de tiempo en la evolución de algunas características de olas obtenidas a partir de los espectros. Se utilizó el software WAFO desarrollado por Wafo group de la *Lund University of Technology*, Suecia,[1] para calcular los espectros y las características espectrales. El software

[1] El software WAFO está disponible en http://www.maths.lth.se/matstat/wafo.

WAFO es una herramienta de Matlab para el análisis estadístico y simulación de oleaje aleatorio. El análisis y comparación de estos dos métodos se realiza en el capítulo 3.

También analizaremos los datos de tormentas. Los datos fueron obtenidos de la plataforma North Alwin situada al norte del Mar del Norte, a unos 160Km al este de Shetland Islands ($60^o48{,}5'N, 1^o44{,}17'$E) con una profundidad del agua de aproximadamente 130 metros. Hay 3 medidores de altura de olas tipo Thorn EMI Infrarrojo montados en la plataforma y sus alturas están entre 25 y 35 metros sobre el nivel medio del agua. Los datos fueron registrados continua y simultáneamente a 5Hz y divididos en intervalos de 20 minutos para los cuales se calcularon los estadísticos altura significativa H_m, período de pico T_p y los momentos espectrales. Se mantuvieron todos los registros de elevación de la superficie con $H_m > 3m$. Los datos fueron transmitidos a la Universidad Heriot-Watt en una base diaria.

Analizamos un solo conjunto de datos. Este consiste en una serie de registros de 20 minutos de duración, muestreados por el altímetro a 5Hz y ocurridos entre la medianoche del 23 de diciembre y las 9:00am del 26 de diciembre de 1999, y consistente de 4840 minutos de registros. Se tiene que estos datos incluyen dos períodos relativamente largos en los que aumenta H_m y una sección en la cual ocurren dos picos en un tiempo relativamente corto. Dado que existen algunos intervalos cortos de tiempo en los cuales hay registros perdidos, dividimos este en 5 conjuntos que cubren la tormenta. Estos conjuntos los llamamos *Storm1999a, Storm1999b, Storm1999c, Storm1999d* y *Storm1999e*. En este caso, utilizaremos los algoritmos SLEX y HHT para realizar un análisis espectral de los mismos, el capítulo 4 trata este tema.

Finalmente, consideraremos los registros de olas de la Estación 144 en St. Petersburg, Florida, EE.UU., con las siguientes características: Latitud 27°20.42' N, Longitud 84°16.50' W, profundidad del agua: 94.0(m). La serie temporal tiene una frecuencia de muestreo de 1.280Hz. Tomamos dos conjuntos de datos, el primero inicia el 20 de enero de 2009 a las 2:11:00am y termina el 21 de enero de 2009 a las 2:10:59am, el segundo conjunto inicia el 01 de marzo de 2009 a la 01:11:00pm y termina el 02 de marzo de 2009 a las 11:10:59am. A estos conjuntos los llamamos *Estación 14401* y *Estación 14403*.

Los datos de la Estación 144 fueron obtenidos de la página web del *Coastal Data Information Program (CDIP), Integrative Oceanographic Division*, operado por el *Scripps Institution of Oceanography* bajo el patrocinio de la *U.S. Army Corps of Engineers* y el *California Department of Boating and Waterways* (`http://cdip.ucsd.edu/`).

En el capítulo 5, vamos a realizar pruebas de hipótesis no paramétricas de gaussianidad para todos los conjuntos de datos descritos anteriormente. El objetivo es determinar el comportamiento gaussiano o no de las olas en un período de tiempo determinado, y ajustar modelos no gaussianos y no-lineales a los registros de las mismas. Realizaremos el ajuste y las pruebas de hipótesis para los registros de las olas. También para validar el algoritmo de prueba de hipótesis y validez de los resultados realizaremos simulaciones gaussianas

con las funciones propias de los software Matlab y R, así como un algoritmo desarrollado basándonos en el procedimiento propuesto por Grigoriu (2009 [10]), para simulación de procesos gaussianos.

En los apéndices se darán los fundamentos teóricos para el desarrollo de las pruebas de hipótesis de gaussianidad. En el primer capítulo de esta tesis daremos los fundamentos y notaciones necesarios para entender el resto de los capítulos antes descritos.

Capítulo 1

Preliminares

1.1. Conceptos básicos y notaciones utilizadas

En esta sección estableceremos la notación a usar en todo este trabajo, así mismo daremos algunas definiciones y teoremas útiles para el estudio del mar. Empezaremos definiendo procesos estacionarios tanto en el caso real como complejo, ya que la representación espectral que definiremos posteriormente requiere de procesos complejos.

Definición 1.1.1 *Un proceso estocástico $X(t)$ es **estrictamente estacionario** si para cualquier n y cualesquiera instantes $t_1, \ldots, t_n$ en el dominio del proceso todas la distribuciones n-dimensionales de*

$$X(t_1 + h), \ldots, X(t_n + h)$$

*son independientes de h. Es llamado **débilmente estacionario** si tiene media constante, $\mathbb{E}[X(t)] = m$, y si su función de covarianza*

$$r(t) = Cov(X(s + t), X(s))$$

es una función del tiempo t.

Definición 1.1.2 *Un proceso a valores complejos*

$$X(t) = Y(t) + iZ(t)$$

*es **estrictamente estacionario** si todas las distribuciones $2n$-dimensionales de*

$$Y(t_1 + h), Z(t_1 + h), \ldots, Y(t_n + h), Z(t_n + h)$$

son independientes de h. Es llamado ***débilmente estacionario*** *o* ***estacionario de segundo orden*** *si* $\mathbb{E}[X(t)] = m$ *es constante y*

$$\mathbb{E}[X(s)\overline{X(t)}] = r(s-t) + |m|^2$$

solo depende de la diferencia de tiempo $s - t$.
La función de covarianza $r(h) = \mathbb{E}[(X(s+h)+m)\overline{(X(s)-m)}$ *es* ***Hermitiana****, es decir,*

$$r(h) = \overline{r(h)}$$

La siguiente definición será útil para ver que toda función de covarianza es positiva definida

Definición 1.1.3 *Una función a valores complejos* $r(t)$ *dada en* $(-\infty, \infty)$ *se dice que es* ***positiva definida*** *si para cada entero* $n \geq 1$, $t_1, \ldots, t_n \in \mathbb{R}$ *y complejos* $z_1, \ldots, z_n$ *se tiene que*

$$\sum_{j,k=1}^{n} r(t_j - t_k) z_j \bar{z}_k \geq 0 \tag{1.1}$$

En particular, se asume que la suma en (1.1) es un número real.

De la definición anterior, se tiene que una propiedad característica de una función de covarianza es que es **positiva definida** en el sentido siguiente: Sea $t_1, \ldots, t_n$ cualquier conjunto finito de tiempos, y tomamos números complejos arbitrarios $a_1, \ldots, a_n$. Por simplicidad suponemos $\mathbb{E}[X(t)] = 0$, entonces

$$\sum_{j,k}^{n} a_j \bar{a}_k r(t_j - t_k) = \mathbb{E}\left[\sum_{j,k}^{n} a_j X(t_j)\overline{a_k X(t_k)}\right] \tag{1.2}$$

$$= \mathbb{E}\left|\sum_{j=1}^{n} a_j X(t_j)\right|^2 \geq 0 \tag{1.3}$$

Teorema 1.1.1 *Cada función* $r(t)$ *positiva definida, posiblemente compleja, es la función de covarianza para un proceso estrictamente estacionario. Por lo tanto, la clase de funciones de covarianza es igual a la clase de funciones positiva definida.*

Como vimos, las funciones de covarianzas para procesos estacionarios están caracterizadas por la propiedad de ser positiva definida. De cursos elementales también sabemos que las funciones de covarianza son transformadas de Fourier de sus *distribuciones espectrales*. A continuación formularemos y demostraremos esta declaración, pero antes demostraremos el siguiente Lema:

Lema 1.1.1 *Sea* $r(t)$ *una función continua positiva definida. Entonces*

I. $r(0) \geq 0$

II. $\bar{r}(t) = r(-t), \quad |r(t)| \leq r(0)$ *y en particular,* $r(t)$ *es una función acotada.*

III. *Si* $\int_{-\infty}^{\infty} |r(t)|dt < \infty$, *entonces* $\int_{-\infty}^{\infty} r(t)dt \geq 0$

IV. *Para cada* $x \in \mathbb{R}$, *la función* $e^{itx}r(t)$, *como función de* t, *es positiva definida.*

Demostración. I Se sigue de la definición de función positiva definida para funciones a valores complejos con $n = 1, z = 1$

IV También es trivial de (1.1) si reemplazamos z_k con $z_k e^{it_k x}$

Para demostrar II, tomemos $n = 2, t_1 = t, t_2 = 0, z_1 = z, z_2 = \lambda$, donde λ es real. Entonces (1.1) se convierte en

$$r(0)(|z|^2 + \lambda^2) + \lambda r(t)z + \lambda r(-t)\bar{z} \geq 0 \tag{1.4}$$

se sigue inmediatamente que $r(t)z + r(-t)\bar{z}$ es real para cada complejo z. Más aún, dado que $r(-t)\bar{z} + \bar{r}(-t)z = 2\mathbf{Re}r(-t)\bar{z}$ es real, el número $(r(t) - \bar{r}(-t))z$ es real para cada complejo z, lo cual es solo posible cuando $r(t) - \bar{r}(-t) = 0$.

Por lo tanto, de (1.1) con $z = \bar{r}(t)$ obtenemos

$$r(0)|r(t)|^2 + r(0)\lambda^2 + 2\lambda|r(t)|^2 \geq 0$$

para todo real λ. Se sigue que $|r(t)|^4 - r^2(0)|r(t)|^2 \leq 0$. Por I se sabe que $r(0) \geq 0$, en el caso $r(0) = 0$ entonces se tiene la igualdad que $r(t) = 0$ para todo t lo cual verifica trivialmente lo requerido, en el caso $r(0) > 0$ se tiene que $|r(t)|^2 < r^2(0)$ de donde $r(t)$ es acotada para todo t, Esto demuestra la parte II.

Para la parte III, recordando que r es continua y su integral es el límite de las sumas de Riemann. Viendo dt y ds como z_j y $\bar{z}_k$ respectivamente, de (1.1) tenemos que

$$\int_{-N}^{N}\int_{-N}^{N} r(t-s)dtds \sim \sum_{i,j} r(t_i - t_j)\Delta t_i \Delta t_j \geq 0,$$

$$0 \leq \frac{1}{N}\int_{-N}^{N}\int_{-N}^{N} r(t-s)dtds = \int_{-\infty}^{\infty} r(t)(2 - \frac{|t|}{N})I_{|t|\leq 2N}dt$$

donde la igualdad se sigue cambiando variables $t - s = t', t + s = s'$. Por el Teorema de Convergencia Dominada de Lebesgue la última integral converge a $2\int_{-\infty}^{\infty} r(t)dt$. Esto demuestra la parte III y finaliza la demostración del lema. □

Teorema 1.1.2 *(Teorema de Bochner). Una función continua $r(t)$ es positiva definida, y por consiguiente una función de covarianza, si y sólo si, existe una función real no-decreciente, continua a la derecha y acotada $F(\omega)$ tal que*

$$r(t) = \int_{-\infty}^{\infty} e^{i\omega t} dF(\omega) \tag{1.5}$$

La función $F(\omega)$ es la función de distribución espectral del proceso, y tiene todas las propiedades de una función de distribución estadística excepto que $F(\infty) - F(-\infty) = r(0)$ no necesita ser igual a uno. La función $F(\omega)$ sólo se define hasta una constante aditiva, y usualmente se toma $F(-\infty) = 0$.

A continuacion incluimos algunos resultados y definiciones necesarios para la exposición de nuestros resultados.

Teorema 1.1.3 ***Lema de Fatou***. *Sea $\{f_n\}_{n=1}^{\infty}$ una sucesión de funciones medibles, $f_n \geq 0\ \forall n$. Entonces se tiene que*

$$\int (\liminf_{n\to\infty} f_n) dx \leq \liminf_{n\to\infty} \int f_n(x) dx.$$

Definición 1.1.4 a) *Sea $(\Omega, \mathfrak{F}, P)$ un espacio de probabilidad. Una transformación T medible sobre $(\Omega, \mathfrak{F}, P)$ se dice que preserva la medida si*

$$P(T^{-1}A) = P(A) \qquad \text{para todo} \quad A \in \mathfrak{F} \tag{1.6}$$

b) *Dado un espacio medible $(\Omega, \mathfrak{F})$ y una transformación T medible sobre $(\Omega, \mathfrak{F})$, una medida de probabilidad P se dice invariante si se cumple (1.6)*

Definición 1.1.5 *Sea $(\Omega, \mathfrak{F}, P)$ un espacio de probabilidad. Sea una transformación T que preserva la medida de Ω en sí misma.*

a) *Una variable aleatoria x sobre $(\Omega, \mathfrak{F}, P)$ se dice invariante bajo T si $x(\omega) = x(T\omega)$ para casi todo $\omega \in \Omega$*

b) *Un conjunto $A \in \mathfrak{F}$ se dice invariante bajo T si $T^{-1}A = A$*

Definición 1.1.6 *Una transformación T que preserva la medida sobre $(\Omega, \mathfrak{F}, P)$ es llamada **ergódica** si todo conjunto invariante $A \in \mathfrak{J}$ tiene $P(A) = 0$ ó $P(A) = 1$; es decir, todo conjunto invariante es trivial. Algunas veces se utiliza el término **métricamente transitivo** en lugar de **ergódico**.*

Para poder obtener un modelo manejable que nos permita estudiar el mar, es necesario hacer algunas suposiciones:

1. Supondremos que el proceso que sirve de modelo es estacionario. Sabemos que las condiciones del mar cambian con el tiempo, y con ellas los parámetros de las distribuciones estadísticas de la altura de olas. En este caso, y en general para cualquier proceso aleatorio, el término se refiere no a las olas sino a sus propiedades estadísticas. Esto quiere decir que la distribución de $X(t+h)$ es la misma para cualquier valor de h, y en particular, es siempre igual a la de $X(0)$. Más aún, para cualquier n y cualesquiera instantes de tiempo $t_1, t_2, ..., t_n$ la distribución del vector $X(t_1+h), X(t_2+h), ..., X(t_n+h)$ es independiente del valor de h.

2. El nivel medio del mar es 0 y mediremos las variaciones respecto a él. Esto quiere decir que el proceso que consideramos es centrado, esto es, $\mathbb{E}[X(t)] = 0$.

3. Las trayectorias del proceso X son continuas. Tenemos que X es en realidad una función sobre el espacio producto $[0,\infty) \times \Omega$ con la propiedad de que para cada $t \in [0,\infty)$ fijo, $X(t,\cdot)$ es medible. Si fijamos $\omega \in \Omega$ obtenemos una función

$$X(\cdot,\omega) : [0,\infty) \to \mathbb{R}$$

que se conoce como una trayectoria del proceso. Pedimos que para casi todo $\omega \in \Omega$ (es decir, con probabilidad 1) esta función sea continua.

4. El proceso es ergódico. El hecho de que el proceso sea ergódico, nos permite sustituir los valores esperados (teóricos) por promedios temporales (empíricos), teniéndose así que:

$$\begin{aligned} E(X(t)) &\equiv \int_\Omega X(t,\omega)dP(\omega) = \lim_{t\to\infty} \frac{1}{t}\int_0^t X(u)du. \\ Cov(X(t), X(t+h)) &\equiv \int_\Omega X(t,\omega)X(t+h,\omega)dP(\omega) \\ &= \lim_{t\to\infty} \frac{1}{t}\int_0^t X(u)X(u+h)du. \end{aligned}$$

5. El proceso que representa la altura de las olas es Gaussiano; es decir, la distribución de la altura de la ola en un punto dado y en un instante de tiempo t tiene la siguiente función de distribución:

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

donde σ^2 es la varianza de la distribución. Aún más, para cualquier valor de n y cualesquiera instantes de tiempo $t_1, t_2, ..., t_n$ la distribución del vector $X(t_1+h), X(t_2+h), ..., X(t_n+h)$ tiene densidad Gaussiana:

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

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

Lema 1.1.2 *Sea $X(t)$ un proceso Gaussiano estacionario y sea $r(t)$ su función de covarianza. Entonces la función de covarianza satisface*

$$r(s,t) = Cov(X(s), X(t)) = E[X(s)X(t)] = r(|s-t|).$$

En particular, si $s = t$

$$r(s,s) = Var(X(s)) = E[X^2(s)] = r(0).$$

La función de covarianza r es par y por lo tanto, si es diferenciable en 0, la derivada debe ser nula. Más aún, si r tiene dos derivadas en el origen, la segunda derivada debe ser negativa: $r''(0) < 0$.

1.1.1. La representación espectral

La función de covarianza de cualquier proceso estacionario es positiva definida, como se muestra en (1.2), (1.3) y el Teorema 1.1.1: Para cualquier n y cualesquiera $z_1, ..., z_n$,

$$\sum_{i,j=1}^{n} r(t_i - t_j) z_i z_j = \sum_{i,j=1}^{n} E(X(t_i)X(t_j)) z_i z_j = E\left(\sum_{i=1}^{n} X(t_i) z_i\right)^2 \geq 0$$

y por el Teorema de Bochner (Teorema 1.7) es la Transformada de Fourier de una función de distribución que llamaremos S, es decir, r tiene una representación espectral de la forma:

$$r(t) = \int_{-\infty}^{\infty} e^{i\omega t} dS(\omega) \tag{1.7}$$

donde S se conoce como la función de distribución espectral. Su derivada s, si existe, es la densidad espectral y se conoce también como el espectro.

Si la función de covarianza es integrable, entonces (1.7) es invertible

$$S(\omega) = \frac{2}{\pi} \int_0^{\infty} \cos(\omega t) r(t) dt \tag{1.8}$$

Usando la representación espectral tenemos

$$\begin{aligned} r'(t) &= \int_{-\infty}^{\infty} -\omega \sin(\omega t) dS(\omega) \\ r''(t) &= \int_{-\infty}^{\infty} -\omega^2 \cos(\omega t) dS(\omega) \end{aligned}$$

y en particular

$$r''(0) = -\int_{-\infty}^{\infty} \omega^2 dS(\omega)$$

La integral anterior se conoce como el segundo momento espectral, y lo denotamos por

$$m_2 = \int_{-\infty}^{\infty} \omega^2 dS(\omega) = -r''(0) \tag{1.9}$$

Si r no es dos veces diferenciable en 0, entonces:

$$m_2 = \int_{-\infty}^{\infty} \omega^2 dS(\omega) = \infty$$

La existencia del segundo momento espectral está asociada a la regularidad de las trayectorias del proceso. Cuando $m_2 < \infty$ la función de covarianza tiene el siguiente desarrollo cerca del origen:

$$r(t) = \sigma^2 - \frac{m_2 t^2}{2} + o(t^2)$$

Definición 1.1.7 *Un proceso estocástico, $X(t)$, se dice que es* ***continuo en media cuadrática*** *(o L^2-continuo) en tiempo t, si*

$$X(t+h) \overset{m.c}{\to} X(t)$$

cuando $h \to 0$, es decir, $\mathbb{E}[(X(t+h) - X(t))^2] \to 0$
Es llamado ***diferenciable en media cuadrática*** *con derivada $Y(t)$ si*

$$\frac{X(t+h) - X(t)}{h} \overset{m.c}{\to} Y(t)$$

cuando $h \to 0$. Por supuesto, el proceso $Y(t)$ se llama derivada *(en media cuadrática) de $X(t)$ y se denota $X'(t)$.*

Criterio 1.1.1 ***Criterio de Loéve:*** *Una sucesión x_n converge en media cuadrática si y sólo si*

$$\mathbb{E}(x_m x_n) \text{ tiene límite finito } c$$

cuando m y n tienden a infinito independientemente.

Lema 1.1.3 *Sea $r(t) = \int_{-\infty}^{\infty} e^{i\omega t} dF(\omega) = \int_{-\infty}^{\infty} \cos(\omega t) dF(\omega)$, entonces*

a.

$$\lim_{t \to 0} \frac{2(r(0) - r(t))}{t^2} = m_2 = \int_{-\infty}^{\infty} \omega^2 dF(\omega) \leq \infty$$

b. Si $m_2 < \infty$ entonces $r''() = -m_2$ y $r''(t)$ existe para todo t.

c. Si $r''(0)$ existe y es finito, entonces $m_2 < \infty$ y por (b) $r''(t)$ existe para todo t.

Demostración. Si $m_2 < \infty$ (a) se sigue de

$$\frac{2(r(0) - r(t))}{t^2} = \int_{-\infty}^{\infty} \omega^2 \frac{1 - \cos(\omega)}{\omega^2 t^2/2} dF(\omega)$$

por el Teorema de Convergencia Dominada, dado que $0 \leq \frac{1-\cos(\omega t)}{\omega^2 t^2/2} \leq 1$.
Si $m_2 = 0$, el resultado se obtiene del Lema de Fatou, ya que

$$\lim_{t \to 0} \frac{1 - \cos(\omega t)}{\omega^2 t^2/2} = 1.$$

Para demostrar (b), suponga que $m_2 < \infty$. Entonces es posible derivar dos veces bajo el signo integral en $r(t) = \int_{-\infty}^{\infty} \cos(\omega t) dF(\omega)$ para obtener que $r''(t)$ existe y

$$-r''(t) = \int_{-\infty}^{\infty} \omega^2 \cos(\omega t) dF(\omega),$$

y en particular $-r''(0) = m_2$.
Para (c), suponga que $r(t)$ tiene segunda derivada continua en el origen, y es diferenciable implícitamente cerca del origen. Aplicando el teorema del valor medio, existen $\theta_1, \theta_2 \in (0, 1)$ tal que

$$\frac{2(r(0) - r(t))}{t^2} = -r''((\theta_1 - \theta_2)t) \to -r''(0)$$

cuando $t \to 0$. Entonces, la parte (a) muestra que $m_2 < \infty$ y queda demostrado el Lema. □

Teorema 1.1.4 *Un proceso estacionario $X(t)$ es diferenciable en media cuadrática si y solo si su función de covarianza $r(t)$ es dos veces continuamente diferenciable en un entorno de $t = 0$. El proceso derivado $X'(t)$ tiene función de covarianza*

$$r_{X'}(t) = Cov(X'(s + t), X'(s)) = -r''(t).$$

Demostración. Para la parte "si" usemos el Criterio de Loéve, y demostremos que, si $h, k \to 0$ independientes uno de otro, entonces

$$\mathbb{E}\left[\frac{X(t+h) - X(t)}{h}, \frac{X(t+k) - X(t)}{k}\right] = \frac{1}{hk}(r(h-k) - r(h) - r(-k) + r(0)) \tag{1.10}$$

tiene límite finito c.
Definamos

$$\begin{aligned} f(h,k) &= r(h) - r(h-k) \\ f_1'(h,k) &= \frac{\partial}{\partial h} f(h,k) = r'(h) - r'(h-k) \\ f_{12}''(h,k) &= \frac{\partial^2}{\partial h \partial k} f(h,k) = r''(h-k) \end{aligned}$$

Aplicando el teorema del valor medio, vemos que existen $\theta_1, \theta_2 \in (0,1)$ tales que (1.10) es igual a

$$\begin{aligned} -\frac{f(h,k)-f(0,k)}{hk} &= -\frac{f_1'(\theta_1 h, k)}{k} \\ &= -\frac{f_1'(\theta_1 h, 0) - k f_{12}''(\theta_1 h, \theta_2 k)}{k} \\ &= -f_{12}''(\theta_1 h, \theta_2 k) = -r''(\theta_1 h - \theta_2 k). \end{aligned} \tag{1.11}$$

Dado que $r''(t)$ es continua por hipótesis, esta tiende a $-r''(0)$ cuando $h, k \to 0$, el cual es el límite que requiere el criterio de Loéve.

Para demostrar la parte "solo si" usaremos el Lema 1.1.3. Si $X(t)$ es diferenciable en media cuadrática, $(X(t+h) - X(t))/h \xrightarrow{m.c} X'(t)$ con $\mathbb{E}(|X(t)|^2)$ finito y

$$\mathbb{E}(|X'(t)|^2) = \lim_{h\to 0} \mathbb{E}\left[\left|\frac{X(t+h)-X(t)}{h}\right|^2\right] = \lim_{h\to 0}\frac{2(r(0)-r(h))}{h^2} = m_2 < \infty.$$

La parte (b) del Lema 1.1.3 muestra que $r''(t)$ existe.
Para ver que la función de covarianza de $X'(t)$ es igual a $-r''(t)$, tomemos el límite de

$$\begin{aligned} &\mathbb{E}\left(\frac{X(s+t+h)-X(s+t)}{h}\,\overline{\frac{X(s+t)-X(s)}{k}}\right) \\ =\ & \frac{1}{hk}(r(t+h-k) - r(t+h) - r(t-k) + r(t)) \\ =\ & -f_{12}''(t+\theta_1 h, \theta_2 k) = -r''(t+\theta_1 h - \theta_2 k) \to -r''(t) \end{aligned}$$

para algún θ_1, θ_2 cuando $h, k \to 0$. □

Como consecuencia del Teorema 1.1.4, tenemos el siguiente corolario:

Corolario 1.1.1 *Sea $X(t)$ un proceso Gaussiano estacionario tal que $m_2 < \infty$, entonces*

$$E(X'(t)) = 0, \qquad Var(X'(t)) = -r''(0) = m_2$$

Este proceso $X'(t)$ es Gaussiano, independiente de $X(t)$ para t fijo y tiene función de covarianza

$$Cov(X'(t), X'(t+h)) = -r''(h) \tag{1.12}$$

La densidad conjunta de X y X' es

$$p(u,z) = \frac{1}{2\pi\sigma\sqrt{m_2}} e^{\left\{-\frac{1}{2}\left(\frac{u^2}{\sigma^2}+\frac{z^2}{m_2}\right)\right\}} \tag{1.13}$$

En el estudio de las olas el espectro juega un papel fundamental y se interpreta como es usual, como la distribución de la energía por frecuencia. Observamos que $r(0) = Var(X(t)) = \int_{-\infty}^{\infty} S(\omega)d\omega$, de modo que la varianza del proceso representa la energía total. El espectro es simétrico y normalmente sólo se considera la parte positiva, y se renormaliza multiplicándola por 2. Por lo general se usa la misma notación S para el espectro renormalizado.

1.1.2. Características de las olas

Altura significativa

La medida más importante de la severidad del mar es la altura significativa. Esta medida trata de indicar la altura de las olas más altas que uno puede encontrarse durante un período razonable de tiempo. Significativa quiere decir que es suficientemente alta como para tener efecto sobre un barco o sobre una estructura colocada en el océano. Una de las definiciones generalmente aceptadas es la siguiente:

Definición 1.1.8 *La altura significativa de un estado del mar se define como*

$$H_m = 4\sqrt{Var(X(t))}. \tag{1.14}$$

Características basadas en cruces del nivel medio

Otras características importantes también tienen más de una definición. La definición usual está basada en los cruces del cero. Sea $X(t)$ el proceso que modela las olas en un punto del espacio; $X(t)$ representa la altura sobre el nivel medio del mar. Supongamos que $X(t)$ cruza el nivel medio hacia abajo en los instantes $t_1, t_2, ..., t_n$. El tiempo entre dos cruces sucesivos del nivel medio hacia abajo definen el período de la ola. Usaremos la notación $T_{d,k}$:

$$T_{d,k} = t_{k+1} - t_k.$$

La distancia vertical entre el máximo y el mínimo valor del proceso en este intervalo se define como la altura de la ola. Usaremos la notación $H_{d,k}$.

Una cresta a_C es el máximo valor de X para t en un intervalo entre dos cruces sucesivos hacia abajo del nivel medio: $t_k < t < t_{k+1}$. De manera similar un *seno* a_S es el (valor absoluto del) valor mínimo en el mismo intervalo de tiempo. Para distinguir diferentes crestas y senos en los intervalos sucesivos ponemos un índice k en cada valor: $(a_{C,k}, a_{S,k})$. De esta forma se tiene que

$$H_{d,k} = a_{C,k} - a_{S,k}.$$

Características basadas en extremos sucesivos

Otra definición posible considera todos los extremos de la ola. Supongamos que la ola tiene una sucesión de mínimos locales $(X_{m,k})$ y máximos locales consecutivos $(X_{M,k})$. Una ola min-max es el par $(X_{m,k}, X_{M,k})$ de valores mínimo y máximo consecutivos. La altura de la ola min-max es la diferencia entre estos valores; esto es $H_k = X_{M,k} - X_{m,k}$, mientras que el período min-max es la diferencia de tiempos correspondiente. De manera similar se definen las olas max-max y min-min.

1.1.3. Otras características

El momento espectral de orden n se define como

$$m_n = \int_0^\infty \omega^n S(\omega) d\tau. \tag{1.15}$$

En la sección 1.1.1 definimos el segundo momento espectral y vimos que para un proceso Gaussiano, su existencia está relacionada con la regularidad de las trayectorias. Esto es cierto en general: la existencia de momentos de orden superior está asociada a una mayor regularidad de las trayectorias.

Los momentos espectrales no son los únicos parámetros asociados a la densidad espectral de interés en el estudio del mar. Ya vimos que la altura significativa es otro de ellos. En términos de los momentos espectrales la altura significativa es

$$H_m = 4\sqrt{Var(X(t))} = 4\sqrt{m_0}. \tag{1.16}$$

A partir del espectro vemos que la frecuencia media está dada por

$$\frac{m_1}{m_0}.$$

Si el espectro está concentrado alrededor de una frecuencia dominante, la frecuencia media da el período medio. Otro parámetro de interés es el ancho espectral, que se define por

$$\epsilon = \sqrt{1 - \frac{m_2^2}{m_0 m_4}}. \tag{1.17}$$

Capítulo 2

Una comparación de dos métodos para el análisis espectral de olas

2.1. Introducción

En este capítulo estudiaremos dos métodos de segmentación de series temporales. Para implementar esta aproximación es necesario tener maneras de detectar los cambios de estado en el proceso, y por consiguiente las características espectrales de la estructura de covarianza de un proceso estacionario. Por lo tanto, es razonable estudiar métodos basados en la detección de cambios en los espectros. Dos de tales métodos son: Detection of Changes by Penalized Contrasts (DCPC) Lavielle (1998 [17], 1999 [18]), Lavielle y Ludeña (2000 [19]), y Smooth Localized complex EXponentials (SLEX) Ombao et. al (2002 [22]). Ambos métodos han sido implementados por sus autores en Matlab y han sido exitosamente usados en otras áreas, particularmente para el análisis de electroencefalogramas.

Para comparar su desempeño consideraremos tres conjuntos de datos. Los dos primeros corresponden a olas en situación normal mientras que el tercer conjunto corresponde a los datos del huracán Camille, los cuales tienen una situación de alta no-estacionaridad. En todos los casos compararemos los intervalos de estacionaridad obtenidos con ambos métodos y también estudiaremos tres características espectrales: Energía Total, Valor Máximo y Frecuencia Dominante. Trazaremos un gráfico y compararemos la evolución de las tres características espectrales para ambos métodos con los tres conjuntos de datos.

En las dos siguientes secciones daremos una breve descripción de ambos métodos, luego consideraremos

los dos conjuntos de datos correspondientes a condiciones normales y finalmente estudiaremos los datos del huracán Camille.

2.2. El método DCPC

El problema de estimación de puntos de cambios de una sucesión de un proceso aleatorio estacionario a trozos ha recibido considerable atención en la literatura (véase, por ejemplo, Brodsky & Darkhovsky (1993) [5], Basseville & Nikiforov (1993) [3]). Describiremos aquí brevemente un método propuesto por M. Lavielle (1998 [17], 1999 [18]) y estudiado en detalle por Lavielle y Ludeña (2000 [19]).

Consideremos una sucesión de variables aleatorias reales $Y_1, \ldots, Y_n$ y supongamos que la distribución del proceso depende de un parámetro θ que cambia abruptamente en algunos instantes desconocidos $(t_j, 1 \leq j \leq K)$, donde K es también desconocido. Para estimar K y los puntos de cambio $(t_j, 1 \leq j \leq K)$ usaremos una función de contraste penalizado de la forma

$$J(t, y) + \beta \text{pen}(t). \tag{2.1}$$

El término $J(t, y)$ estima el punto de cambio mientras que el término de penalización previene al algoritmo de sobreestimar el número de puntos de cambio. Este último sólo depende de la dimensión $K(t)$ del modelo y crece con K. El parámetro de penalización β ajusta el balance entre la minimización de $J(t, y)$ lo cual requiere típicamente valores grandes de K y la minimización de $\text{pen}(t)$ la cual va en dirección contraria.

El principio general propuesto en el algoritmo DCPC es el siguiente: Para cada $1 \leq k \leq K$, sea $U(Y_{t_{k-1}}, \ldots, Y_{tk}; \theta)$ una función de contraste usada para la estimación del valor del parámetro desconocido θ. El estimador de contraste mínimo $\hat{\theta}(Y_{t_{k-1}}, \ldots, Y_{t_k})$ calculado para el k-ésimo segmento de t se define como la solución del siguiente problema de minimización:

$$U(Y_{t_{k-1}}, \ldots, Y_{t_k}; \hat{\theta}(Y_{t_{k-1}}, \ldots, Y_{t_k})) \leq U(Y_{t_{k-1}}, \ldots, Y_{t_k}; \theta), \tag{2.2}$$

para todo $\theta \in \Theta$. Para $1 \leq k \leq K$ sea

$$C(Y_{t_{k-1}} + 1, \ldots, Y_{t_k}) = \frac{1}{n} U(Y_{t_{k-1}}, \ldots, Y_{t_k}; \hat{\theta}(Y_{t_{k-1}}, \ldots, Y_{t_k})) \tag{2.3}$$

entonces la función de contraste se define como

$$J(t, y) = \sum_{k=1}^{K} C(y_{t_{k-1}}, \ldots, y_{t_k}) \tag{2.4}$$

donde $t_0 = 0$ y $t_k = n$.

Usando este principio general, diferentes funciones de contraste se pueden usar de acuerdo a la situación. En el caso de cambios en la distribución espectral, consideraremos que la energía del proceso en cierta banda de frecuencia $[\lambda_j, \mu_j), 1 \leq j \leq J$, cambia repentinamente. Para cada k y cada $u \in [0, \pi]$ sea

$$I_k(u) = \frac{1}{2\pi n_k} \left| \sum_{t=t_{k-1}+1}^{t_k} Y_t e^{itu} \right|^2 \tag{2.5}$$

el periodograma de la sucesión (Y_j) en la banda de frecuencia $[\lambda_j, \mu_j)$ y sea

$$F_{kj} = \int_{\lambda_j}^{\mu_j} I_k(u) du \tag{2.6}$$

la energía de $(Y_{t_{k-1}}, \ldots, Y_{t_k})$ en la banda de frecuencia $[\lambda_j, \mu_j)$. El contraste usado para detectar los puntos de cambio es

$$C(y_{t_{k-1}}, \ldots, y_{t_k}) = -\frac{n_k}{n} \sum_{j=1}^{J} F_{kj}^2. \tag{2.7}$$

El algoritmo DCPC ha sido implementado en Matlab para diferentes criterios: Cambios en media, varianza, media y varianza, función de distribución y espectros. Este se puede descargar de la página web personal de M. Lavielle: http://www.math.u-psud.fr/∼lavielle/programs/index.html.

2.3. El método SLEX

El algoritmo auto-SLEX es un procedimiento estadístico que divide una serie temporal en segmentos que son aproximadamente estacionarios y automáticamente elige un parámetro de suavizado para la estimación de los espectros que cambian en el tiempo. El método está basado en la transformada SLEX (Smooth Localized complex EXponential), la cual usa los vectores SLEX y el cual está muy relacionado con la transformada clásica de Fourier. El método se presenta en Ombao et. al (2002 [22]) y aquí seguimos su presentación. El algoritmo ha sido implementado en Matlab y está disponible en la página web: www.stat.uiuc.edu/∼ombao.

Como es bien sabido, las funciones de Fourier son apropiadas para representar procesos aleatorios estacionarios, dado que ellas están localizadas en frecuencia y las propiedades espectrales de los procesos estacionarios son invariantes en el tiempo, sin embargo ellas no pueden representar procesos con propiedades espectrales que evolucionan en el tiempo. Para lidiar con el problema de localización en tiempo, se aplican ventanas suaves de soporte compacto, sin embargo la funciones allí dejan de ser ortogonales. Es bien sabido

que no existe una ventana suave tal que la base de vectores de Fourier en ella sean ortogonales y localizadas en tiempo y frecuencia al mismo tiempo. Las funciones SLEX evitan este problema usando un operador de proyección, en vez de una ventana, sobre las exponenciales complejas. Resulta que la acción del operador de proyección sobre una función periódica es equivalente a aplicar dos ventanas suaves especialmente construidas para la base de funciones de Fourier.

Las funciones en la base SLEX $\phi_\omega(u)$ son de la forma

$$\phi_\omega(u) = \Psi_+(u)\exp(i2\pi\omega u) + \Psi_-\exp(-i2\pi\omega u) \tag{2.8}$$

donde $\omega \in [-1/2, 1/2]$ y $\Psi_+(u)$ y $\Psi_-(u)$ son funciones suaves a valores reales específicas que definiremos luego. Las funciones en la base SLEX tienen soporte en $[-\delta, 1+\delta]$, donde $0 < \delta < 0{,}5$. Entonces las funciones SLEX se solapan en diferentes bloques diádicos pero permanecen ortogonales.

La base de funciones SLEX generaliza directamente la base de vectores ortogonales SLEX para representación de series temporales. Sean $a_0 < a_1$ dos tiempos enteros, $|S| = a_1 - a_0$ y el solapamiento $\varepsilon = [\delta|S|]$, donde $[\cdot]$ denota la parte entera. El soporte $\bar{S}$ de los vectores SLEX en el bloque S consiste de puntos de tiempo definidos en S y el solapamiento: $\bar{S} = \{a_0 - \varepsilon, \ldots, a_0, \ldots, a_0, \ldots, a_1 - 1, a_1 - 1 + \varepsilon\}$. Una base de vectores SLEX definido en el bloque S tiene elementos $\{\phi_{S,\omega_k,t}\}$ con

$$\begin{aligned} \phi_{S,\omega_k}(t) &= \phi_{\omega_k}((t-a_0)/|S|) \\ &= \Psi_{S,+}((t-a_0)/|S|)\exp\{i2\pi\omega_k(t-a_0)\} \\ &\quad +\Psi_{S,-}((t-a_0)/|S|)\exp\{-i2\pi\omega_k(t-a_0)\} \end{aligned} \tag{2.9}$$

donde $\omega_k = k/|S|, k = -|S|/2+1, \ldots, |S|/2$. La ventana se puede representar en términos de la función de levantamiento de corte r:

$$\begin{aligned} \Psi_{S,+}(t) &= r^2\left(\frac{t-a_0}{\varepsilon}\right) r^2\left(\frac{a_1-t}{\varepsilon}\right) \\ \Psi_{S,-}(t) &= r\left(\frac{t-a_0}{\varepsilon}\right) r\left(\frac{a_0-t}{\varepsilon}\right) - r\left(\frac{t-a_1}{\varepsilon}\right) r\left(\frac{a_1-t}{\varepsilon}\right). \end{aligned} \tag{2.10}$$

En la implementación específica usaremos r como

$$r(u) = \text{sen}\left(\frac{\pi}{4}(1+u)\right) \tag{2.11}$$

donde $u \in [-1, 1]$.

Auto-SLEX también usa el Algoritmo de Mejor Bases (BBA, *siglas en inglés*) de Coifman y Wickerhauser (1992 [7]) para elegir la mejor segmentación usando una función de costo definida en términos de los logaritmos de los periodogramas SLEX. Primero, se calcula el espectro SLEX para todo el conjunto de datos, entonces se divide en dos el conjunto de datos y se calcula el espectro SLEX para cada mitad. Se calcula el costo de cada configuración y el algoritmo elige la configuración de menor costo. El costo de una configuración dada es determinado por el criterio de Complejidad Penalizado Kullback-Leibler

$$Cost(j,b) = \sum_{k=-\frac{M_j}{2}+1}^{\frac{M_j}{2}} \log\det\left(I_{j,b,k}\right) + \beta(j,b)\sqrt{M_j} \tag{2.12}$$

donde Ij, b, k es la matriz del periodograma suavizado, M_j es la longitud del bloque al nivel j, $M_j = T/2^j$ donde T es la longitud del vector de datos y b es el índice del bloque, $b = 0, 1, \ldots, 2^j - 1$.

Primero, se calcula el espectro SLEX para todo el conjunto de datos, entonces se divide el conjunto en 2 segmentos y se calcula el espectro SLEX para cada uno. Se calcula el costo de cada configuración y luego el algoritmo escoge la configuración de menor costo. Este procedimiento se repite hasta que se consigue la mejor configuración o se alcanza el tamaño mínimo para los intervalos. El espectro SLEX se calcula usando el algoritmo FFT y el conjunto de datos debe tener longitud una potencia de 2. Dado que los subintervalos se obtienen de divisiones sucesivas en 2 del conjunto original, la longitud de los intervalos obtenidos también tienen longitud una potencia de 2, y sus puntos iniciales y finales son sumas de potencias de 2. En la Figura 2.1 se puede observar una subdivisión del intervalo para $j = 2$, donde los bloques en tono claro son los que tienen menor costo y los oscuros son descartados para la configuración final.

Dado que existe un tamaño mínimo para los intervalos, relacionado con el conjunto más pequeño de datos requerida para una buena estimación del espectro, lo cual en nuestro caso es $2^{10} = 1024$, existe una limitación en la precisión con la cual el algoritmo puede detectar los puntos de cambios en la serie temporal. Más detalles se pueden conseguir en Ombao et al. (2002 [22]).

2.4. Análisis de datos de olas para mares en estado normal

Consideremos dos conjuntos de datos correspondientes a 3 días en septiembre de 2005, iniciando el 1^{ro} de septiembre a las 0h, para dos boyas desplegadas por Costal Data Information Program, Integrative Oceanography Division, operado por el Scripps Institution of Oceanography (http://cdip.ucsd.edu/): Estación 067 Isla San Nicolás en las costas de California con una profundidad de 360m y la Estación 106 en la Bahía Waimea,

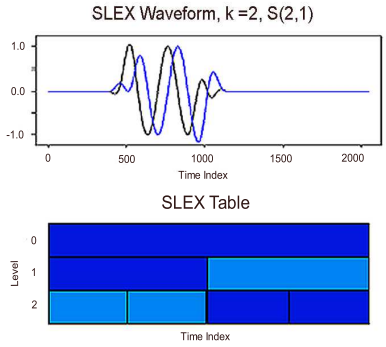

Figura 2.1: Subdivisión del archivo.

Hawaii con una profundidad de 200m. En ambos casos la frecuencia de muestreo es 1.28Hz. Usaremos ambos métodos de segmentación para estos conjuntos de datos.

2.4.1. Estación 067

Presentamos en la Figura 2.2 los datos para la Estación 067 junto con la segmentación producida por ambos métodos. En la parte superior están los puntos de cambios obtenidos con el algoritmo DCPC y en la parte inferior los obtenidos con el algoritmo SLEX. Los valores están dados en las Tablas 2.1 y 2.2.

00:46:57	4:20:17	6:06:57	7:53:37	11:26:57
15:00:17	16:45:31	17:40:17	18:33:37	22:06:57
23:53:37	25:40:17	27:26:57	28:20:17	29:13:37
36:20:17	43:26:57	45:13:37	47:00:17	48:46:57
50:33:37	54:06:57	57:40:17	61:13:37	64:46:57
68:20:17	71:53:37			

Tabla 2.1: Puntos de cambios dados por SLEX para la Estación 067 (min).

Como se puede ver de la Figura 2.2 y las tablas, el algoritmo DCPC produce más segmentos: 45 vs. 27 y los puntos de cambios están en diferentes posiciones. De hecho, ninguno de los 27 puntos de cambio de SLEX está a menos de 5 minutos de los obtenidos con DCPC, 2 tienen 10 minutos de diferencia, otros 5 con 20 minutos y 6 más con una diferencia de media hora. En general las segmentaciones difieren. La Tabla 2.3 muestra un

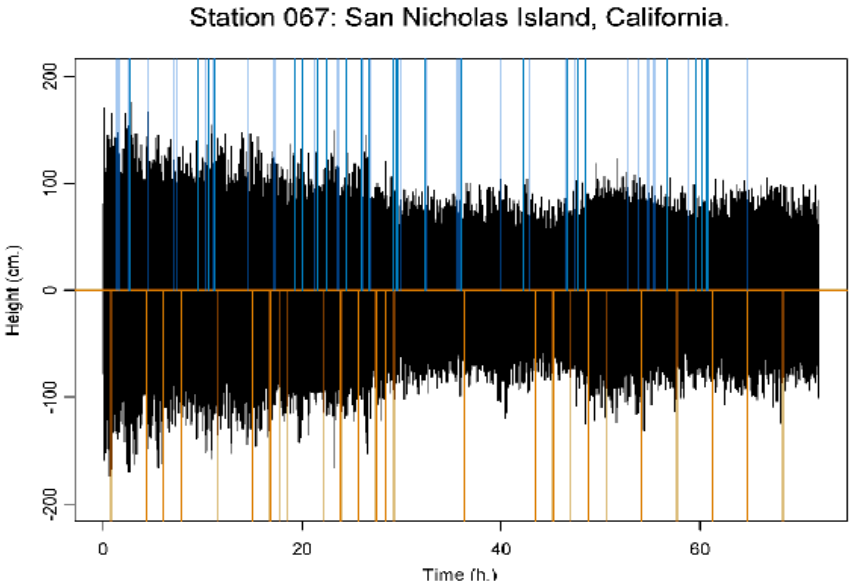

Figura 2.2: Altura de olas para la Estación 067. La segmentación DCPC se muestra en la parte superior la segmentación SLEX en la parte inferior.

1:22:54	1:42:22	2:37:58	4:35:23	7:02:01
7:21:43	9:31:54	10:16:09	10:36:59	11:10:06
14:39:37	17:13:11	19:13:50	20:01:35	21:10:15
21:35:10	22:29:56	23:36:54	24:29:24	26:00:46
26:45:18	29:07:20	29:31:29	29:50:56	32:24:39
35:37:51	35:56:25	40:02:01	42:15:02	42:48:13
46:36:28	47:25:24	47:44:25	48:31:35	52:40:32
53:49:19	54:47:27	55:22:34	56:41:26	58:50:57
59:34:22	60:09:35	60:42:26	64:41:54	71:53:37

Tabla 2.2: Puntos de cambio dados por DCPC para la Estación 067 (min).

análisis de estadísticos básicos para la longitud de los intervalos con ambos algoritmos.

Calculamos los espectros para cada segmento usando el software WAFO, y estudiamos la evolución de distintas propiedades espectrales. Las tres propiedades en que nos enfocamos fueron: Energía Total, Valor

	SLEX	DCPC
Mínimo	49.95	9.42
1er. Cuartil	106.66	33.66
Mediana	106.66	56.86
Promedio	159.76	93.77
3er. Cuartil	213.33	132.31
Máximo	426.66	422.28
Varianza	9900.0	7224.1

Tabla 2.3: Estadísticos Básicos para la Longitud de los Intervalos, Estación 067.

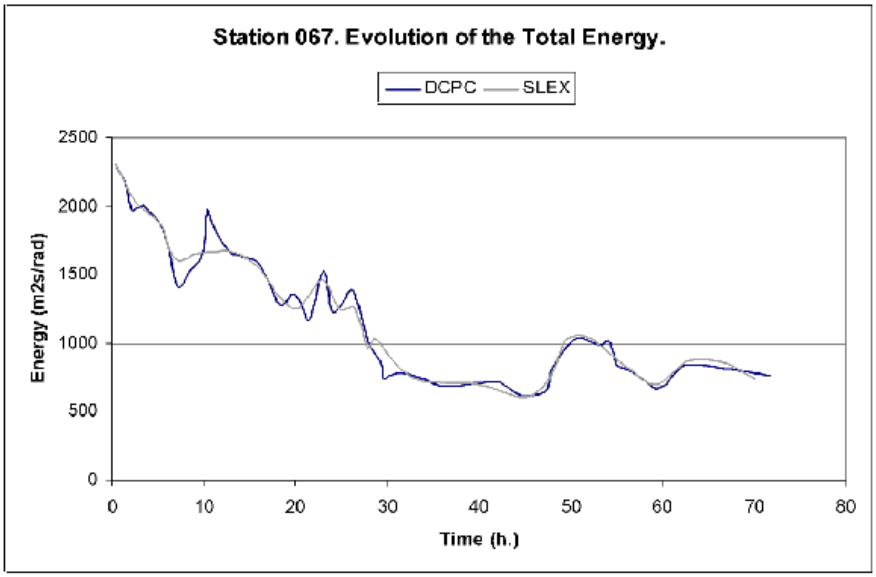

Figura 2.3: Evolución de la Energía Total para la Estación 067.

Máximo del Espectro y la Frecuencia correspondiente al valor máximo (Frecuencia Dominante). Presentamos en las Figuras 2.3, 2.4 y 2.5 la evolución de estas cantidades obtenidas con ambos métodos de segmentación. Como se puede ver, en general ambas curvas siguen un patrón similar, pero dado que DCPC tiende a producir intervalos más pequeños, éste detecta cambios que no presenta el método SLEX. Esto se puede ver en las Figuras 2.3 y 2.4 alrededor de 10h. y en la Figura 2.5 entre 55 y 65h.

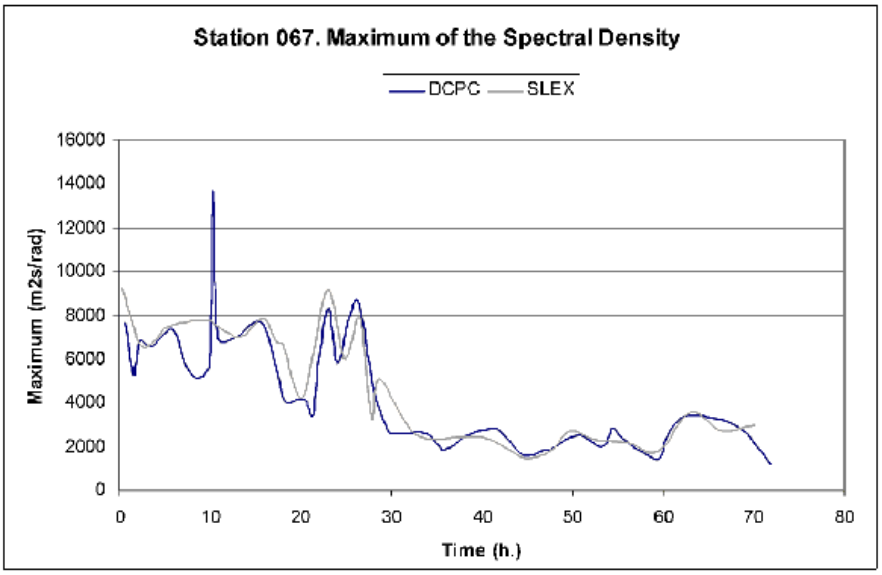

Figura 2.4: Evolución del Valor Máximo de la Densidad Espectral para la Estación 067.

Viendo en la Figura 2.2 y las Tablas 2.1 y 2.2, notamos que los intervalos producidos por el algoritmo SLEX son con frecuencia divididos en intervalos más pequeños por DCPC, pero en algunos casos sucede lo contrario. Por ejemplo, el último intervalo DCPC es dividido en 2 por SLEX. Es interesante realizar una comparación de los espectros para estos casos. Consideremos primero el intervalo SLEX con puntos finales 29:13:37 y 36:20:17, y los intervalos DCPC con puntos finales 29:07:20.63, 29:31:29.06, 29:50:56.25, 32:24:39.69, 35:37:51.88 y 35:56:25.94. Los correspondientes espectros se muestran en la Figura 2.6. La Tabla 2.4 muestra los valores de las tres propiedades consideradas: Energía Total, Máximo de Energía y Frecuencia Dominante. Los intervalos son llamados SLEX1 y DCPCn con $n = 1, \ldots, 5$.

Como se puede ver, la frecuencia dominante se mantiene aproximadamente constante a través de los intervalos alrededor de 0.66. La energía total en DCPC1 es mayor que la energía total en SLEX1 pero decrece en los siguientes 3 intervalos, siendo aproximadamente igual a la de SLEX1 y en el último intervalo de DCPC esta decrece. El valor máximo de energía muestra un patrón similar. Finalmente, existe un segundo pico en el espectro de SLEX1 que también aparece en el espectro de DCPC1, desapareciendo en DCPC2 pasando a una frecuencia mayor en DCPC3 volviendo a la misma frecuencia en DCPC4 y desapareciendo en DCPC5.

Es interesante notar que DCPC5 es un intervalo corto, de menos de 19 min de duración el cual incluye el punto final derecho de SLEX1.

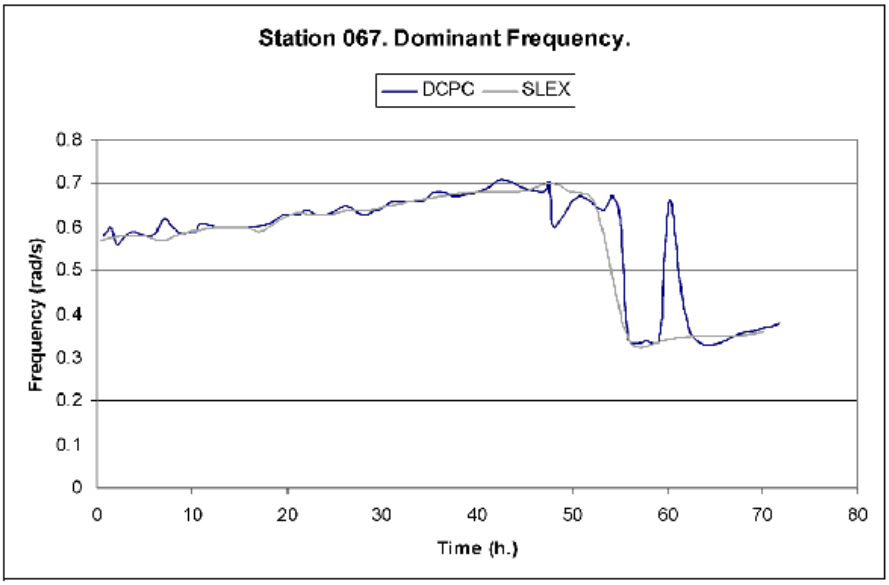

Figura 2.5: Evolución de la Frecuencia Dominante para la Estación 067.

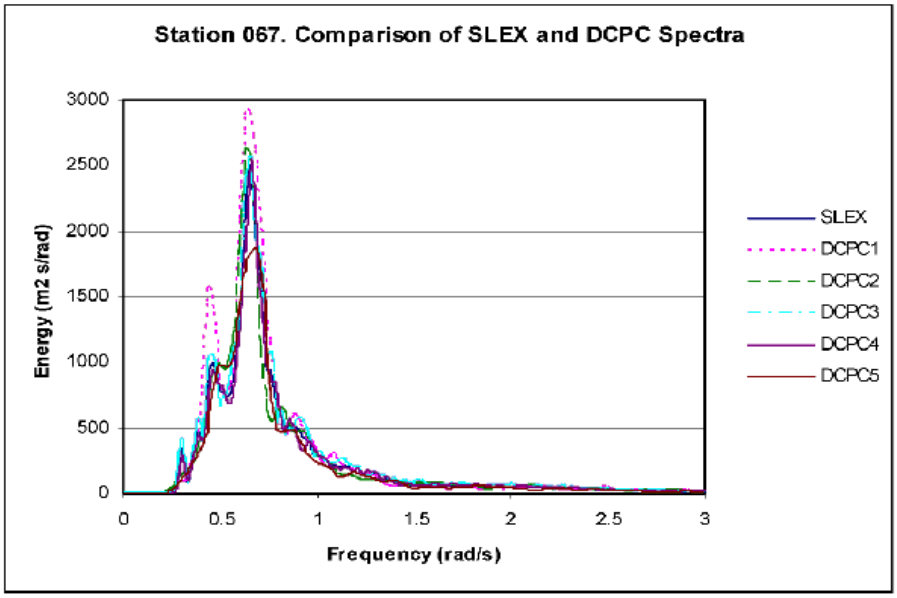

Figura 2.6: Comparación de los espectros SLEX y DCPC.

Intervalos	Energía Total (m^2s/rad)	Valor Máximo (m^2s/rad)	Frecuencia Dominante 1 (rad/s)	Frecuencia Dominante 2 (rad/s)
SLEX1	752.80	2509.20	0.66	0.46
DCPC1	862.20	2936.60	0.64	0.45
DCPC2	738.60	2628.90	0.64	
DCPC3	784.30	2568.00	0.66	0.76
DCPC4	736.70	2549.40	0.66	0.46
DCPC5	688.40	1870.00	0.68	

Tabla 2.4: Comparación de las Propiedades Espectrales.

Ahora consideremos el intervalo DCPC dividido en dos por el algoritmo SLEX. El intervalo DCPC tiene puntos extremos 64:41:54.84 y 71:53:37.19, mientras que los puntos extremos de los intervalos SLEX son 64:46:57.19, 68:20:17.19 y 71:53:37.19. Los espectros correspondientes se muestran en la Figura 2.7. La Tabla 2.5 muestra los valores de la energía total, valor máximo de energía y frecuencia dominante. Los intervalos son nombrados SLEX1, SLEX2 y DCPC1.

Como se puede ver en la Figura 2.7 y la Tabla 2.5 los espectros SLEX se parecen en forma, y para el primero la energía es mayor que para el intervalo SLEX mientras que para el segundo es menor. El resto de las propiedades se mantienen más o menos constantes.

Intervalos	Energía Total (m^2s/rad)	Valor Máximo (m^2s/rad)	Frecuencia Dominante 1 (rad/s)	Frecuencia Dominante 2 (rad/s)
DCPC1	806.20	2892.40	0.36	0.71
SLEX1	865.00	2719.00	0.35	0.71
SLEX2	738.80	2965.80	0.36	0.70

Tabla 2.5: Comparación de Propiedades Espectrales.

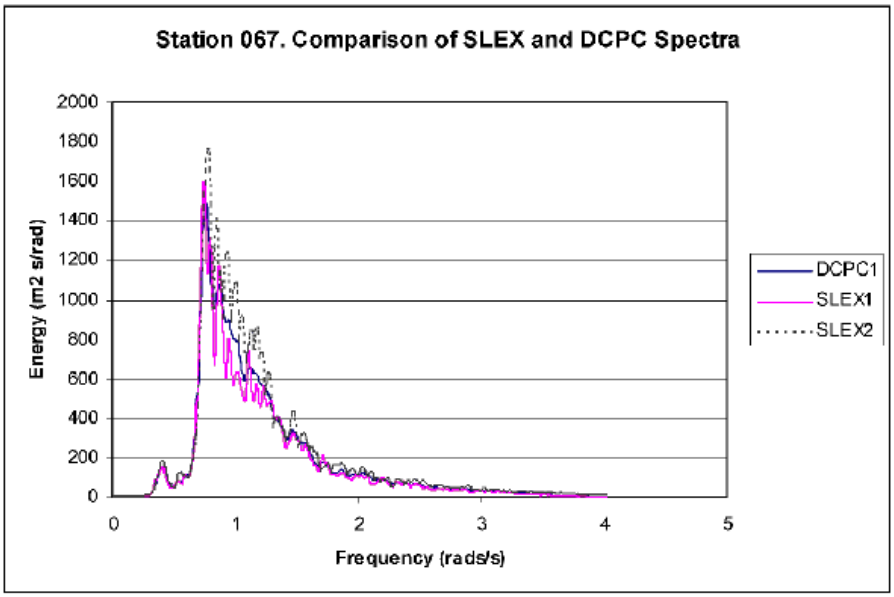

Figura 2.7: Comparación de los espectros SLEX y DCPC.

2.4.2. Estación 106

Realizamos un análisis similar para los datos de la Estación 106. La Figura 2.8 muestra los registros de alturas de olas junto con los puntos de cambio determinados por ambos algoritmos. Estos puntos de cambios se listan en las Tablas 2.6 y 2.7.

0:53:20	4:26:40	8:00:01	15:06:40	16:53:20
18:40:00	2:26:40	22:13:20	29:20:00	31:06:40
32:53:20	36:26:40	40:00:00	40:53:20	41:20:00
41:46:40	42:40:00	43:33:20	45:20:00	47:06:40
50:40:00	52:26:40	54:13:20	56:00:00	57:46:40
64:53:20	66:40:00	68:26:40	70:13:20	

Tabla 2.6: Cortes SLEX para la Estación 106 (min).

Nuevamente, el algoritmo DCPC proporciona más cambios que SLEX: 46 vs. 30. En este caso 4 de los cortes conseguidos con SLEX están a menos de 5 minutos de los conseguidos con DCPC, 6 tienen 10 minutos, 5 con

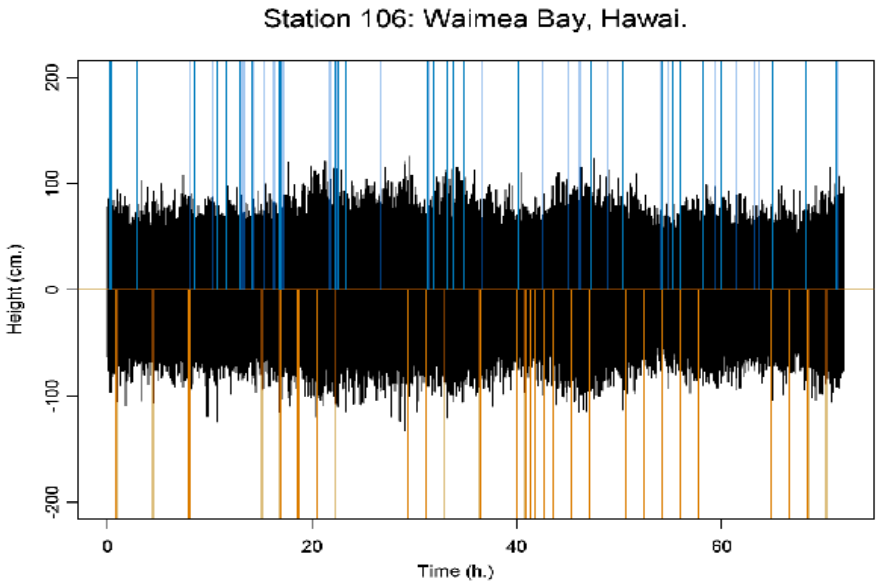

Figura 2.8: Altura de olas para la Estación 106. La segmentación DCPC se muestra en la parte superior y la segmentación SLEX en la parte inferior.

0:19:05	2:55:36	8:05:37	8:30:54	10:19:13
10:43:12	11:37:22	12:59:15	13:19:49	14:11:26
15:20:59	16:18:16	17:09:02	21:46:47	22:14:12
22:32:08	23:16:22	26:38:35	31:19:17	31:53:26
33:09:50	33:47:24	34:45:59	36:41:30	40:08:13
42:33:41	45:07:46	46:09:01	47:19:35	48:57:30
50:19:56	54:09:13	54:50:14	55:14:27	55:57:46
58:10:02	59:19:57	59:57:07	61:21:34	63:10:04
63:35:07	65:03:36	68:14:47	71:14:24	

Tabla 2.7: Cortes DCPC para la Estación 106 (min).

20 minutos y 2 con más de 30 minutos. Así que de nuevo las segmentaciones difieren aunque no tan marcadas como el caso anterior. La Tabla 2.8 muestra los estadísticos básicos para la longitud de los intervalos.

Las tres figuras siguientes muestran la evolución de la Energía Total, Máximo de la Densidad Espectral y

	SLEX	DCPC
Mínimo	26.66	17.2
1er. Cuartil	106.66	37.26
Mediana	106.66	69.73
Promedio	144.00	93.9
3er. Cuartil	186.67	128.08
Máximo	426.67	310.01
Varianza	11946.6	5933.58

Tabla 2.8: Estadísticos Básicos para la Longitud de los Intervalos, Estación 106 (min).

Frecuencia Dominante para la Estación 106. Como se puede ver, las observaciones hechas en el caso previo (Estación 067) son también válidas aquí, aunque las diferencias son menos marcadas.

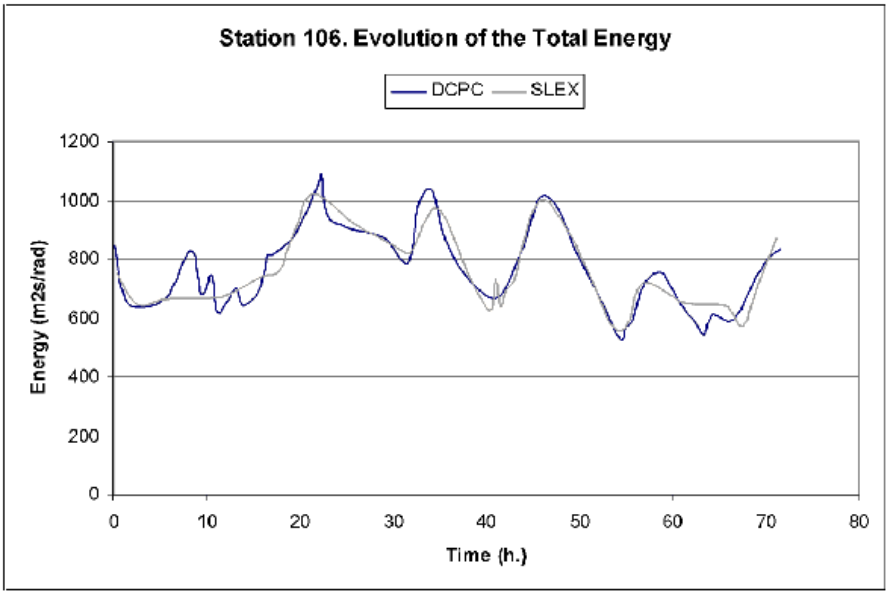

Figura 2.9: Evolución de la Energía Total para la Estación 106.

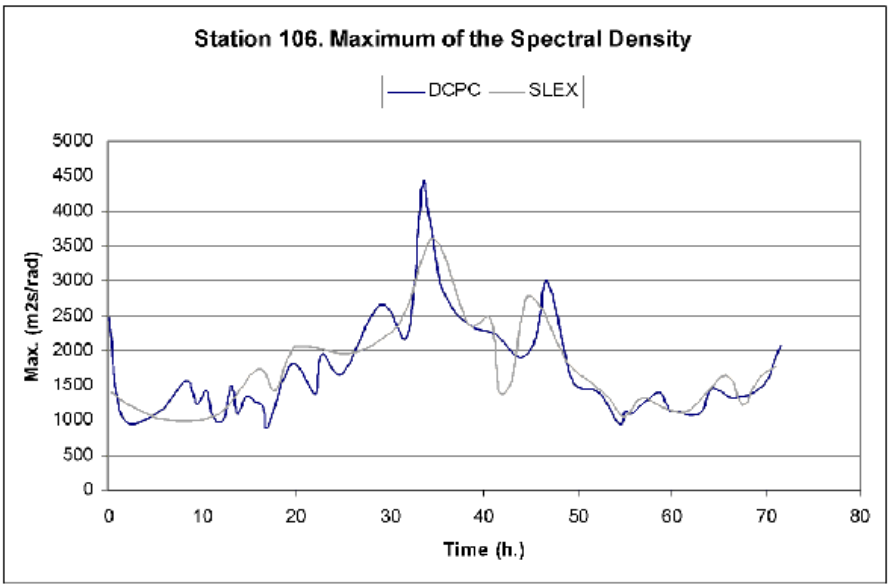

Figura 2.10: Evolución del Valor Máximo de la Densidad Espectral para la Estación 106.

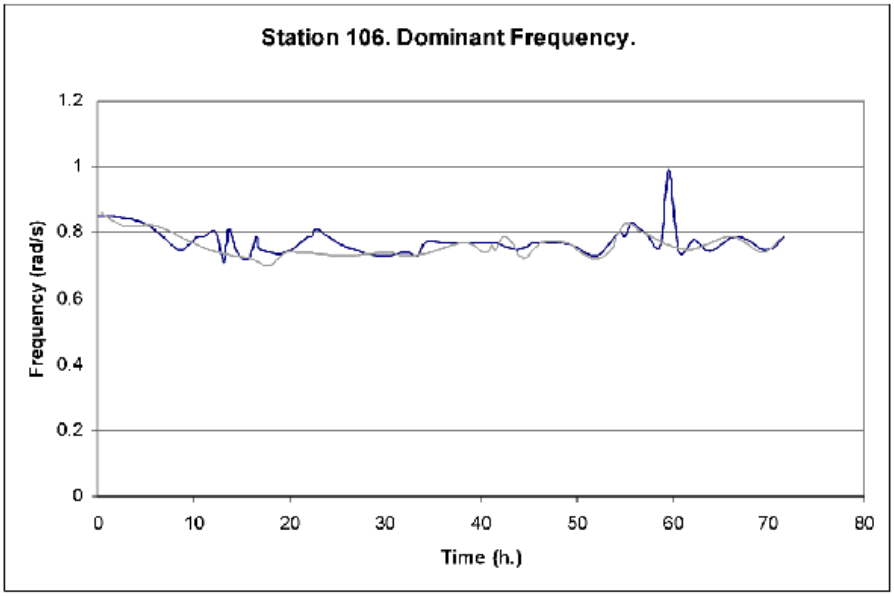

Figura 2.11: Evolución de la Frecuencia Dominante para la Estación 106.

2.4.3. Datos del huracán Camille

Como punto final en la comparación de estos dos métodos consideraremos una situación altamente no-estacionaria: Analizaremos los datos del Huracán Camille. Este conjunto de datos es bien conocido y ha sido considerado previamente por diversos autores (véase por ejemplo, Forristall (1978 [9]) y Guedes Soares et al. (2004 [11]). Aplicamos ambos métodos de segmentación a estos datos y los resultados obtenidos se presentan en la Figura 2.12 y la Tablas 2.9 y 2.10.

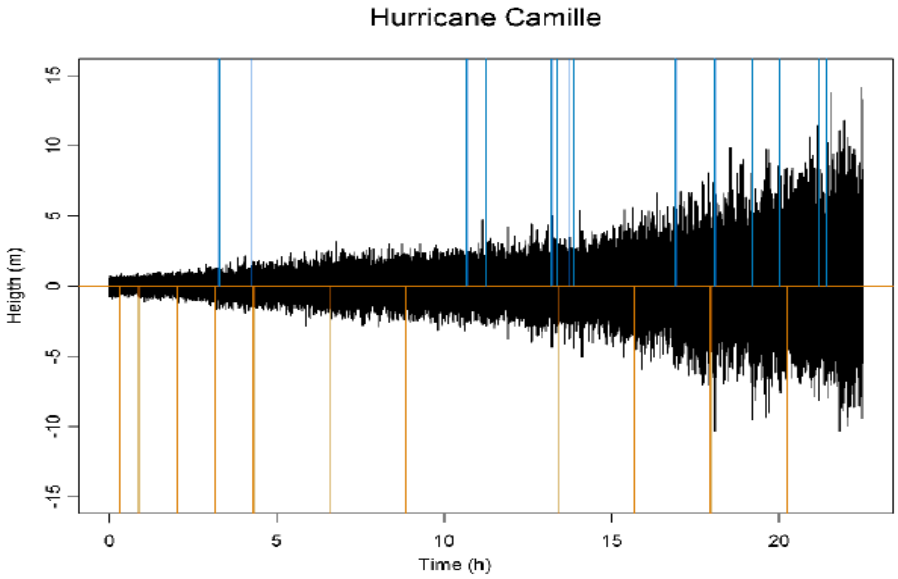

Figura 2.12: Altura de olas para el Huracán Camille. La segmentación DCPC se muestra en la parte superior y la segmentación SLEX en la parte inferior.

0:18:48	0:52:56	2:01:12	3:09:28	4:17:44
6:34:16	8:50:48	13:23:52	15:40:24	17:56:56
20:13:28				

Tabla 2.9: Cortes SLEX para el Huracán Camille (min).

Esta vez el número de puntos de cambios es similar 11 vs. 13, pero la localización es de nuevo diferente, excepto para 4 puntos de cambios que están razonablemente cercanos. Los intervalos SLEX tienden a estar

3:16:57	4:16:00	10:40:14	11:14:39	13:11:54
13:42:35	13:52:34	16:54:56	18:04:59	19:12:24
20:00:59	21:11:13	21:24:27		

Tabla 2.10: Cortes DCPC para el Huracán Camille (min).

uniformemente espaciados mientras que DCPC produce intervalos pequeños y grandes. La Tabla 2.11 muestra los estadísticos básicos para la longitud de los intervalos. Como se puede ver de esta tabla la distribución para los intervalos SLEX está más concentrada.

	SLEX	DCPC
Mínimo	18.8	9.98
1er. Cuartil	68.27	27.36
Mediana	136.53	65.55
Promedio	112.5	90.0
3er. Cuartil	136.53	93.74
Máximo	273.07	384.23
Varianza	4548.9	9888.4

Tabla 2.11: Estadísticos Básicos para la Longitud de los Intervalos, Huracán Camille (min).

Comparamos la evolución de la Energía Total, Valor Máximo de la Densidad Espectral y Frecuencia Dominante para ambos métodos. Los resultados se muestran en la Figuras 2.13 a 2.15

De nuevo, para cada gráfico las curvas tienen patrones similares, excepto al final en la Figuras 2.13 y 2.14, donde el algoritmo SLEX falla al detectar un cambio que ocurre alrededor de la hora 20. Por otra parte, el algoritmo SLEX detecta cambios en la frecuencia dominante que ocurren al principio de la serie mientras que DCPC no los detecta.

En la Figura 2.12 podemos observar que el primer intervalo DCPC es dividido en 4 por el algoritmo SLEX, mientras que el último intervalo SLEX es dividido en 3 subintervalos por el algoritmo DCPC. Ahora compararemos los correspondientes espectros en ambas situaciones. Consideremos primero el intervalo DCPC con puntos extremos 0 y 3:16:57 y los intervalos SLEX con puntos extremos 0:18:48, 0:52:56, 2:01:12 y 3:09:28. Los correspondientes espectros se muestran en la Figura 2.16 y la Tabla 2.12 muestra los valores para las 3 propiedades espectrales.

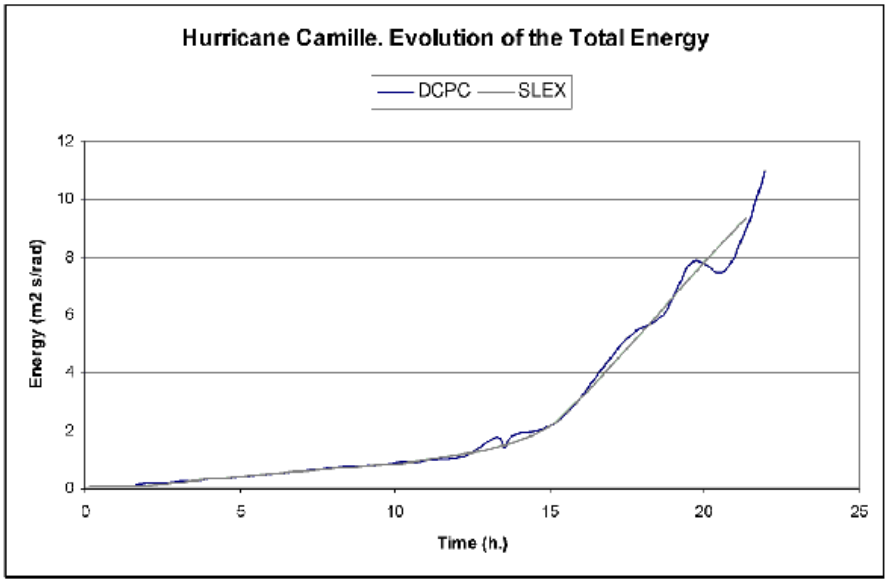

Figura 2.13: Evolución de la Energía Total para el Huracán Camille.

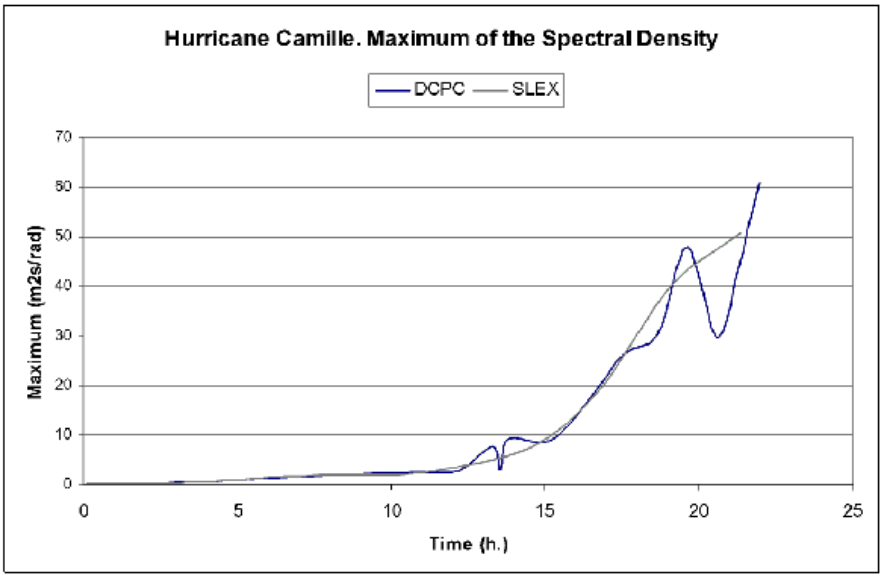

Figura 2.14: Evolución del Valor Máximo de la Densidad Espectral para el Huracán Camille.

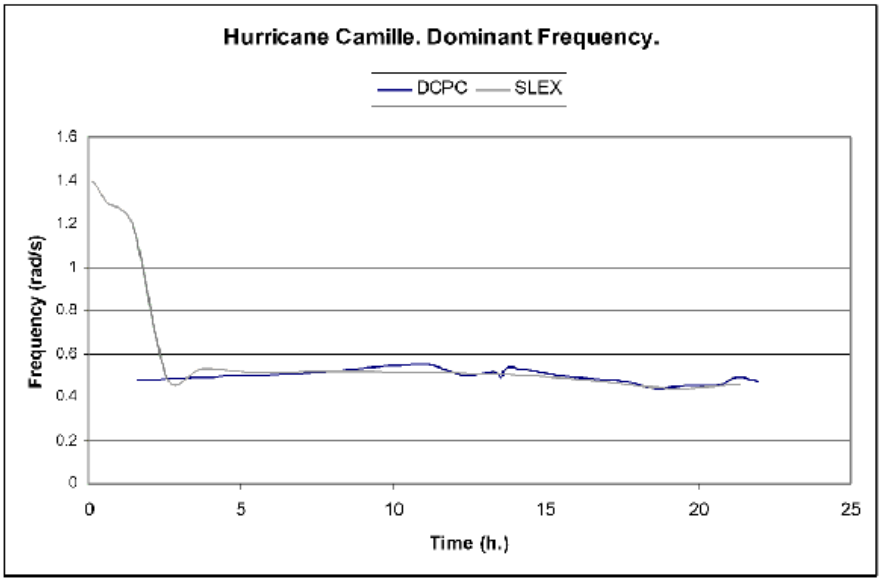

Figura 2.15: Evolución de la Frecuencia Dominante para el Huracán Camille.

Intervalos	Energía Total (m^2s/rad)	Valor Máximo (m^2s/rad)	Frecuencia Dominante 1 (rad/s)	Frecuencia Dominante 2 (rad/s)
DCPC1	0.127	0.159	0.48	1.2
SLEX1	0.0779	0.157	1.4	1.2
SLEX2	0.0791	0.165	1.3	1.2
SLEX3	0.111	0.168	1.2	0.46
SLEX4	0.169	0.382	0.48	1.2

Tabla 2.12: Comparación de las Propiedades Espectrales para el primer intervalo DCPC.

Como se puede ver, en los espectros SLEX la frecuencia dominante va decayendo y para el último intervalo esta coincide con la frecuencia dominante para el intervalo DCPC. Los 5 espectros tienen aproximadamente las mismas frecuencias dominantes, 0.48 y $1.2 rad/s$, pero sus importancias relativas son diferentes. El gráfico muestra que todos los espectros tienen al menos 4 picos.

La Energía Total y el Valor Máximo también cambian, la energía en los espectros SLEX crece y el valor en el

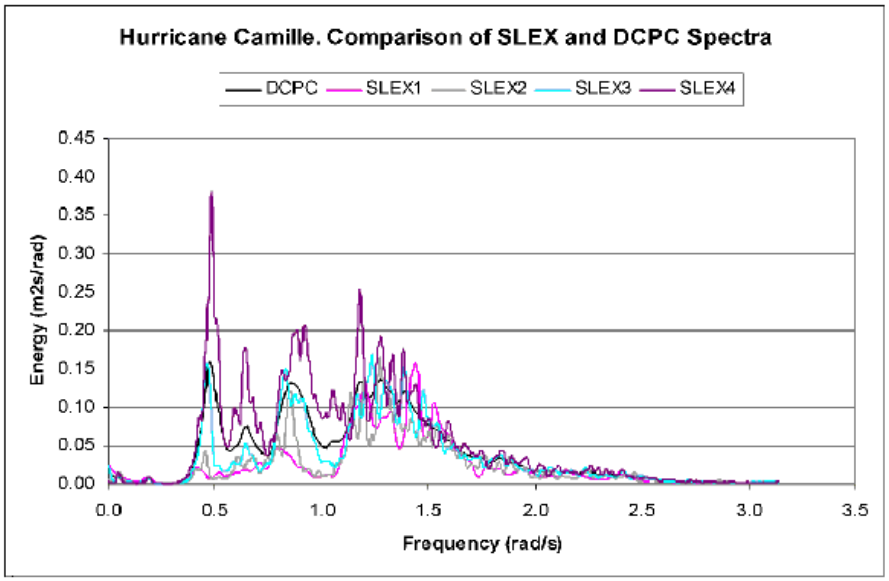

Figura 2.16: Comparación de los espectros DCPC y SLEX para el primer intervalo DCPC.

espectro DCPC está entre los valores del tercer y cuarto espectro SLEX. El valor máximo para el cuarto espectro SLEX es el doble que el del resto de los espectros.

El último intervalo SLEX con puntos extremos 20:13:28 y 22:30:00, es dividido en 3 subintervalos por el algoritmo DCPC, con puntos extremos 20:00:59, 21:11:13, 21:24:27 y 22:30:00. Los espectros se muestran en la Figura 2.17, la Tabla 2.13 muestra los valores de las propiedades espectrales.

Intervalos	Energía Total (m^2s/rad)	Valor Máximo (m^2s/rad)	Frecuencia Dominante 1 (rad/s)	Frecuencia Dominante 2 (rad/s)
SLEX1	6.35	50.94	0.46	
DCPC1	7.47	29.83	0.45	0.43
DCPC2	8.84	43.44	0.49	
DCPC3	10.982	60.93	0.47	

Tabla 2.13: Comparación de la Propiedades Espectrales para el último intervalo SLEX.

En este caso la frecuencia dominate permanece aproximadamente constante y el mayor cambio ocurre en

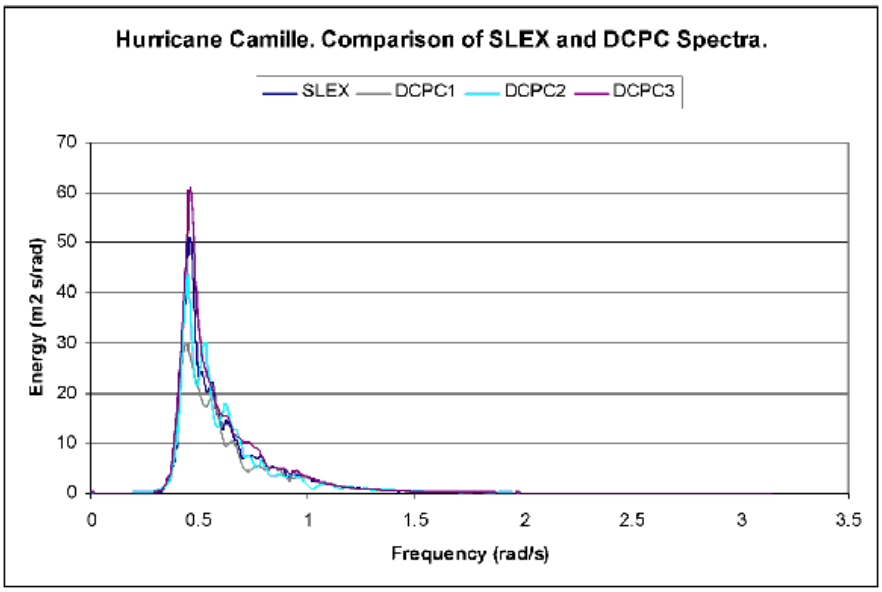

Figura 2.17: Comparación de los espectros SLEX y DCPC para el último intervalo SLEX.

la energía reflejada tanto en la Energía Total como en el Valor Máximo de la Densidad Espectral. Las formas de los diferentes espectros son similares.

2.5. Conclusiones

Hemos considerado dos métodos para detectar los puntos de cambio en una serie temporal: Detection of Changes by Penalized Contrasts (DCPC) y Smooth Localized complex EXponentials (SLEX). Estos algoritmos fueron utilizados para tres conjuntos de datos, dos de ellos provenientes de un mar en condiciones *normales* (Estaciones 067 y 106) y uno de un huracán (Huracán Camille).

Los resultados obtenidos por estos métodos difieren. En condiciones normales DCPC tiende a producir más puntos de cambios y por lo tanto intervalos más pequeños. Para los datos del Huracán Camille el número de puntos de cambios es aproximadamente el mismo, sin embargo SLEX produce intervalos más uniformes en longitud.

Para cada intervalo estimamos la densidad espectral y también analizamos algunas características espectrales: Energía Total, Valor Máximo de la Densidad Espectral y Frecuencia Dominante. El patrón general para la evolución es la misma para ambos métodos en los tres casos, pero DCPC parece capturar más variaciones

que SLEX. Esto es probablemente al hecho de que produce más puntos de cambios.

En algunos casos también comparamos los espectros para intervalos que eran subdivididos por el otro método.

El algoritmo SLEX es más rápido y fácil de usar y nos da más información de la que usamos aquí (Los espectros SLEX son calculados automáticamente y almacenados, son dados los gráficos de frecuencia vs. tiempo donde las diferentes *intensidades* para las frecuencias se muestran en un código de color. Esto da una idea de la evolución espectral de los datos analizados). El algoritmo DCPC es más lento y no trabaja muy bien con conjuntos de datos muy grandes, pero por otra parte no tiene restricciones en que la longitud del registro sea una potencia de 2.

Desde nuestro punto de vista ninguno de los dos métodos es completamente satisfactorio pero esto es probablemente debido no a un defecto en los métodos de segmentación sino de la suposición que los cambios en el patrón de olas ocurren abruptamente.

Es importante destacar que nuestras conclusiones están basadas solo en 3 conjuntos de datos. Se requerirá más investigación para validar nuestras conclusiones, especialmente en lo que tiene que ver con huracanes.

Capítulo 3

Una comparación de procedimientos de segmentación y análisis de la evolución de parámetros espectrales

3.1. Introducción

En este capítulo consideraremos la evolución de los espectros de energía de olas durante un período de un año con datos de una estación situada en Waimea Bay, Hawaii, EEUU. Usando los registros de alturas de olas calcularemos los espectros de energía cada 15 minutos de manera de determinar cambios en períodos pequeños en la evolución de algunas características de olas obtenidas a partir de los espectros. Se utilizó el software WAFO para calcular los espectros y las características espectrales.

Soukissian & Samalekos (2005)[32] han propuesto un método de segmentación para alturas significativas basado en la determinación de períodos de estabilidad, crecimiento y decrecimiento utilizando técnicas de series temporales. Su método consiste en hacer regresión lineal local donde los puntos iniciales y finales son los puntos extremos (máximos y mínimos locales) de la serie temporal y luego definiendo una función de costo para determinar la mejor configuración de los intervalos. Aplicaremos este procedimiento a otras características espectrales y compararemos los resultados con un método de segmentación diferente descrito a continuación.

El segundo método de segmentación está basado en valores medios sobre ventanas móviles y usando una banda con ancho fijo para determinar los puntos de cambio en el registro de olas. Los intervalos durante el

cual los valores se mantienen dentro de las bandas alrededor de la media se considerarán estables mientras que aquellos que estén por encima (o debajo) se considerarán crecientes (o decrecientes). De esta manera determinaremos los intervalos de estabilidad, crecimiento y decrecimiento de los registros. Ambos métodos fueron implementados en MATLAB.

Las características espectrales que consideraremos serán:

1. Alturas de olas significativas ($H_m = 4\sqrt{Var(X(t))}$),
2. Momento espectral de orden cero ($m_0 = \int_0^\infty S(\omega)d\omega$),
3. Momento espectral de orden dos ($m_2 = \int_0^\infty \omega^2 S(\omega)d\omega$) y
4. Períodos de pico espectral ($Tp = 2\pi\sqrt{m_0/m_2}$).

Al calcular las características espectrales se observaron picos que afectan el cálculo de las rupturas, por lo tanto suavizamos el registro de datos usando un filtro de promedio móvil finito de orden 5, para eliminar los ruidos locales.

Sea q un entero no negativo y consideramos el promedio móvil a ambos lados,

$$W_t = (2q+1)^{-1} \sum_{j=-q}^{q} X_{t-j}, \tag{3.1}$$

del proceso $\{X_t\}$ definido por $X_t = \bar{x}_t + Y_t, t = 1, \ldots, n$, donde $\mathbb{E}[Y_t] = 0$. Entonces para $q+1 \leq t \leq n-q$,

$$W_t = (2q+1)^{-1} \sum_{j=-q}^{q} m_{t-j} + (2q+1)^{-1} \sum_{j=-q}^{q} Y_{t-j} \simeq m_t, \tag{3.2}$$

suponiendo que $\bar{x}_t$ es aproximadamente lineal sobre el intervalo $[t-q, t+q]$ y que el promedio de los términos de error sobre este intervalo es cercano a cero.

El promedio móvil nos provee entonces con el siguiente estimador

$$\hat{x}_t = (2q+1)^{-1} \sum_{j=-q}^{q} X_{t-j}, q+1 \leq t \leq n-q. \tag{3.3}$$

Véase Brockwell y Davis (1996)[4] para más detalles.

La serie temporal analizada proviene de la Estación 10601 en Waimea Bay, Hawaii, EEUU, ya descrita. El resto del capítulo está organizado como sigue. En la siguiente sección describimos el algoritmo de Soukissian & Samalekos, y aplicamos este método a la serie temporal. Luego describiremos el algoritmo de bandas y sus aplicaciones a la serie temporal. En las siguientes dos secciones hacemos un análisis de los resultados obtenidos y damos las conclusiones.

3.2. Algoritmo de Soukissian

Consideremos una serie temporal de alturas de olas significativas $H_m = h_1, h_2, \ldots, h_n$ con n términos; el objetivo es hallar una k-segmentación de H_m, es decir, $H_m = H_{m_1}, H_{m_2}, \ldots, H_{m_k}$ con H_{m_i} intervalos disjuntos y no-superpuestos.

El primer paso en la segmentación de la serie temporal es definir un modelo de representación que aproxime los datos de cada segmento. Una vez hallado el modelo de representación, se evalúa la calidad de la aproximación con una función de costo que minimice el error de representación. Para esto usaremos un modelo de regresión lineal (Charbonnier (2005)[6]) y el error de representación lo definiremos basado en la suma de los cuadrados de las distancias entre los valores reales de la serie temporal y los valores fijados por el modelo.

El costo total de la k-segmentación es

$$COST = \sum_{i=1}^{k} cost(i,k),$$

donde *cost(i,k)*, $1 \leq i \leq k$ es el costo del i-ésimo segmento de la k-segmentación.

Se emplea el modelo de regresión lineal porque el registro de datos de H_m presenta tendencias de crecimiento y decrecimiento monótonos y este modelo se ajusta bien para detectar estos desarrollos.

El algoritmo inicia creando una segmentación fina de $n-1$ segmentos y n puntos de rupturas $[t_1, t_2]$, $[t_2, t_3]$, $\ldots$, $[t_{n-1}, t_n]$ a partir del registro completo. Esta primera partición se hace en base a los extremos locales (máximos y mínimos locales). Entonces estima un modelo de regresión lineal para cada segmento y calcula el error de representación en cada uno de ellos.

Suponiendo que $[s_i, e_i]$ donde s y e denotan los puntos inicial y final para el segmento i, los valores fijados del registro $h_{s_i}, h_{s_{i+1}}, \ldots, h_{e_{i-1}}, h_{e_i}$ se calculan con el modelo de regresión lineal descrito como sigue

$$h_i = \alpha_i + \beta_i t + \epsilon_t, \tag{3.4}$$

o

$$\hat{h}_i = a_i + b_i t, \tag{3.5}$$

para $s_i < t < e_i$ y $h_{s_i} < h_t < h_{e_i}$, donde t es el tiempo (variable independiente), h_t es la altura significativa (variable dependiente) en tiempo t, ϵ_t es el error aleatorio y a_i, b_i son los estimadores de α_i y β_i respectivamente. El parámetro a_i representa el corte vertical (valor en el origen) y b_i la pendiente (coeficiente de regresión) de la recta de regresión li-neal. Luego el algoritmo iterativamente concatena los pares de segmentos de menor costo hasta que el error de representación es menor que un máximo error definido por el usuario. Este proceso nos

da todos los intervalos crecientes y decrecientes. Para extraer los estados estacionarios del mar procedemos a aplicar el siguiente criterio, el cual para los intervalos crecientes es de la forma $h_{e_i} \leq h_{s_i}(100+p)\,\%$ y para los intervalos decrecientes es de la forma $h_{s_i} \leq h_{e_i}(100+p)\,\%$ (véase Soukissian y Samalekos (2006)[32];Labeyrie (1990)[16] y Athanasoulis et. al (1992)[2]). En las tablas 3.1 a 3.4 presentamos los resultados obtenidos para la Estación 106

	Creciente	Decreciente	Estacionario	Total
No. intervalos	337	368	446	1151
Mínimo (min)	15	15	15	15
1^{er} cuartil (min)	60	75	15	45
Mediana (min)	120	150	90	120
Promedio (min)	159,21	180,82	114,99	148,98
3^{er} cuartil (min)	220	240	165	210
Máximo (min)	720	900	675	900
Desv. estándar (min)	131,84	140,79	118,66	132,88

Tabla 3.1: Estadísticos para la Altura significativa (H_m), máx-error=0,015, p=4 % (min).

	Creciente	Decreciente	Estacionario	Total
No. intervalos	384	428	373	1185
Mínimo (min)	15	15	15	15
1^{er} cuartil (min)	45	71.25	15	45
Mediana (min)	105	135	45	105
Promedio (min)	153,20	175,20	100,98	144,71
3^{er} cuartil (min)	195	225	135	195
Máximo (min)	1020	1260	675	1260
Desv. estándar (min)	147,91	162,84	116,39	147,84

Tabla 3.2: Estadísticos para el momento espectral de orden cero (m_0), máx-error=0,001, p=6 % (min).

Note que para cada característica espectral elegimos un valor distinto para el máximo error y para p, porque la escala de cada característica espectral es diferente. La elección de estos valores es empírico, depende

	Creciente	Decreciente	Estacionario	General
No. intervalos	402	415	372	1189
Mínimo (min)	15	15	15	15
1^{er} cuartil (min)	60	60	15	45
Mediana (min)	120	135	45	105
Promedio (min)	157,09	174,90	96,09	144,22
3^{er} cuartil (min)	180	210	135	180
Máximo (min)	960	1275	810	1275
Desv. estándar (min)	151,30	181,96	117,11	157,03

Tabla 3.3: Estadísticos para el momento espectral de orden dos, (m_2), máx-error=0,0025, p=7 % (min).

	Creciente	Decreciente	Estacionario	General
No. intervalos	375	366	417	1158
Mínimo (min)	15	15	15	15
1^{er} cuartil (min)	60	75	15	45
Mediana (min)	120	135	105	120
Promedio (min)	158,68	166,52	122,37	148,08
3^{er} cuartil (min)	210	225	195	210
Máximo (min)	855	855	705	855
Desv. estándar (min)	128,77	125,26	118,73	125,54

Tabla 3.4: Estadísticos para los períodos de pico espectral (T_p),máx-error=0,07, p=3 % (min).

de la experiencia del investigador y de las características de la serie temporal. El rango para H_m es [0,8298; 5,3873], para m_0 es [0,0431; 1,82], para m_2 es [0,041; 2,1926] y para T_p es [3,4958; 12,2582]. También calculamos la pendiente de cada estado del mar y cada característica espectral. Los valores medios los presentamos en la tabla 3.5. Para dar una idea de las correspondientes distribuciones mostramos los boxplots de las mismas en las Figuras 3.9 a 3.12.

Característica Espectral	pendiente creciente	pendiente decreciente
Altura significativa, H_m	0,035	-0,030
Momento de orden cero, m_0	0,013	-0,012
Momento de orden dos, m_2	0,017	-0,016
Período de pico espectral, T_p	0,090	-0,079

Tabla 3.5: Valores medios de las pendientes de cada estado del mar para cada característica espectral.

3.3. Algoritmo de Bandas

Ahora describiremos el algoritmo de bandas. Este procedimiento de segmentación está basado en el cálculo de valores medios sobre ventanas móviles con un ancho de banda fijo. Iniciamos calculando la media para los dos primeros registros del archivo de datos y entonces añadimos sucesivamente nuevos puntos y recalculamos la media.

Sea m_i el valor medio de $X_1, \ldots, X_i$ y sea $2h$ el ancho de banda elegido. Si el siguiente punto X_{i+1} pertenece al intervalo $[m_i - h, m_i + h]$ recalculamos la media añadiendo el nuevo punto X_{i+1} para obtener m_{i+1} y el nuevo intervalo es $[m_{i+1} - h, m_{i+1} + h]$. Este proceso continúa hasta que el nuevo punto no pertenezca al intervalo, en cuyo caso es marcado como punto de corte. Los puntos previos forman un intervalo estacionario. El proceso inicia nuevamente. Si puntos sucesivos caen por encima (o debajo) del correspondiente ancho de banda, ellos determinan un intervalo creciente (o decreciente). El algoritmo es como sigue:

1. Lee los datos del archivo de característica espectral
2. Hacer $j = 1$, n=longitud del archivo de datos.
3. Mientras $j < n - 2$,
   - *a*) $m_i = media(p_i, \ldots, p_f)$, donde p_i es el punto inicial y p_f el punto final
   - *b*) Si $p_f + 1 \in [m_i - h, m_i + h]$ entonces $p_f = p_f + 1$;
   - *c*) sino
     1) Marcar p_f como punto de corte
     2) Hacer $p_i = p_f$ y $p_f = p_f + 1$;

```
    d) fin

4. Ir a 3
```

En las tablas 3.6 a 3.9 presentamos los resultados de la aplicación de este algoritmo a los datos de la Estación 106.

	Creciente	Decreciente	Estacionario	General
No. intervalos	792	848	1179	2819
Mínimo (min)	15	15	15	15
1^{er} cuartil (min)	30	45	30	30
Mediana (min)	45	45	45	45
Promedio (min)	55,63	56,76	67,26	60,84
3^{er} cuartil (min)	60	75	90	75
Máximo (min)	360	240	600	600
Desv. estándar (min)	35,40	29,43	66,01	49,63

Tabla 3.6: Estadísticos para la Altura significativa (H_m), ancho de banda = 0,125 (min).

	Creciente	Decreciente	Estacionario	General
No. intervalos	748	798	1054	2600
Mínimo (min)	15	15	15	15
1^{er} cuartil (min)	30	45	30	30
Mediana (min)	45	45	45	45
Promedio (min)	55,99	57,31	79,58	65,96
3^{er} cuartil (min)	63,75	75	90	75
Máximo (min)	375	240	1110	1110
Desv. estándar (min)	36,75	30,18	102,39	71,00

Tabla 3.7: Estadísticos para el momento espectral de orden cero (m_0), ancho de banda = 0,035 (min).

En este caso, como con el algoritmo de Soukissian y por la misma razón, el valor del parámetro de ancho de banda es diferente para cada característica espectral. El valor de ancho de banda h para cada característi-

	Creciente	Decreciente	Estacionario	General
No. intervalos	740	773	1009	2522
Mínimo (min)	15	15	15	15
1^{er} cuartil (min)	30	45	30	30
Mediana (min)	45	45	45	45
Promedio (min)	59,23	61,07	79,74	68,00
3^{er} cuartil (min)	75	75	90	75
Máximo (min)	300	300	1020	1020
Desv. estándar (min)	39,27	41,06	114,46	79,37

Tabla 3.8: Estadísticos para el momento espectral de orden dos (m_2), ancho de banda = 0,06 (min).

	Creciente	Decreciente	Estacionario	General
No. intervalos	809	751	1114	2674
Mínimo (min)	15	15	15	15
1^{er} cuartil (min)	30	45	30	30
Mediana (min)	45	45	45	45
Promedio (min)	56,68	57,82	73,80	64,13
3^{er} cuartil (min)	75	75	90	75
Máximo (min)	210	255	705	705
Desv. estándar (min)	31,88	32,67	73,10	53,83

Tabla 3.9: Estadísticos para el período de pico espectral (T_p), ancho de banda =0,28 (min).

ca espectral fue elegido después de muchas pruebas con diferentes valores de h y analizando el número de intervalos, duración de los intervalos, puntos iniciales y finales de estos, etc.

De nuevo calculamos las pendientes para cada estado y cada característica espectral, los valores medios para las pendientes crecientes y decrecientes los mostramos en la tabla 3.10. Los correspondientes boxplots están dados en las Figuras 3.9 a 3.12.

Característica Espectral	pendiente creciente	pendiente decreciente
Altura significativa, H_m	0,054	-0,051
Momento de orden cero, m_0	0,018	-0,017
Momento de orden dos, m_2	0,028	-0,026
Período de pico espectral, T_p	0,121	-0,119

Tabla 3.10: Valores medios de las pendientes de cada estado del mar para cada característica espectral.

3.4. Análisis de los resultados

En el análisis de los datos de olas con ambos algoritmos podemos ver que el número de puntos de corte así como la distribución de la longitud de los intervalos son diferentes. El algoritmo de bandas muestra más puntos de cortes que el algoritmo de Soukissian. Más aún, los puntos de corte no siempre coinciden unos con otros. En la siguiente Tabla (Tabla 3.11) damos el número total de puntos de corte y el número de puntos que coinciden. La cuarta columna de la Tabla 3.11 muestra los porcentajes de puntos de corte obtenidos con el algoritmo de Soukissian que coinciden con los obtenidos con el algoritmo de bandas. Como se puede observar la característica espectral con mayor porcentaje de coincidencia fue la Altura Significativa y en la Figura 3.1 se muestran tres segmentos de la serie temporal con los puntos de cortes de ambos algoritmos para la Altura Significativa.

	Algoritmo de Bandas	Algoritmo de Soukissian	Número de puntos que coinciden	Porcentaje
H_m	2819	1151	938	91,49 %
m_0	2600	1185	957	82,64 %
m_2	2522	1189	974	82,19 %
T_p	2674	1158	956	80,40 %

Tabla 3.11: Número de puntos de corte que tienen coincidencia.

Las tablas 3.12 y 3.13 muestran el número de puntos de cortes que coinciden entre dos diferentes características espectrales para cada algoritmo. El porcentaje que se muestra en la cuarta columna es con relación a la característica con mayor número de puntos de cortes para cada caso.

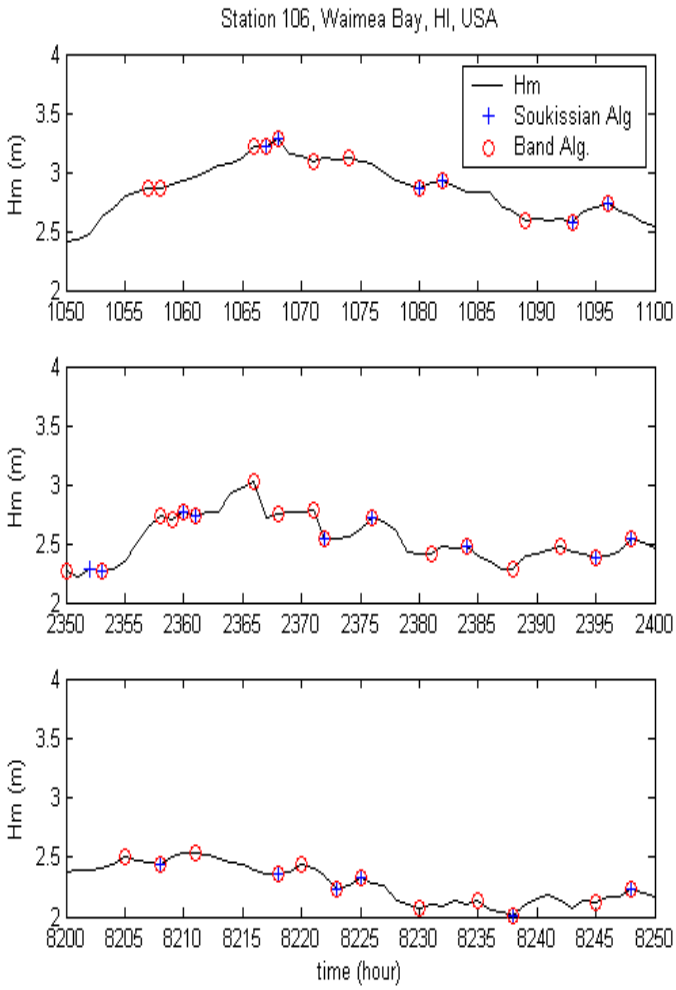

Figura 3.1: Tres segmentos de la altura significativa calculada de los datos de la Estación 106.

En los boxplots (Figuras 3.2 a 3.5) se pueden comparar la segmentaciones obtenidas con ambos algoritmos. Para el algoritmo de banda la duración de los intervalos estacionarios es mayor que la duración de los períodos de crecimiento y decrecimiento para todas las características espectrales consideradas, mientras que para el algoritmo de Soukissian es al contrario. Para el algoritmo de banda los períodos crecientes y decrecientes tienen similar distribución mientras que para el algoritmo de Soukissian los períodos de decrecimiento tienden a ser más largos que los períodos de crecimiento. Las distribuciones, sin embargo, varían con la característica

Característica Espectral	Número de puntos de cortes	Número de puntos que coinciden	Porcentaje
H_m	2819	2402	85,21 %
m_0	2600		
m_0	2600	799	29,88 %
T_p	2674		
m_2	2522	894	33,43 %
T_p	2674		
H_m	2819	1112	39,45 %
m_2	2522		
H_m	2819	882	31,29 %
T_p	2674		
m_0	2600	1607	61,81 %
m_2	2522		

Tabla 3.12: Una comparación de los resultados de segmentación para diferentes características espectrales obtenidos con el algoritmo de banda.

espectral considerada. Por otra parte, para ambos algoritmos el porcentaje del número de diferentes tipos de intervalos son similares para todas las características espectrales como se puede ver en la Tabla 3.14.

En las Figuras 3.6 a 3.8, una para cada estado, mostramos boxplots para todas las características espectrales para ambos algoritmos. En las Figuras 3.9 a 3.12, una para cada parámetro, mostramos boxplots para los valores absolutos de las pendientes de los intervalos crecientes y decrecientes obtenidos con ambos algoritmos. Como se puede ver de las Figuras 3.9 a 3.12, las pendientes difieren de un algoritmo a otro, pero las pendientes crecientes y decrecientes son muy similares para cada algoritmo. En todos los casos la distribución de las pendientes decrecientes parecen tener menos dispersión que la distribución para las pendientes crecientes.

Característica Espectral	Número de puntos de cortes	Número de puntos que coinciden	Porcentaje
H_m	1151	930	78,48 %
m_0	1185		
m_0	1185	198	16,71 %
T_p	1158		
m_2	1189	246	20,69 %
T_p	1158		
H_m	1151	282	23,72 %
m_2	1189		
H_m	1151	192	16,58 %
T_p	1158		
m_0	1185	283	23,80 %
m_2	1189		

Tabla 3.13: Una comparación de los resultados de segmentación para diferentes características espectrales obtenidos con el algoritmo de Soukissian.

Como se puede ver de las Tablas 3.1 y 3.6 y la Figura 3.8 la duración de los intervalos para el algoritmo de bandas es menor que la duración de los intervalos obtenidos con el algoritmo de Soukissian para todas las características espectrales. Como un ejemplo, de las Tablas 3.1 y 3.6 podemos ver que la duración media de los intervalos estacionarios para la altura significativa obtenidos con el algoritmo de bandas es 67,26min., mientras que para los obtenidos con el algoritmo de Soukissian es 114.99min. De hecho, la duración media para todos los intervalos obtenidos con el algoritmo de bandas es 60,84min. y de 148,98min. para los obtenidos con el algoritmo de Soukissian para la altura significativa.

En la Figura 3.13 podemos ver un intervalo donde la segmentación para la altura significativa es similar en el número de puntos de cortes, hay 12 para el algoritmo de Soukissian y 18 para el algoritmo de bandas. En estos intervalos se puede ver que ambos algoritmos detectan algunos puntos de cortes donde los datos alcanzan un máximo o un mínimo local, pero no en los mismos puntos. Mientras que el algoritmo de Soukissian sólo detecta puntos de cortes en los máximos y mínimos locales el algoritmo de bandas puede detectar puntos de cortes en cualquier punto (máximos locales, mínimos locales o cualquier otro punto).

Altura de olas significativas, H_m				
	Algoritmo de Soukissian		Algoritmo de Bandas	
	Núm. de Interv.	Porcentaje	Núm. de Interv.	Porcentaje
Creciente	337	29,28 %	792	28,10 %
Decreciente	368	31,97 %	848	30,08 %
Estacionario	446	38,75 %	1179	41,82 %
Momento Espectral de orden cero, m_0				
	Algoritmo de Soukissian		Algoritmo de Bandas	
	Núm. de Interv.	Porcentaje	Núm. de Interv.	Porcentaje
Creciente	384	32,41 %	748	28,77 %
Decreciente	428	36,12 %	798	30,69 %
Estacionario	373	31,48 %	1054	40,54 %
Momento Espectral de orden dos, m_2				
	Algoritmo de Soukissian		Algoritmo de Bandas	
	Núm. de Interv.	Porcentaje	Núm. de Interv.	Porcentaje
Creciente	402	33,81 %	740	29,34 %
Decreciente	415	34,90 %	773	30,65 %
Estacionario	372	31,29 %	1009	40,01 %
Periodo de pico espectral, T_p				
	Algoritmo de Soukissian		Algoritmo de Bandas	
	Núm. de Interv.	Porcentaje	Núm. de Interv.	Porcentaje
Creciente	375	32,38 %	809	30,25 %
Decreciente	366	31,61 %	751	28,09 %
Estacionario	417	36,01 %	1114	41,66 %

Tabla 3.14: Porcentajes del número de diferentes tipos de intervalos para las 4 características espectrales obtenidos con ambos algoritmos para la Estación 106.

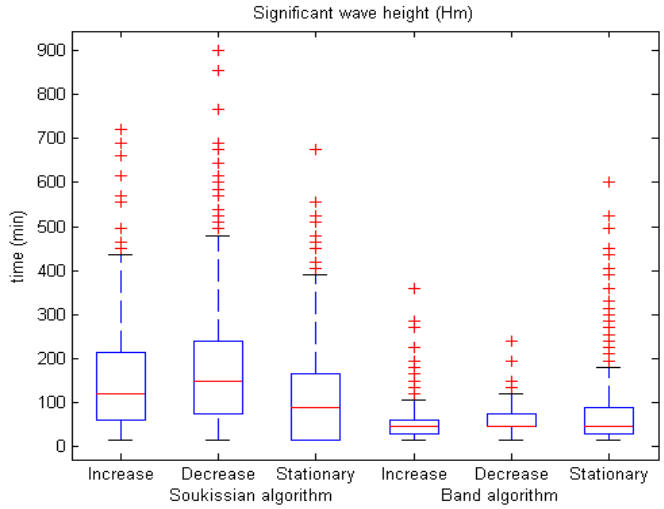

Figura 3.2: Boxplot para la segmentación de la Altura Significativa H_m obtenidas con el algoritmo de Soukissian (izquierda) y con el algoritmo de Bandas (derecha).

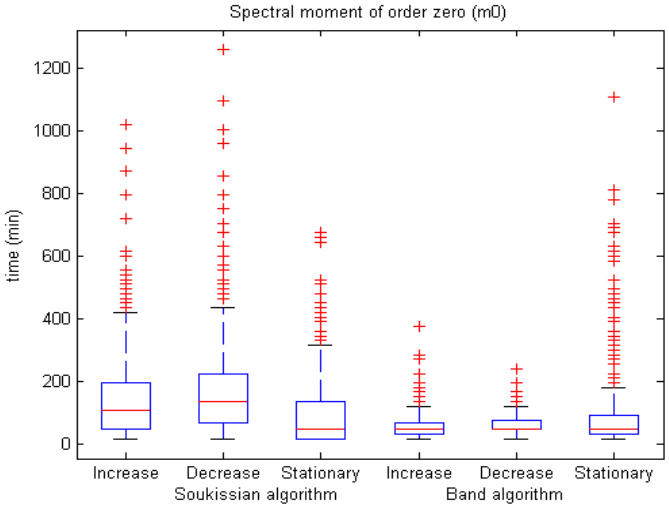

Figura 3.3: Boxplot para la segmentación del Momento Espectral de orden cero m_0 obtenidas con el algoritmo de Soukissian (izquierda) y con el algoritmo de Bandas (derecha).

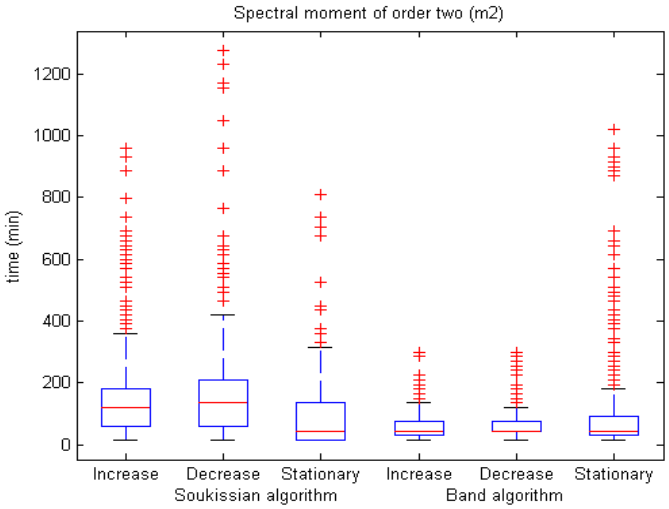

Figura 3.4: Boxplot para la segmentación del Momento Espectral de orden dos m_2 obtenidas con el algoritmo de Soukissian (izquierda) y con el algoritmo de Bandas (derecha).

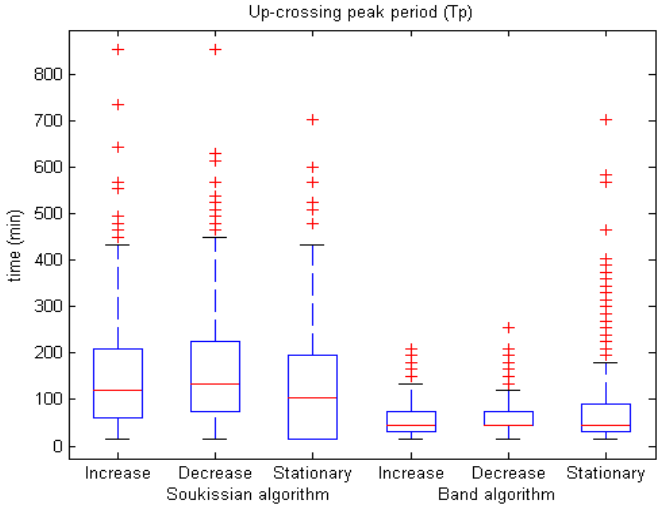

Figura 3.5: Boxplot para la segmentación del Periodo de Pico Espectral T_p obtenidas con el algoritmo de Soukissian (izquierda) y con el algoritmo de Bandas (derecha).

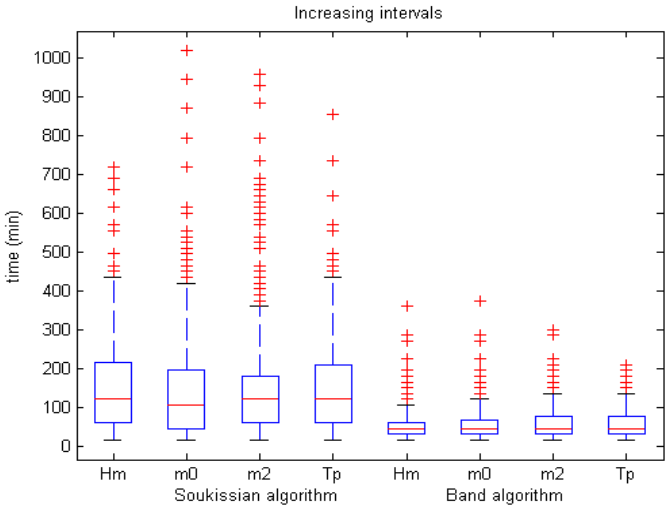

Figura 3.6: Boxplot para los intervalos crecientes para todas las características espectrales obtenidas con ambos algoritmos.

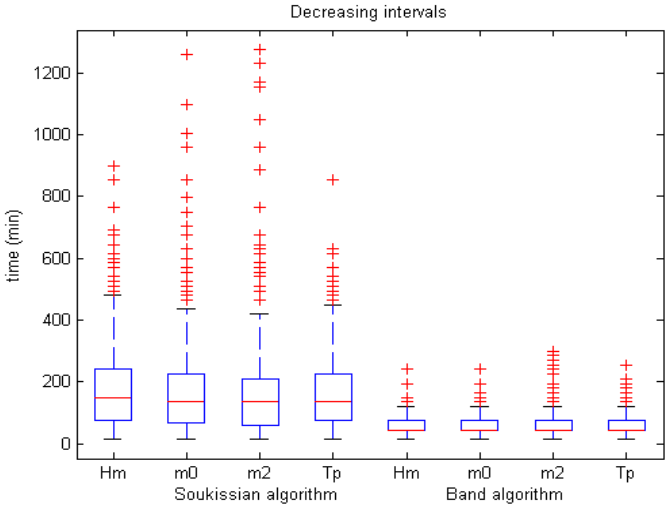

Figura 3.7: Boxplot para los intervalos decrecientes para todas las características espectrales obtenidas con ambos algoritmos.

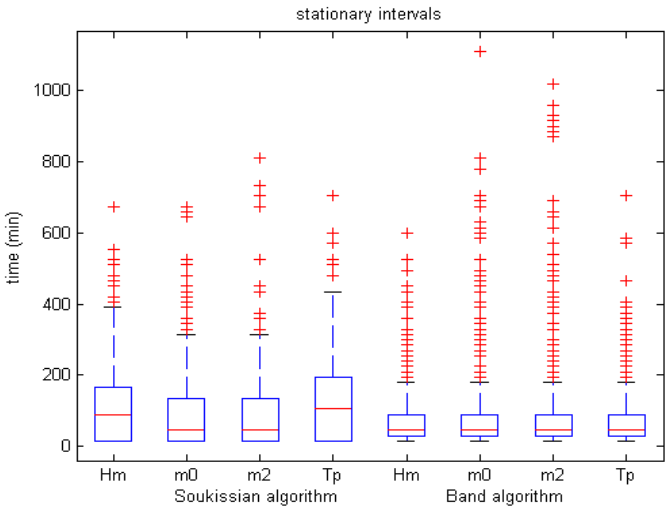

Figura 3.8: Boxplot para los intervalos estacionarios para todas las características espectrales obtenidas con ambos algoritmos.

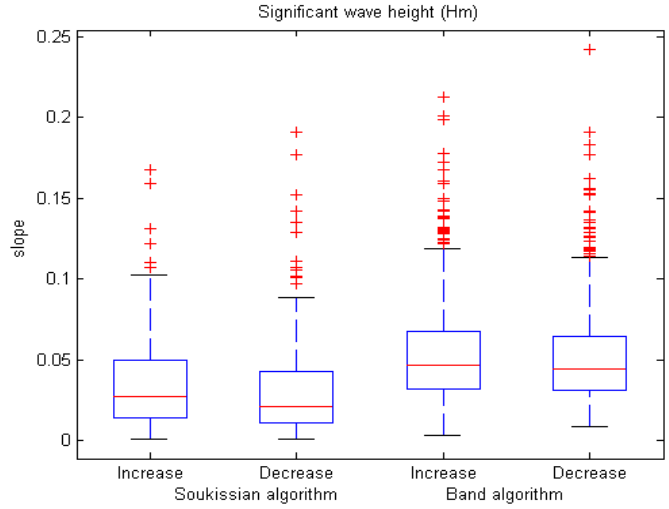

Figura 3.9: Boxplot para el valor absoluto de las pendientes de los intervalos crecientes y decrecientes de la altura significativa H_m obtenidas con el algoritmo de Soukissian (izquierda) y con el algoritmo de Bandas (derecha).

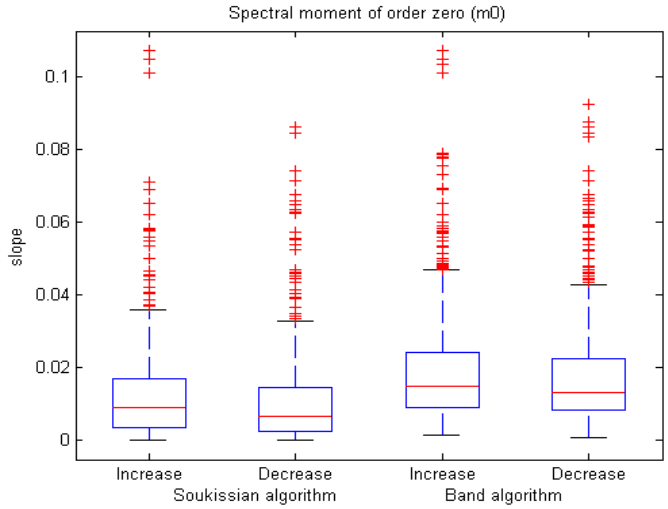

Figura 3.10: Boxplot para el valor absoluto de las pendientes de los intervalos crecientes y decrecientes del momento espectral de orden cero m_0 obtenidas con el algoritmo de Soukissian (izquierda) y con el algoritmo de Bandas (derecha).

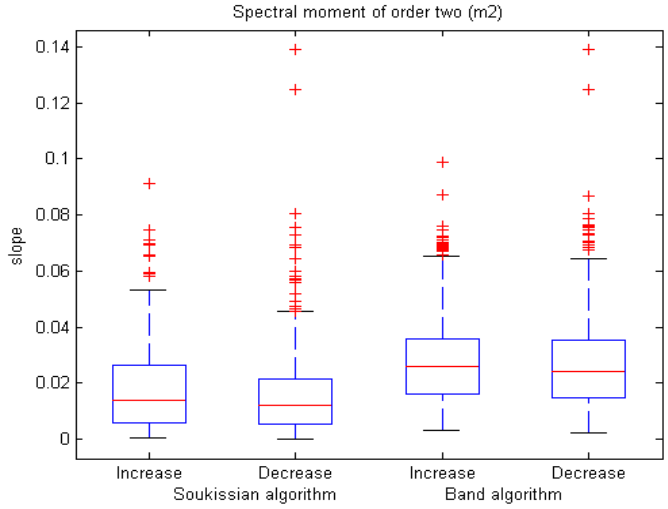

Figura 3.11: Boxplot para el valor absoluto de las pendientes de los intervalos crecientes y decrecientes del momento espectral de orden dos m_2 obtenidas con el algoritmo de Soukissian (izquierda) y con el algoritmo de Bandas (derecha).

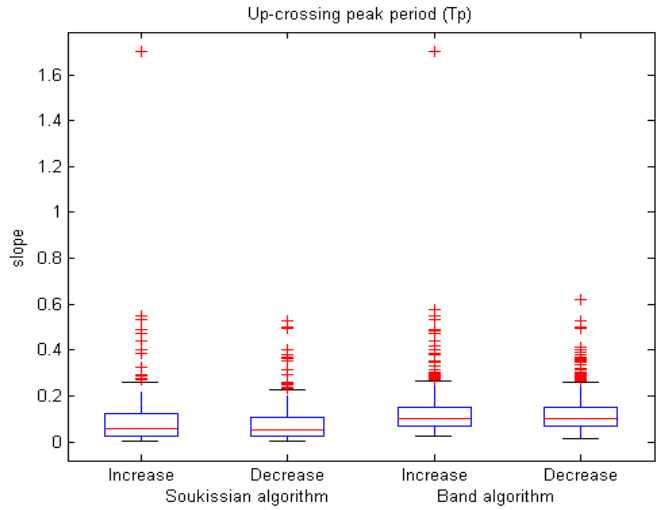

Figura 3.12: Boxplot para el valor absoluto de las pendientes de los intervalos crecientes y decrecientes del periodo de pico espectral T_p obtenidas con el algoritmo de Soukissian (izquierda) y con el algoritmo de Bandas (derecha).

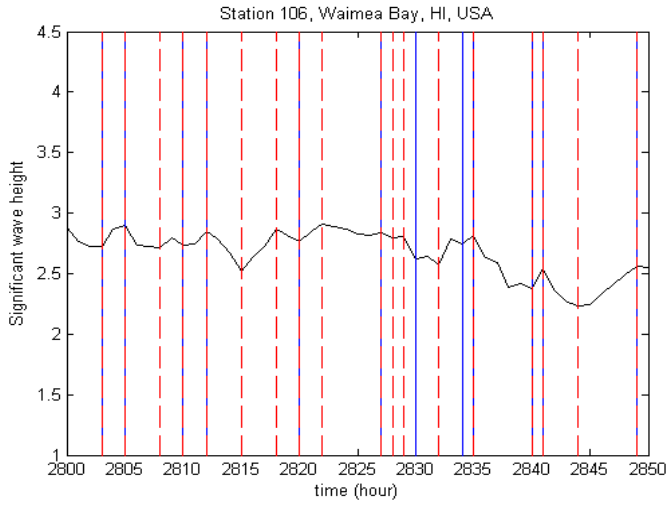

Figura 3.13: Segmentación para la altura significativa, Estación 106. La segmentación de Soukissian se muestra en azul (línea sólida), la segmentación de bandas en rojo (línea punteada).

3.5. Conclusiones

Hemos considerado dos procedimientos de segmentación para detectar puntos de cambio en una serie temporal: El algoritmo de Soukissian y el algoritmo de bandas. Estos algoritmos fueron usados para el conjunto de datos de las Estación 106 para cuatro características espectrales diferentes.

Los resultados fueron diferentes con relación al número de intervalos o puntos de cambios y la distribución de la duración de los intervalos. Los intervalos obtenidos con el algoritmo de Soukissian tiene una duración mayor que los obtenidos con el algoritmo de bandas para todas las características espectrales, pero la distribución de la duración es regular para ambos algoritmos.

Una desventaja del algoritmo de Soukissian es que toma más tiempo realizar la corrida para conjuntos de datos muy grandes. A pesar de esto, el algoritmo trabaja bien para detectar los puntos de cambio en un conjunto grande de datos. Por otra parte, el algoritmo de bandas es más rápido y fácil de usar, aunque tiende a conseguir muchos más intervalos que el algoritmo de Soukissian. No obstante para el algoritmo de bandas queda como futuro trabajo buscar maneras de fijar los parámetros de ancho de banda automáticamente para evitar la subjetividad de elección por parte del usuario.

Para diferentes características espectrales pero para el mismo algoritmo obtuvimos resultados similares con relación al número de intervalos o puntos de cortes para cada tipo de intervalo (crecientes, decrecientes y estacionarios). La altura significativa es la característica espectral más utilizada para determinar la segmentación. Una ventaja de usar la altura significativa es que esta nos puede dar una idea de la evolución del mar aún antes de calcular los puntos de cortes, y lo mismo vale para los períodos de pico espectral. Pero, en vista de lo analizado anteriormente, vemos que no existen diferencias significativas en la distribución como un todo (aunque habría que ver más en detalle las distribuciones), pero si hay diferencias en la ubicación de los cortes, como muestran las tablas 3.12 y 3.13.

Para ambos algoritmos los parámetros son fijados por el usuario, pero una vez hecho esto, el cálculo de los intervalos o puntos de cortes es automático, lo cual evita la subjetividad de elección de los intervalos. Estamos buscando maneras de fijar los parámetros automáticamente pero no hemos obtenido resultados satisfactorios.

Desde nuestro punto de vista ambos algoritmos trabajan bien para detectar los puntos de cambio en una serie temporal cuando el mar está en condiciones *normal*, es necesario estudiar el mar en presencia de condiciones extremas, como tormentas o huracanes, de manera de establecer si ellos trabajan bien o no bajo tales condiciones.

Capítulo 4

Estudio de los espectros de energía usando la Transformada de Hilbert-Huang para la segmentación de tormentas dado por el algoritmo SLEX

4.1. Introducción

En este capítulo vamos a utilizar el algoritmo SLEX (Smooth Localized complex EXponential) propuesto por Ombao et al. (2002) [[22]] para detectar los cambios en los datos de tormentas que describiremos luego. También haremos uso del algoritmo HHT (Hilbert-Huang Transform) para calcular los espectros marginales de cada segmento estacionario detectado con SLEX. El primer método ha sido ideado para el estudio de series temporales y conseguir segmentación de las mismas en intervalos estacionarios, y ya lo hemos descrito en el capítulo 2, sección 2.3; el segundo método fue ideado para el estudio de ondas no estacionarias y no lineales, descomponiendo la señal en una serie de componentes a distintas frecuencias y tiempos. Ya que las olas marinas pueden verse como una serie temporal y al mismo tiempo presentan las complicaciones de no estacionaridad y no linealidad, será útil el estudio de los datos de tormentas con ambos algoritmos. Utilizaremos el algoritmo SLEX para determinar los estados estacionarios de las mismas, una vez obtenido los cortes o

rupturas haremos uso tanto del algoritmo SLEX como HHT para el análisis espectral de la segmentación, con el objetivo de comparar ambos métodos. La aplicación de ambos métodos (SLEX y HHT) es independiente una de otra.

Los datos que utilizaremos son los de la tormenta del Mar del Norte descritos en la introducción. Dado que existen algunos intervalos cortos de tiempo en los cuales hay registros perdidos, dividimos esta serie en 5 conjuntos que cubren la tormenta. En la Tabla 4.1 damos una lista de los 5 conjuntos junto con algunas características básicas de olas: Altura de ola significativa, H_m; período medio de olas, T_{m01}; período de pico espectral, T_p; y parámetro de ancho de banda espectral, ν. En la Figura 4.1 mostramos la evolución de la altura de ola significativa durante la tormenta.

El capítulo está estructurado de la siguiente manera: En la sección 4.2 describimos el algoritmo HHT, primero describimos el proceso de Descomposición en Modos Empíricos (EMD por sus siglas en inglés), luego desarrollamos brevemente el Análisis Espectral de Hilbert. En la sección 4.3 analizaremos los datos de la tormenta; primero realizaremos la segmentación utilizando el algoritmo SLEX, segundo estudiaremos los espectros de cada segmentación usando los algoritmos SLEX y HHT y luego haremos un estudio de eventos importantes, esto es, un análisis de los espectros de energía para olas con altura mayor a 6.5m. Finalmente presentamos las conclusiones.

	Duración	H_m	T_{m01}	T_p	ν
Storm1999a	8h. 40m	5.34	9.31	11.87	0.509
Storm1999b	6h	3.72	8.48	11.22	0.514
Storm1999c	18h	5.07	8.25	10.50	0.510
Storm1999d	24h	5.87	8.99	11.70	0.492
Storm1999e	24h	5.10	8.75	11.70	0.506

Tabla 4.1: Características básicas de los 5 intervalos de datos.

4.2. Análisis HHT

4.2.1. El método de descomposición en modos empíricos

La transformada de Hilbert-Huang (HHT) fue propuesta por Huang et. al (1998, 1999, 2003) como un método adecuado para el análisis espectral de procesos no-estacionarios y no-lineales. El algoritmo HHT consiste

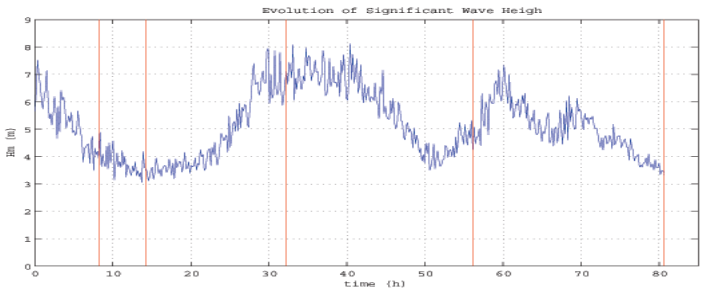

Figura 4.1: Evolución de la altura de ola significativa.

en dos partes: Un algoritmo empírico llamado Descomposición de Modo Empírico (EMD por sus siglas en inglés), usado para descomponer una serie temporal en oscilaciones características individuales conocidas como funciones de modo intrínsecas (IMF, siglas en inglés) y el análisis espectral de Hilbert (HSA, siglas en inglés). El algoritmo EMD está basado en la suposición de que cada señal consiste de diferentes modos de oscilación basados en diferentes escalas de tiempo, de modo que cada IMF representa uno de estos modos oscilatorios. Cada IMF debe satisfacer dos criterios:

1. El número de extremos locales y cruces de ceros debe ser igual o diferir a lo sumo en uno.

2. En cada instante, la media de las envolventes definidas por los máximos locales y las correspondientes a los mínimos locales debe ser cero.

Estas dos condiciones son requeridas para evitar inconsistencias en la definición de las frecuencias instantáneas. Una IMF representa un modo oscilatorio simple que es homóloga a una función armónica simple, pero más general: En vez de una amplitud y una frecuencia constantes, como en un componente armónico simple, la IMF puede tener amplitud y frecuencias variables como función del tiempo. Una vez que la señal es descompuesta, se aplica la Transformada de Hilbert a cada IMF. La Transformada de Hilbert $y(t)$ de una función $x(t)$ se define como $(1/\pi)$ veces la convolución de $x(t)$ con la función $1/t$.

$$y(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1}{t-s} x(s) ds \tag{4.1}$$

donde la integral se toma como el valor principal de Cauchy. Entonces, si $z(t)$ es la señal analítica asociada a $x(t)$, tenemos para todo t,

$$z(t) = x(t) + iy(t) = A(t) \exp(i\theta(t)) \tag{4.2}$$

con $A(t) = \sqrt{x^2(t) + y^2(t)}$ y $\theta(t) = \arctan(y(t)/x(t))$.
La frecuencia instantánea se define ahora como la derivada de la función de fase de la señal analítica $z(t)$:

$$\omega(t) = \frac{d\theta(t)}{dt} \tag{4.3}$$

Una vez que la señal ha sido descompuesta en IMF's y se ha obtenido la Transformada de Hilbert para cada una, la señal $x(t)$ se puede representar como

$$x(t) = \Re \sum_{j=1}^{n} A_j(t) \exp\left(i \int \omega_j(t) dt\right) \tag{4.4}$$

la cual es una forma generalizada del desarrollo de Fourier para $x(t)$ en la cual tanto la amplitud como la frecuencia son funciones del tiempo. La distribución tiempo-frecuencia de la amplitud o de la amplitud al cuadrado es definida como el espectro de amplitud de Hilbert o el espectro de energía de Hilbert respectivamente. Para más detalles véase Huang y Shen (2005) [13].

4.2.2. Análisis espectral de Hilbert

Después de realizar la transformada de Hilbert sobre cada componente IMF, los datos originales se pueden expresar como la parte real $\mathbb{R}$ de la siguiente forma:

$$x(t) = \mathbb{R} \left\{ \sum_{j=1}^{n} A_j(t) \exp\left[i \int \omega_j(t) dt\right] \right\}. \tag{4.5}$$

Aquí, el residual r_n ha sido quitado a propósito, porque éste es una función monótona o una constante. Aunque la transformada de Hilbert puede tratar la tendencia monótona como parte de una oscilación más grande; la energía envuelta en el residual que representa una medida de compensación podría dominar. En consideración a la incertidumbre de la tendencia de oscilación más grande, y en el interés de obtener la información contenida en otras energías bajas pero componentes claramente oscilatorias, el componente de no IMF final deberá ser excluido. Sin embargo, podría se incluido si consideraciones físicas justifican su inclusión.

La ecuación (4.5) nos da tanto la amplitud como la frecuencia de cada componente como función del tiempo. Los mismos datos desarrollados en la representación de Fourier serían

$$x(t) = \mathbb{R} \left[\sum_{j=1}^{n} a_j e^{i\omega_j(t)t} \right], \tag{4.6}$$

con a_j y ω_j constantes. El contraste entre (4.5) y (4.6) es claro: las IMF's representan un desarrollo generalizado de Fourier. Las variables amplitud y la frecuencia instantánea no sólo suministran gran mejoramiento en la eficiencia del desarrollo, sino también permiten que el desarrollo se ajuste a datos no-lineales y no-estacionarios. Con el desarrollo de las IMF, la amplitud y la modulación de frecuencia están claramente separadas. Por lo tanto, la restricción de que la amplitud y la frecuencia sean constantes en el desarrollo de Fourier han sido sustituidos con una variable amplitud y una representación de frecuencia. Esta distribución tiempo-frecuencia de la amplitud es designada como el *Espectro de amplitud de Hilbert*, y denotado $H(\omega, t)$, o simplemente *Espectro de Hilbert*.

Con el espectro de Hilbert ya definido, podemos también definir el espectro marginal $h(\omega)$ como

$$h(\omega) = \int_0^T H(\omega, t)dt. \tag{4.7}$$

El espectro marginal ofrece una medida de la contribución de cada valor de frecuencia a la amplitud (o energía) total. Este espectro representa el acumulado sobre todo el conjunto de datos en un sentido probabilístico.

4.3. Análisis de los datos

4.3.1. Segmentación con SLEX

Utilizando el algoritmo SLEX para cada uno de los conjuntos de datos, obtuvimos los cortes en cada uno. Para ello usamos diversos valores de los parámetros hasta obtener la mejor segmentación de cada uno de los conjuntos de datos: Valores para el parámetro de *Penalización*: 2; 1.5; 1; 0.5; 0.25; 0.125; 0.1; 0.05 y 0.025. Parámetro de *suavizado*: 0.1; 0.02; 0.015; 0.01; 0.005 y 0. Finalmente escogimos los valores de *Penalización 0.25* y *Suavizado* 0, dado que la segmentación obtenida con estos valores resultó la más aceptable; es decir, se obtuvieron intervalos con una duración media entre 16 y 20 minutos aproximadamente, como se puede ver en la Tabla 4.3. Para los otros valores de prueba, el número de intervalo resultaba muy grande con una duración media de aproximadamente 3 min 30 seg., lo cual no parece razonable, aún con el mar agitado debido a la presencia de una tormenta fuerte como la que estamos analizando.

Como el algoritmo SLEX funciona con archivos de longitud una potencia de 2, estudiamos dos situaciones: 1) Truncamos cada uno de los conjuntos de datos en la mayor potencia de 2 cuyo valor fuera menor o igual que la longitud de los archivos y 2) Completamos con ceros al final de cada conjunto de datos a la menor potencia

Archivo	**Duración**	**Longitud** (No. registros)	**Archivo Truncado**	**Archivo Completado**
Storm1999a	8h.40m	156000	$2^{17} = 131072$	$2^{18} = 262144$
Storm1999b	6h	108000	$2^{16} = 65536$	$2^{17} = 131072$
Storm1999c	18h	324000	$2^{18} = 262144$	$2^{19} = 524288$
Storm1999d	24h	432000	$2^{18} = 262144$	$2^{19} = 524288$
Storm1999e	24h	432000	$2^{18} = 262144$	$2^{19} = 524288$

Tabla 4.2: Longitud de cada uno de los conjuntos de datos.

de 2 tal que fuera mayor o igual a la longitud del archivo. En la Tabla 4.2 se muestran las longitudes de los conjuntos de datos así como las potencias correspondientes a los datos truncados y completados.

Al realizar ambas segmentaciones podemos observar que en los intervalos comunes, la segmentación es la misma. En la figura 4.2 podemos observar la segmentación obtenida con SLEX, en la parte superior mostramos los cortes con la completación de ceros del archivo y en la parte inferior con los datos truncados para el archivo *Storm1999c*. Para los archivos *Storm1999b, Storm1999d* y *Storm1999e* aún cuando la segmentación es la misma en los intervalos comunes, se perdían casi el 40 % de los datos en los archivos truncados, por lo que después de realizar el análisis con cada conjunto de datos los mejores resultados los obtuvimos realizando la completación a la siguiente potencia de dos. El máximo nivel para los periodogramas de estos archivos fue: Storm1999a=7; Storm1999b=7; Storm1999c=9; Storm1999d=9 y Storm1999e=9. La razón de ellos es para que el tamaño de los bloques no sea menor a 1024 registros que representan una duración de 3.41min, ya que intuitivamente no parece lógico que los estados del mar cambien en tan poco tiempo.

En el borde de los archivos, es decir, donde termina el registro original y comienza la completación de ceros se observan algunos cortes que tienen intervalos de duración real menor a 3.41seg. (1024 datos), como se observa en la figura 4.3, por lo tanto para evitar esta situación, el último corte a considerar es el previo al final del archivo como se muestra en la figura 4.3.

Para el primer conjunto de datos *Storm1999a*, obtuvimos 30 puntos de corte (29 segmentos, están incluidos los extremos); para el segundo conjunto *Storm1999b*, 24 puntos de cortes (23 segmentos); para el tercer conjunto *Storm1999c*, 60 puntos de cortes (59 segmentos); para el cuarto conjunto *Storm1999d*, 72 puntos de cortes (71 segmentos) y para el último conjunto *Storm1999e*, 72 puntos de cortes (71 segmentos). En la Tabla 4.3 mostramos algunos estadísticos básicos para la duración o longitud de los intervalos de la segmentación de cada conjunto de datos.

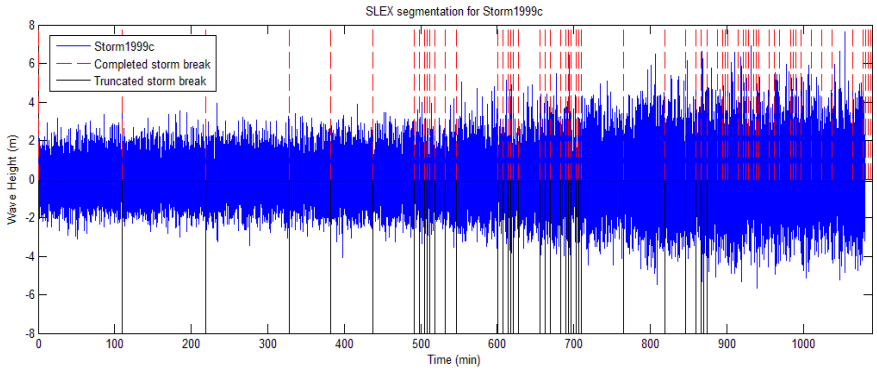

Figura 4.2: Segmentación del archivo *Storm1999c* obtenida con SLEX.

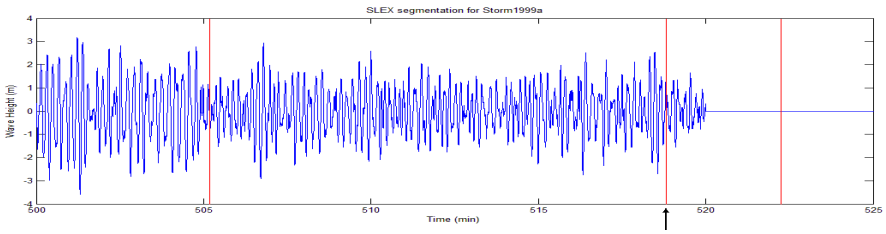

Figura 4.3: Segmentación en el borde del archivo *storm1999a*.

En la Figura 4.4 podemos observar un boxplot para las longitudes de los intervalos en la segmentación de cada conjunto de datos.

Archivo	storm1999a	storm1999b	storm1999c	storm1999d	storm1999e
Mínimo	00:03:04.04	00:03:04.04	00:03:04.04	00:03:04.04	00:03:04.04
1er. cuartil	00:06:08.08	00:06:08.08	00:03:04.04	00:06:08.08	00:06:08.08
Promedio	00:18:17.17	00:15:52.52	00:18:09.09	00:20:09.09	00:20:10.10
3er. cuartil	00:27:32.32	00:13:16.16	00:13:16.16	00:27:32.32	00:27:32.32
Máximo	00:54:04.04	00:54:04.04	01:49:08.08	01:49:08.08	01:49:08.08
Varianza	00:00:08.08	00:00:08.08	00:00:29.29	00:00:25.25	00:00:26.26

Tabla 4.3: Estadísticos básicos para la longitud de los intervalos (min), Tormenta 1999.

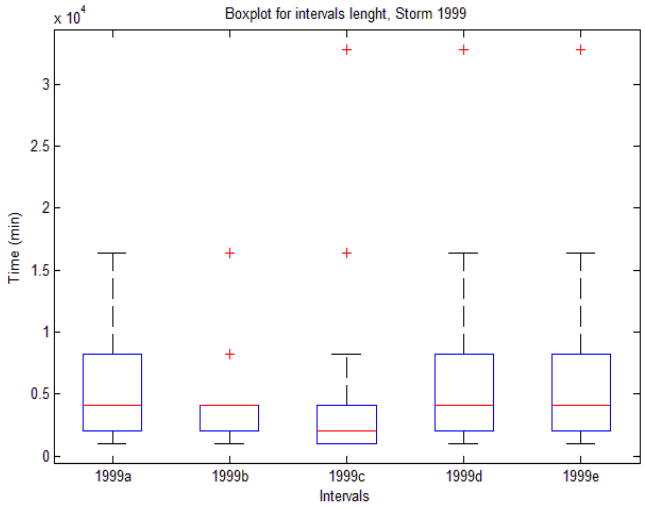

Figura 4.4: Boxplot para la longitud de los segmentos obtenidos con SLEX.

4.3.2. Análisis de los espectros con SLEX y HHT

En esta sección realizaremos un análisis de los espectros de energía con los algoritmos SLEX y HHT, cabe destacar que la aplicación de estos algoritmos es independiente uno del otro; nuestro objetivo aquí es ver qué aporta cada algoritmo al análisis espectral del registro de olas. Realizamos el análisis de los datos de la tormenta con HHT. Los 5 archivos de datos fueron descompuestos en Funciones de Modo Intrínseco (IMFs) usando el proceso de Descomposición en Modos Empíricos. El software usado fue HHT-DPS, desarrollado por la NASA. El número de IMFs obtenidos varía con cada conjunto de datos y parece incrementarse con

la longitud del mismo, como se puede ver en la Tabla 4.4. Esto probablemente se deba al hecho de que en intervalos más grandes es posible detectar frecuencias más bajas en el archivo de datos. Calculamos el espectro marginal para cada intervalo obtenido con el algoritmo SLEX de cada conjunto de datos.

Archivo	Duración	Núm. IMF's
Storm1999a	8h.40min.	17
Storm1999b	6h.	15
Storm1999c	18h.	19
Storm1999d	24h.	21
Storm1999e	24h.	20

Tabla 4.4: Número de IMFs para cada conjunto de datos.

Para cada conjunto de datos calculamos los espectros marginales de Hilbert, los espectros con el software WAFO para cada segmento obtenido con SLEX, así como los máximos de energía y frecuencia dominante. El algoritmo SLEX también calcula el espectrograma SLEX de energías para el archivo de datos, usando la función **slexgram()**; en la Figura 4.6 observamos el espectrograma SLEX para la segmentación del archivo *Storm1999a*. En la Figura 4.5 mostramos los boxplots de los máximos de energías calculados con el algoritmo HHT de cada intervalo de la segmentación dada por SLEX .

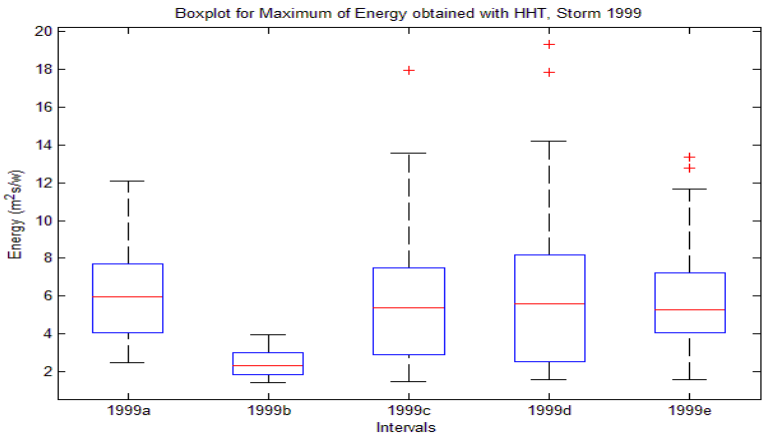

Figura 4.5: Boxplot de los máximos de energías de cada segmento calculada con HHT para la tormenta 1999.

En los puntos de cortes dados por SLEX podemos observar variaciones de la distribución marginal calcu-

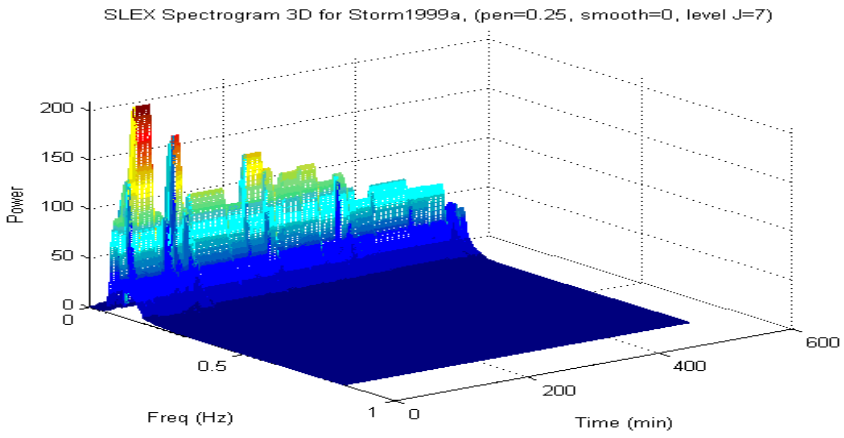

Figura 4.6: Espectrograma SLEX de energías de la segmentación de Storm1999a.

lada con HHT, por ejemplo para el archivo *Storm1999a*, construimos los espectros marginales de Hilbert en un entorno de los puntos de cortes, tomamos 300 registros de cada lado del punto de corte, lo que corresponde a 60 segundos (1min.) de duración, es decir un entorno de 2min. de duración. En las gráficas de los espectros de energía podemos notar que en la mayoría de ellos hay cambios en la energía o distribución de energía; para esta serie tenemos 30 puntos de corte, incluidos los extremos; dado que queremos ver si hay cambio o no en la distribución de energía entre cada segmento, excluimos del estudio los puntos extremos, lo que nos deja 28 puntos de corte. Al realizar el estudio, se puede ver que en 8 de ellos la distribución de energía se mantiene bastante similar, véase la Figura 4.7 donde mostramos la distribución de energía en un entorno del corte 4 dado por SLEX. En el resto observamos cambios significativos en la distribución de la energía, en 8 de los cortes la energía crece (Figura 4.8), en los restantes 12 cortes la energía disminuye (Figura 4.9). Si observamos el espectrograma SLEX del archivo *Storm1999a* podemos notar que efectivamente la energía cambia entre los puntos de corte (véase la Figura 4.6). Una ventaja de realizar el espectro marginal de Hilbert es que la resolución es muy fina y se pueden observar con más detalle los cambios en la distribución de energía de lo que se puede observar con el algoritmo SLEX.

4.3.3. Eventos importantes

Ahora vamos a presentar un análisis de los espectros de energía y su evolución para eventos importantes, esto es, para olas con altura mayor a 6,5m. En la Tabla 4.5 mostramos la ubicación de estos eventos, así como

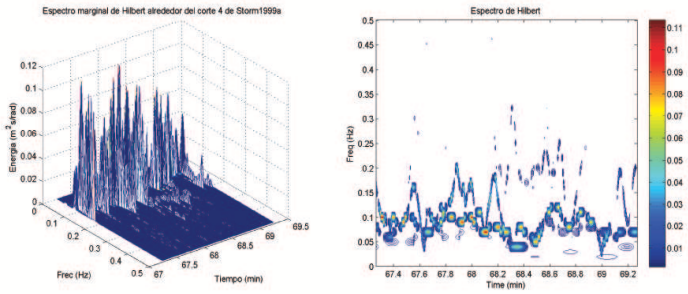

Figura 4.7: Espectro marginal de Hilbert de un entorno del corte 4 de *Storm1999a* dado por SLEX

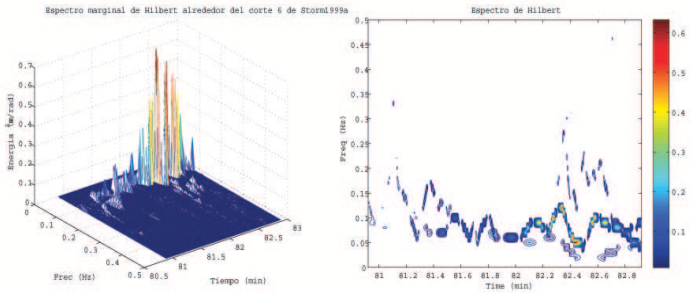

Figura 4.8: Espectro marginal de Hilbert de un entorno del corte 6 de *Storm1999a* dado por SLEX

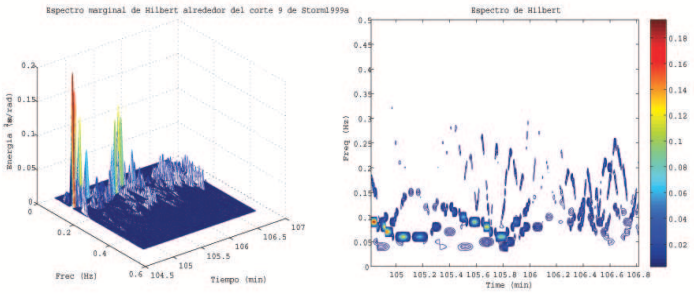

Figura 4.9: Espectro marginal de Hilbert de un entorno del corte 9 de *Storm1999a* dado por SLEX

sus alturas y la energía acumulada en un entorno de las mismas, este entorno cubre 150 registros, tomamos este valor para tener un tiempo de 30 segundos; 15 segundos a cada lado de la ola, para cubrir un entorno de ellas, ya que el período medio de las olas es de aproximadamente 11 segundos.

Tormenta		Registro	Tiempo (min)	Altura (m)	Energía (HHT)
1999a	Ola 1	24705	82.3	6.9294	45.3620
	Ola 2	28562	95.2	6.6129	33.5807
1999c	Ola 1	260365	867.9	6,6237	43.7456
	Ola 2	273146	910.5	6.5699	29.5411
	Ola 3	279587	932.0	6.9384	57.6346
	Ola 4	316510	1055.0	7.6621	44.2784
1999d	Ola 1	22885	76.3	6.9140	17.7316
	Ola 2	23765	79.2	6.8757	33.5376
	Ola 3	24653	82.2	7.2738	24.4441
	Ola 4	70298	234.3	6.6849	36.3667
	Ola 5	74329	247.8	7.2624	33.1469
	Ola 6	77015	256.7	8.5906	49.7123
	Ola 7	89891	299.6	6.8856	33.4308
	Ola 8	92022	306.7	6.9990	39.1099
	Ola 9	171602	572.0	6.7243	29.7569
1999e	Ola 1	51695	172.3	6.6862	38.8597

Tabla 4.5: Eventos importantes, (Olas con alturas $>$ 6,5m.)

En las siguientes Figuras (Figs. 4.10 a 4.14) mostramos, los espectrogramas bidimensionales y tridimensionales de Hilbert para algunos eventos importantes. En la Figura 4.10 (parte superior) podemos observar el espectro de Hilbert 3D del séptimo intervalo de segmentación de *Storm1999a* obtenida con SLEX el cual muestra la evolución de energía para el intervalo. Podemos observar grandes cantidades de energía cerca de los extremos del intervalo localizados aproximadamente en 82 y 95min., que alcanzan una energía de $0{,}6m^2s/rad$ y $0{,}75m^2s/rad$ respectivamente. Estos dos eventos corresponden a las Olas Grandes 1 y 2, y son mostradas en la parte inferior de la Figura 4.11. La energía total acumulada en un entorno de 30 segundos de la Ola Grande 1 es $45{,}3620m^2s/rad$ mientras que para la Ola Grande 2 es de $33{,}5807m^2s/rad$.

La Figura 4.11 muestra un gráfico de contorno en colores para el espectro de Hilbert para las Olas Grandes 1 y 2 de *Storm1999a* en la parte superior y las correspondientes olas en la parte inferior con la misma escala de tiempo. La figura para la Ola Grande 1 muestra que hay una gran cantidad de energía alrededor de la

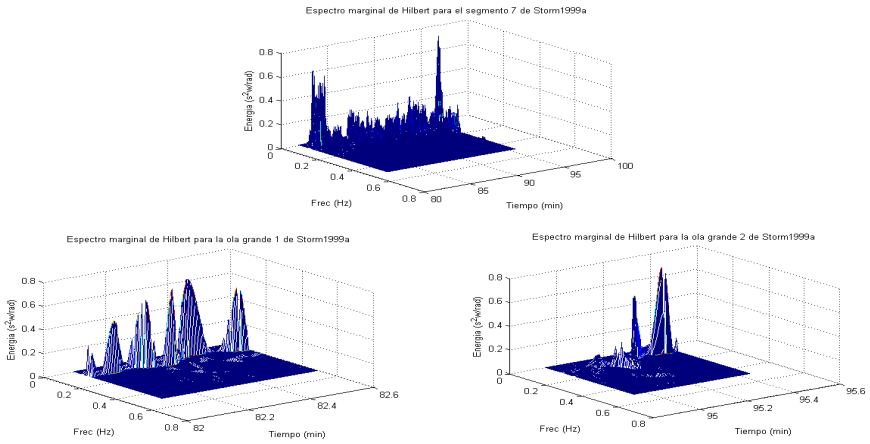

Figura 4.10: Espectro de Hilbert para el intervalo 7 de *Storm1999a* (parte superior). Espectros de Hilbert para las olas grandes 1 y 2 (parte inferior).

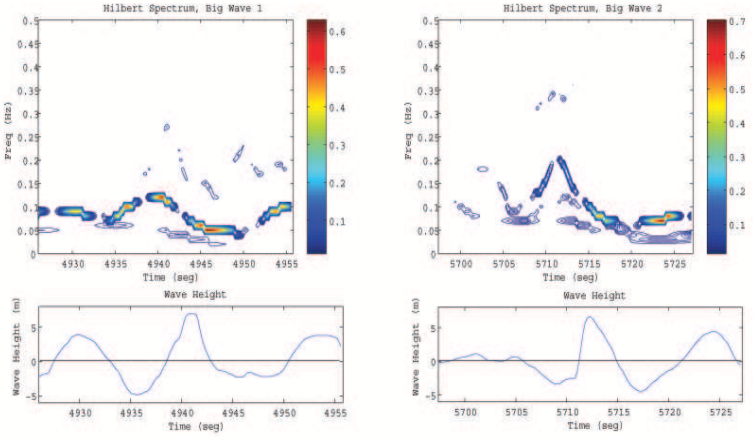

Figura 4.11: Gráfico de contornos de los espectros marginales de Hilbert para las Olas Grandes 1 y 2 de *Storm1999a*

misma. En contraste, el gráfico de la Ola Grande 2 muestra una gran cantidad de energía *después* de la ola, en el intervalo de tiempo entre 5720 y 5725 segundos.

En la Figura 4.12 mostramos los espectros del Hilbert del segundo intervalo de segmentación de *Storm1999d* (parte superior), y las Olas Grandes 1 (parte inferior izquierda) y Ola Grande 2 (parte inferior derecha). En el

espectro de Hilbert del segmento podemos ver la evolución de la energía de las olas; las Olas Grandes 1 y 2 están alrededor de 76 y 79min. respectivamente (Véase la Tabla 4.5), esto es hacia el extremo derecho del intervalo; y en la parte inferior podemos notar la evolución de energía en los entornos de estas olas. La energía total acumulada de la Ola Grande 1 es 17,7316m^2s/rad y la de la Ola Grande 2 es 33,5376m^2s/rad.

En la figura 4.13 mostramos una gráfica de contorno del espectro de Hilbert para el intervalo 2 de la segmentación de *Storm1999d* (parte superior), que cubre 27.3min. En la parte inferior hay un gráfico de alturas de olas para el mismo intervalo usando la misma escala de tiempo. Hay dos olas grandes en este intervalo, la primera ocurre a los 76.3min. y la segunda a los 79.2min. Para la Ola Grande 1 observamos que la amplitud de frecuencias está distribuida entre 0.05 y 0.2Hz en un entorno de 30seg. de la ola. Para la Ola Grande 2 la frecuencia se concentra alrededor de 0.05Hz. En este caso la distribución de frecuencias se mantiene entre 0.05 y 1Hz. Ambas olas son el resultado de la superposición de distintas Funciones de Modo Intrínseco, ninguna de las cuales tiene una cantidad grande de energía asociada (véase Figura 4.14). Este fenómeno fue también observado y comentado en [[28]].

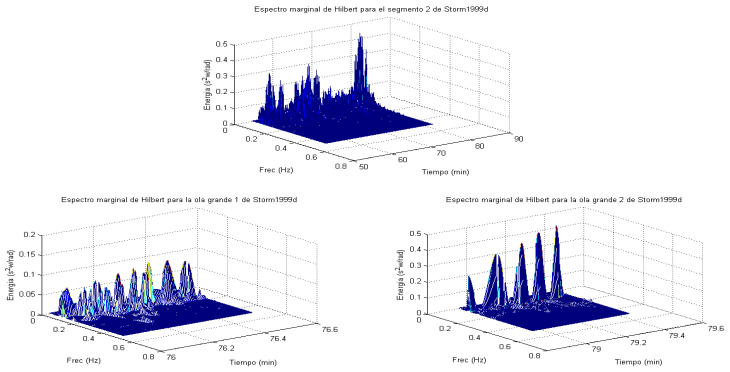

Figura 4.12: Espectro de Hilbert 3-D para el intervalo 2 de *Storm1999d* donde se muestran las Olas Grandes 1 y 2.

4.4. Conclusiones

Hemos realizado un análisis de segmentación y espectral para una serie de datos provenientes de una tormenta en el Mar del Norte en el año 1999, usando los algoritmos SLEX y la Transformada de Hilbert-Huang; la tormenta fue descompuesta en 5 conjuntos de datos con duraciones de 8 a 24 horas como mostramos en la

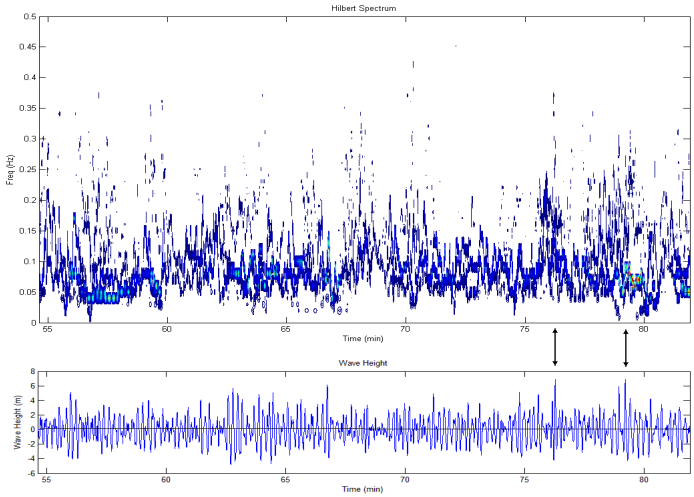

Figura 4.13: Gráfico de contorno del espectro marginal de Hilbert para el intervalo 2 de *Storm1999d* donde se muestran dos olas grandes.

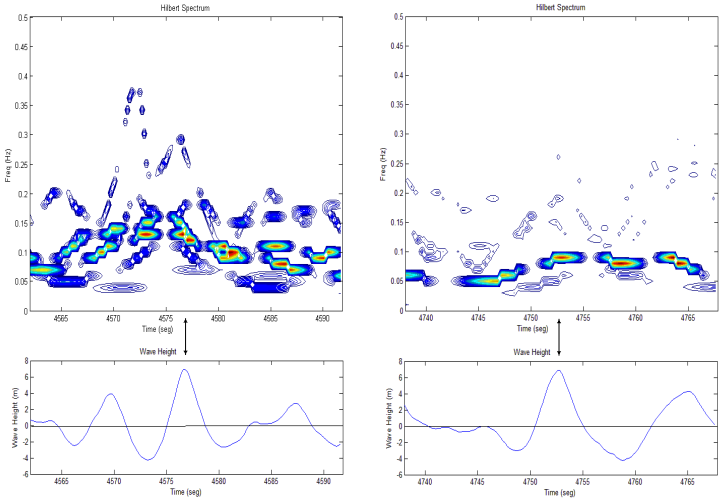

Figura 4.14: Gráfico de contornos de los espectros marginales de Hilbert para las Olas Grandes 1 y 2 de *Storm1999d*.

Tabla 4.1. Cada uno de los conjuntos lo segmentamos usando el algoritmo SLEX, y también los descompusimos en Funciones de Modo Intrínseco usando el software HHT-DPS. Para cada intervalo de la segmentación calculamos los espectros de frecuencia tanto con SLEX como los espectros marginales de Hilbert a partir de las IMF's.

El algoritmo SLEX es muy útil para la segmentación de un conjunto grande de datos como los utilizados en este trabajo. Como los estados del mar se pueden considerar estacionarios, podemos utilizar el algoritmo SLEX, pues este divide una serie temporal en segmentos que son aproximadamente estacionarios. Una desventaja de este algoritmo es que la longitud del archivo de datos debe ser una potencia de 2 ($L = 2^N$). Sin embargo, esta restricción es fácil de solventar con sólo completar con ceros el archivo de datos hasta llegar a una longitud potencia de 2, o truncarlos a la potencia de 2 más cercana a la longitud del archivo. El algoritmo SLEX también nos permite calcular los espectros de cada intervalo de la segmentación, así como un espectrograma del conjunto de datos (véase la Figura 4.6). Como se puede ver de los resultados obtenidos acá, el algoritmo SLEX es bastante útil para hallar la segmentación de un conjunto de datos a la vez que presenta los espectros de energía así como los espectrogramas correspondientes, lo que facilita el estudio o análisis espectral de una serie temporal.

También hicimos uso del algoritmo HHT para el análisis espectral de la tormenta. Este algoritmo ha sido ideado para descomponer una señal no-lineal y no-estacionaria; y durante una tormenta el estado del mar es claramente no-estacionario, lo que nos permite hacer uso del algoritmo sin dificultad. El software HHT-DPS realiza automáticamente la descomposición de la señal en Funciones de Modo Intrínseco, la salida del procedimiento son dos matrices $N \times k$, donde N es la longitud del archivo de datos y k el número de IMF's. Una de las matrices es **IMF** donde se almacenan todos los valores de las Funciones de Modo Intrínseco, y la otra matriz es **HT** donde están almacenados todas las frecuencias asociadas a cada IMF.

El hecho de que la resolución temporal del procedimiento HHT sea muy fina, nos permitió hacer un análisis espectral bastante robusto de cada intervalo de la segmentación de cada conjunto; esto también nos permitió realizar un análisis espectral de eventos importantes, es decir, para olas con altura mayor a 6.5m. Realizamos el análisis en un entorno de la ola que cubre 150 registros (30seg.), obteniendo una muy buena re-solución. Como se observa en los resultados obtenidos en este trabajo, el procedimiento HHT es bastante bueno para el análisis espectral detallado de una serie, donde podemos restringirnos a subconjuntos o intervalos muy pequeños del mismo y aún así podemos calcular los espectros marginales de Hilbert apropiadamente.

Finalmente podemos mencionar que aún cuando cada algoritmo ha sido ideado para una tarea distinta (SLEX para segmentación y HHT para descomposición de la señal) podemos valernos de ambos algoritmos para realizar un análisis espectral detallado de una serie; primero utilizamos el procedimiento HHT para con-

seguir las Funciones de Modo Intrínseco, segundo utilizamos el algoritmo SLEX para realizar la segmentación de la serie original (recuerde que ambos algoritmos trabajan independientemente); una vez realizado estos pasos, en tercer lugar podemos calcular los espectros marginales de Hilbert para cada intervalo de la segmentación obtenida con SLEX, haciendo uso de las IMF's.

Capítulo 5

Análisis de la gaussianidad del mar

En este capítulo, vamos a realizar pruebas de hipótesis de gaussianidad para varios conjuntos de datos: Estación 106 en Waimea Bay, Hawaii, USA. Estación 144 en St. Petersburg, Florida, USA. y Golfo de México, Huracán Camille. El objetivo es determinar el comportamiento gaussiano o no de las olas en un periodo de tiempo determinado. Realizaremos el ajuste y las pruebas de hipótesis para los registros de las olas.

5.1. Soporte teórico

Basaremos el estudio de la gaussianidad de los registros de olas por medio de una prueba de hipótesis basada en la estimación de densidades. Los conceptos y demostraciones que apoyan esta sección están contenidas en los apéndices. Supondremos

1. X_k el proceso de altura de olas es α-mixing.
2. Estimaremos la densidad marginal 1-dimensional por medio de un estimador de núcleo (Método de núcleo de Parzen-Rosenblatt)
$$\hat{f}_n(x) = \frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{x - X_n}{h}\right)$$
donde n es el tamaño de la muestra, K un núcleo de convolución positivo con $\int K dx = 1$ y $h = h(n)$ la ventana de estimación es una sucesión tal que $h(n) \to 0$ cuando $n \to \infty$.

Dados los datos de cada condición del mar, estimaremos las densidades marginales respectivas utilizando el método de núcleo. Usaremos dos tipos de núcleos: núcleo gaussiano y núcleo exponencial.
En general para el ajuste a la densidad estimada se usa el siguiente resultado

$$Y^n(y_j) = \sqrt{nh}\left(\hat{f}_n(y_j) - f(y_j)\right) \to N(0, \sigma^2(y_j)) \tag{5.1}$$

con $h = n^{-\alpha}$, $j = 1, 2, \ldots, m$ y una serie $X_1, X_2, \ldots, X_n$. Si $\alpha = \frac{1}{5}$ entonces la gaussiana límite resulta no centrada. Ahora de la ecuación (5.1) y de la independencia asintótica obtenemos

$$(Y^n(y_1), \ldots, Y^n(y_m)) \to N\left(0, \begin{pmatrix} \sigma^2(y_1) & & 0 \\ & \ddots & \\ 0 & & \sigma^2(y_m) \end{pmatrix}\right)_{nh\to\infty} \tag{5.2}$$

Dado el registro de olas $\{X_1, X_2, \ldots, X_n\}$ procedemos a realizar el ajuste estimado con núcleo $\hat{f}_n(y_j)$. Denotamos $\varphi_{\mu,\sigma^2}(x)$ la densidad gaussiana con media μ y varianza σ^2 y $\varphi_{\bar{X},s^2}(x)$ la densidad gaussiana con parámetros, $\bar{X}$ y s^2, estimados en el segmento. Entonces, también tenemos el resultado

$$\hat{Y}^n(X_j) = \sqrt{nh}\left(\hat{f}_n(X_j) - \varphi_{\bar{X},s^2}(X_j)\right) \to N(0, \underbrace{\sigma^2(X_j, \bar{X}, s^2)}_{\sigma_j^2}) = Y_j \text{ para } j = 1, \ldots, m. \tag{5.3}$$

Luego, se tiene

$$\sum_{j=1}^{m}\left(\frac{1}{\sigma(y_j)}Y_j\right)^2 = \chi_m^2 \tag{5.4}$$

donde χ_m^2 es una variable aleatoria chi-cuadrado con m grados de libertad. Este último estadístico nos servirá para construir la prueba de hipótesis de gaussianidad. En los apéndices desarrollamos estas ideas.

5.2. Análisis de registros de olas lineal

Como indicamos anteriormente, vamos a trabajar con varios conjuntos de datos: Estación 106 en Waimea Bay, Hawaii, USA. Estación 144 en St. Petersburg, Florida, USA. y Golfo de México, Huracán Camille. Estos conjuntos de datos fueron descritos en la introducción de esta tesis.

El primer paso en el análisis fue realizar una segmentación de cada conjunto de datos, en intervalos aproximadamente estacionarios, para ello nos valimos del algoritmo SLEX (para detalles véase Ortega & Hernández [(2006) [27]], Ombao et. al [(2002) [22]], Ombao et. al [(2004) [24]]). El algoritmo SLEX se describe en el capítulo 2, sección 2.3.

También hicimos uso del software WAFO para calcular los espectros de energía de cada intervalo de la segmentación dada por SLEX, de todos los conjuntos de datos. Así mismo calculamos a partir de los espectros

de energías algunas características básicas de olas: Altura significativa, H_m; período medio de olas, T_{m01} y período de pico espectral, T_p. En la Tabla 5.1 damos una lista de los 9 conjuntos de datos junto con estas características básicas, también se muestran los cortes obtenidos con SLEX en cada conjunto de datos.

Archivo	Duración	Cortes	H_m	T_{m01}	T_p
Estación 106	72h	31	1,1847	5,5742	9,0501
Estación 14401	18h21m	30	3,6736	6,8411	8,7219
Estación 14403	17h11m	45	3,5534	6,9026	9,1707
Camille	22h30m	17	5,5884	7,6788	10,7459

Tabla 5.1: Características básicas de los conjuntos de datos

El segundo paso fue estimar las densidades empíricas, para ello utilizamos dos funciones de núcleo: núcleo exponencial y núcleo gaussiano. Para cada intervalo de segmentación de cada uno de los conjuntos de archivos estimamos las densidades empíricas de las densidades marginales unidimensionales del proceso con ambos núcleos. Así mismo calculamos los errores cuadráticos medio entre estas densidades y la función de densidad normal de media cero y varianza uno. Por último realizamos las pruebas de hipótesis de gaussianidad, siendo las hipótesis:

- H_0: X tiene distribución normal $N(\mu, \sigma^2)$
- H_a: X no tiene distribución normal $N(\mu, \sigma^2)$

La prueba fue realizada sobre una malla de 201 puntos en el intervalo $[-5, 5]$, el estadístico de prueba es el dado por (A.12), para ambos núcleos y comparado con una χ^2_{m-1}. En todos los casos usamos $\alpha = 0{,}05$. Los verdaderos parámetros fueron sustituidos por los estimadores definidos en el apéndice.

Todos estos algoritmos fueron implementados en Matlab. Para verificar la validez del algoritmo de prueba de gaussianidad realizamos la prueba sobre 3000 conjuntos de datos gaussianos simulados de la siguiente forma:

Primero, generamos 1000 series de 2000 observaciones con frecuencia 1,280Hz utilizando la función `normrnd` de Matlab para generar números aleatorios a partir de una distribución gaussiana, usamos $\mu = 0, \sigma = 1$.

Generamos otras 1000 series de 2000 observaciones con frecuencia 1,280Hz a partir del espectro JONSWAP con $H_m = 4$ y $T_p = 9$, usando las funciones `jonswap([],[Hm Tp])` y `spec2sdat(S,[2000 1],0.78125)` de WAFO, lo cual nos da una simulación de mar gaussiano.

Las últimas 1000 series de datos las generamos utilizando el espectro TORSETHAUGEN con $H_m = 4$ y $T_p = 9$, usando las funciones `torsethaugen([],[Hm Tp],0)` y `spec2sdat(S,[2000 1],0.78125)` de WAFO las cuales también simulan un mar gaussiano (para detalles, véase (Ochi (1998)[21]) y (WAFO tutorial (2000)[34])).

Para estas 3000 series de datos calculamos las densidades empíricas marginales y luego realizamos la prueba de hipótesis de gaussianidad, en casi todos los casos se aceptó la hipótesis de gaussianidad H_0. Para las primeras 1000 series (generados con N(0,1)) se rechazaron 6 archivos lo que nos da un 99,4 % de aceptación de H_0. Para los siguientes 1000 conjuntos (generados con JONSWAP) se rechazaron 75 series, obteniéndose un 92,5 % de aceptación de H_0. Para las últimas 1000 series (generadas con TORSETHAUGEN) se rechazaron 49 conjuntos, lo que representa un 4,9 % por lo que se aceptan 95,1 % de los conjuntos como gaussianos. De los resultados anteriores, se tiene evidencia de que el nivel de la prueba se ve afectado por la dependencia de las observaciones. Para verificarlo se procedió a realizar la prueba de hipótesis con saltos en los registros, valiéndonos de la función de covarianza como base para los saltos en la toma de los registros, esto es, sea X la serie original y j el salto dado por la función de covarianza, entonces la nueva serie será, para $i = 1, \ldots, [\text{longitud}(X)/j]$, $X_s = X(i * j)$, siendo $[\cdot]$ la función parte entera. Sobre esta serie, realizamos la prueba de hipótesis de gaussianidad. Al calcular la función de covarianza, para las series generadas con ambos espectros, se puede observar que a partir de $j = 14$ la función es casi cero, por lo tanto construimos las series X_s para $j = 1, \ldots, 14$. Al realizar las pruebas de hipótesis se pudo notar que el nivel de la prueba mejora con cada salto. Para las series generadas a partir del espectro JONSWAP el nivel de aceptación va de 93,4 %, en el caso del registro completo, hasta 99,2 % para $j = 7$, de hecho para $j = 2$ el porcentaje de aceptación es de 97,8 %. Para los registros obtenidos a partir del espectro TORSETHAUGEN el nivel de aceptación va de 94,5 % con los registros completos hasta 99,1 % para $j = 7$. Al igual que el caso anterior, para $j = 2$ ya se observa una mejoría en el nivel de aceptación, siendo en este caso de 98,1 %. Se puede concluir que la dependencia entre las observaciones modifica el nivel de la prueba.

También procedimos a realizar la simulación de procesos gaussianos con el software R. Realizamos 2000 simulaciones de procesos gaussianos de longitud 2000 datos. Las primeras 1000 simulaciones las realizamos con la función **simFGN0(n,p)** y las otras 1000 simulaciones con la función **simARMA0(n,p)**, ambas funciones están implementadas en la librería *longmemo* de R. Para estas simulaciones procedimos a realizar las pruebas de hipótesis de gaussianidad obteniéndose los resultados mostrados en la Tabla 5.2

Como se observa en la Tabla 5.2 se tiene una aceptación por encima del 95 % para todas las simulaciones con ambos núcleos, el valor de significancia α usado fue de 0.05. En esta caso también realizamos el mismo procedimiento de prueba en función de los saltos dado por la función de covarianza obteniéndose que el nivel

	Núcleo Exponencial		Núcleo Normal	
Función	Aceptados	p-valor	Aceptados	p-valor
simFGN0	978	0.9377	982	0.9461
simARMA0	959	0.9270	962	0.9406

Tabla 5.2: Pruebas de hipótesis para las simulaciones de procesos gaussianos con R

de la prueba va de 95.9 % hasta 98,8 % para $j = 8$.

Podemos decir que el algoritmo implementado en Matlab es válido para realizar las pruebas de hipótesis de gaussianidad de los registros de alturas de olas que estamos considerando.

5.2.1. Análisis de registros de olas, Estación Waimea Bay

La Figura 5.1 muestra el registro de alturas y los cortes dados por SLEX para la Estación 106. Para detalles sobre el número de cortes conseguidos para este archivo, véase Ortega & Hernández (2006 [27]).

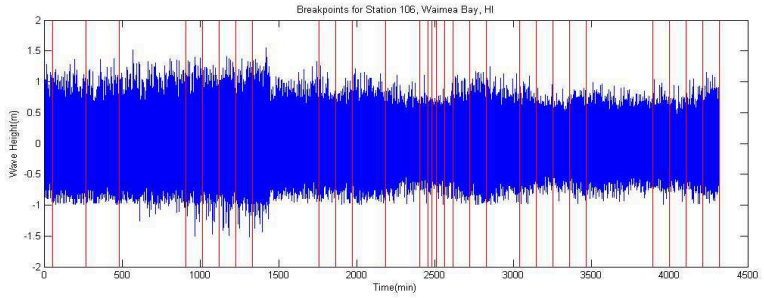

Figura 5.1: Cortes SLEX para la Estación 106

Las condiciones del mar de esta estación son los de la presencia de una tormenta moderada. Para cada uno de los segmentos del archivo *Estacion106* dados por el algoritmo SLEX, realizamos la prueba de hipótesis de gaussianidad. En este caso todos los segmentos resultaron no-gaussianos con un p-valor de 0. También realizamos la prueba de hipótesis de gaussianidad utilizando los saltos de la función de covarianza, para cada una de los segmentos con saltos de 1 a 10, todas las pruebas resultaron no gaussianas.

5.2.2. Análisis de registros de olas, Estación St. Petersburg

La Figura 5.2 muestra el registro de alturas y los cortes dados por SLEX para la Estación 144, meses enero y marzo. La media de las alturas significativas para esta estación está alrededor de 0.65m, al observar el boxplot de los registros de altura significativa anual desplegada por la *Coastal Data Information Program* (CDIP). De allí se puede observar que para los meses de enero a abril la altura significativa está alrededor de 1.1m evidenciando la presencia de tormenta, más aún tomando en cuenta que esta boya se encuentra en las costas del estado de Florida, y los meses de noviembre a marzo es la temporada de lluvia en esa región. Los parámetros usados para hallar la segmentación fueron: Para el caso *Estación 14401*: parámetro de penalización 0.50 y parámetro de suavizado 0.10, con nivel máximo 7. Para el caso de *Estación 14403*: parámetro de penalización 0.25 y parámetro de suavizado 0.10, con nivel máximo 7.

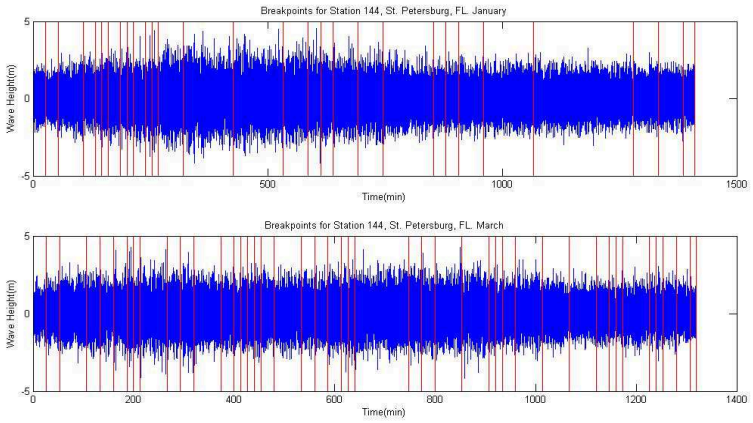

Figura 5.2: Cortes SLEX para la Estación 144, enero y marzo

Para cada uno de los segmentos del archivo *Estación 14401* dados por el algoritmo SLEX, realizamos la prueba de hipótesis de gaussianidad. En este caso los segmentos 1-5, 12-17 y 21-29 resultaron no-gaussianos con un p-valor de 0. Mientras que los intervalos 6-11 y 18-20 resultaron gaussianos con p-valor 1 o cercano a 1, excepto los intervalos 9 (p-valor1=0,6371, p-valor2=0,6543) y 10 (p-valor1=0,1712, p-valor2=0,4383). Denotamos p-valores1 a los p-valores correspondientes a la prueba de hipótesis con las densidades de núcleo exponencial y p-valores2 a los del núcleo normal o gaussiano. La prueba fue realizada sobre una malla de 201 puntos, el estadístico de prueba es el dado por (A.12), para ambos núcleos.

Para los segmentos del archivo *Estación 14403* los resultados fueron: Intervalos no-gaussianos: 1-3, 5-7, 14,

19, 21 y 31-44 con p-valor 0. Los intervalos gaussianos: 4, 8-13, 15-18, 20 y 22-30, con p-valores 1 o cercano a uno excepto los intervalos 9 (p-valor1=0,2016, p-valor2=0,3667), el intervalo 11 con p-valores 0,4211 y 0,6275 respectivamente. Para el intervalo 12 la prueba de hipótesis resultó gaussiana, para la densidad exponencial el valor del estadístico fue $T = 224{,}0880$ comparado con una χ^2 de 200 grados de libertad ($\chi^2_{200;0,05} = 233{,}9942$) con p-valor 0,1165, mientras que para el núcleo normal el valor del estadístico de prueba fue $T = 226{,}9794$ con p-valor 0,0924. En la Figura 5.3, a modo de ilustración, se observa el espectro de energía así como las densidades empíricas estimadas por medio de los núcleos exponencial y normal para el intervalo 12 de la *Estación 14403*. Para los intervalos 20 (p-valor1=0,3659, p-valor2=0,4470) y 23 (p-valor1=0,1269, p-valor2=0,2447).

Al igual que para la estación *Waimea Bay (106)* realizamos la prueba de hipótesis en función de los saltos dados por la función covarianza, se construyeron las series como en la sección 5.2) con saltos del 1 al 10. Para realizar la prueba de hipótesis, escogimos el valor $h = \pm 0{,}3$ como umbral para el cual escoger los saltos, de donde se obtuvo que para todos los intervalos el valor de salto para el cual la cola de la función de covarianza queda entre las cotas $\pm h$ es de 7. Al realizar la pruebas de hipótesis se obtuvieron los siguientes resultados: Para la *Estación 14401* los intervalos adicionales que resultaron gaussianos son: 5, 12, 13, 16, 17 y 21. Para la *Estación 14403* los intervalos gaussianos adicionales son: 3,5, 6, 7, 19, 21, 31, 32 y 33.

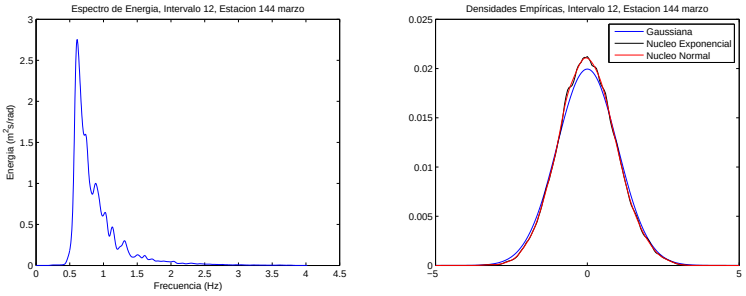

Figura 5.3: Densidad Espectral del intervalo 12 de la *Estación 14403* (izquierda) y densidades gaussiana y empírica (derecha)

5.2.3. Análisis de registros de olas, huracán Camille

La Figura 5.4 muestra el registro de alturas y los cortes dados por SLEX para el Huracán Camille. Véase Ortega & Hernández (2006 [27]) para mayores detalles. Cabe destacar que la segmentación obtenida allí con el algoritmo SLEX nos da 11 puntos de corte, la que mostramos tiene 17 puntos de cortes (16 segmentos). La razón de ello, es que nos valimos de los algoritmos SLEX y DCPC, que como se puede ver en ((2006)

[27]), hay intervalos grandes que son divididos en 2 o más intervalos por el otro algoritmo y que se observan cambios significativos en las características espectrales no detectadas por uno de ellos. Por tal razón se hizo una interpolación de los cortes dados por ambos algoritmos; en la Figura 5.4 estos cortes interpolados se muestran con una línea punteada. Para el huracán Camille también realizamos las pruebas de gaussianidad para cada uno de los segmentos dados por SLEX, en este caso como era de esperar todos los segmentos dieron no-gaussiano con p-valor 0.

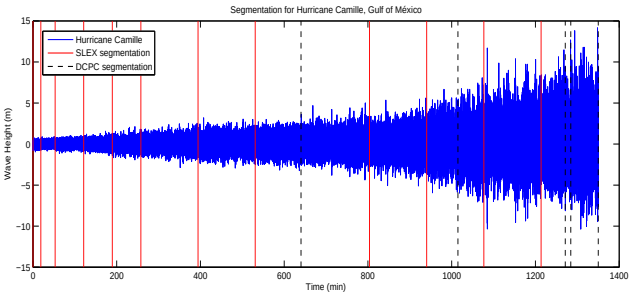

Figura 5.4: Segmentación para el Huracán Camille

5.2.4. Validación de las pruebas de hipótesis

Para comprobar la validez de los resultados de las estaciones, realizamos una simulación de los registros de olas de cada segmento basándonos en el procedimiento propuesto por Grigoriu (2009 [10]), para simulación de procesos gaussianos a partir de procesos no gaussianos.

Sea g la densidad espectral de un proceso X estacionario con segundo momento. La función de covarianza de X tiene la siguiente representación espectral:

$$\rho(\tau) = \int_0^\infty g(\nu) cos(\nu\tau) d\tau \tag{5.5}$$

1. Sea el rango de frecuencias $[0, \bar{\nu}], 0 < \bar{\nu} < \infty$ y sea

$$g_n(\nu) = \sum_{i=1}^{n} g(\nu_{i-1}) 1_{(\nu_{-i-1} < \nu < \nu_i)}$$

una aproximación de g donde $0 = \nu_0 < \nu_1 < \cdots < \nu_n = \bar{\nu}$

2. Sean

$$\begin{aligned}\rho(\tau) &= \int_0^{\infty} g(\nu)\cos(\nu\tau)d\tau \text{ y} \\ \tilde{\rho}_n(\tau) &= \int_0^{\bar{\nu}} g_n(\nu)\cos(\nu\tau)d\nu \\ &= \sum_{i=1}^{n} g(\nu_{i-1})\frac{\operatorname{sen}(\nu_i\tau)\operatorname{sen}(\nu_{i-1}\tau)}{\tau}\end{aligned}$$

funciones de covarianza de g y g_n

Sea $G(t)$ un proceso gaussiano estacionario con densidad espectral g y función de covarianza (5.5)

Consideremos un proceso gaussiano estacionario con media cero y varianza uno, y función de covarianza

$$\rho_n(\tau) = \frac{\tilde{\rho}_n(\tau)}{\tilde{\rho}_n(0)}$$

Dado que g_n es una densidad espectral, $\tilde{\rho}_n$ una función de covarianza, también lo será $\rho_n(\tau) = \frac{\tilde{\rho}_n(\tau)}{\tilde{\rho}_n(0)}$.
Finalmente, el proceso gaussiano simulado teniendo densidad espectral g_n que converge a g, es de la forma

$$Z_n(t) = \sum_{k=1}^{n} \sigma_k \left(A_k \cos(\nu_k t) + B_k \operatorname{sen}(\nu_k t)\right) \tag{5.6}$$

donde $\nu_k > 0$ son distintas frecuencias, A_k y B_k son variables aleatorias $N(0,1)$ y σ_k son constantes que se obtienen de la densidad espectral. La función de covarianza de Z_n es $\mathbb{E}[Z_n(t+\tau)Z_n(t)] = \sum_{k=1}^{n} \sigma_k^2 \cos(\nu_k\tau)$

Valiéndonos del procedimiento descrito arriba, realizamos las simulaciones de los registros de olas. Al realizar las pruebas de hipótesis de gaussianidad y tomando en cuenta la media μ y varianza σ^2 de cada segmento del registro de la estación 106 se obtuvieron los siguientes resultados:
Para la estación 106, se aceptaron casi todas las simulaciones como gaussianas a excepción de las simulaciones para los segmentos 2, 4 y 20. Para la estación 14401 se rechazaron las simulaciones correspondientes a los segmentos 14, 15, 25 y 26. Para la estación 14403, se aceptaron como gaussianas todas las simulaciones excepto las correspondientes a los intervalos 20, 25, 39 y 43. Finalmente para el huracán Camille se rechazaron las correspondientes a los segmentos 8, 9, 10, 11, 13 y 16, y las demás resultaron gaussianas.

Para la estación 106 y el huracán Camille, realizamos 100 simulaciones de cada uno de los segmentos obtenidos con SLEX, obteniendo los siguientes resultados: Para la estación 106, se obtuvo que el porcentaje de aceptación de las simulaciones como gaussianas está entre 71 % y 99 % con p-valor promedio entre 0,3271 y 0,9908, para el segmento 4 hay una excepción, ya que sólo se aceptó el 44 % de las simulaciones como gaussianas con un p-valor promedio de 0,2410. Para el huracán Camille, los porcentajes de aceptación de simulaciones gaussianas fue entre 59 % y 99 % con p-valor promedio entre 0,1738 y 0,9784. Similar al caso de la estación 106,

hay una excepción para el segmento 9, donde solo se aceptaron como gaussianas 4 simulaciones con p-valor promedio 0,3830, es decir, se rechazó el 96 % de las simulaciones de series de tiempo como gaussianas.

5.3. Conclusiones

Hemos realizados pruebas de hipótesis de gaussianidad para varios conjuntos de registros de alturas de olas con distintas condiciones del mar: tormenta moderada (Estación 106, Waimea Bay, Hawaii), lluvia fuerte en mares llanos (Estación 144, San Petersburg, Florida) y en presencia de huracán (Huracán Camille, Golfo de México).

El primer paso fue realizar una segmentación de cada uno de los conjuntos de datos utilizando el algoritmo SLEX para obtener segmentos o intervalos aproximadamente estacionarios (véase (2011)[12] y (2006)[27] para detalles de la segmentación obtenida.)

Realizamos la prueba de hipótesis sobre cada intervalo de la segmentación de cada conjunto de datos; se puede notar que el comportamiento de las olas del mar en presencia de tormentas o huracanes resulta un proceso no gaussiano como se ve en los casos de la Estación 106 en Waimea Bay, Hawaii y el huracán Camille en el Golfo de México. También se pudo comprobar que en mares llanos el comportamiento de las olas resulta ser un proceso no gaussiano como en el caso de la Estación 144 en San Petersburg, Florida. Las pruebas fueron realizadas con un nivel de significancia del 95 %.

También se puede observar que al considerar las series con saltos en los registros dados por la función de covarianza, la aceptación de la prueba de hipótesis mejora considerablemente, esto debido a que hay una mayor independencia entre los registros de olas.

El algoritmo de prueba de hipótesis fue validado realizando simulaciones gaussianas con distintos métodos. Realizando simulaciones gaussianas N(0,1), con el software Matlab, también usando el software R, y la herramienta WAFO de Matlab para crear simulaciones de olas gaussianas. En todos los casos, las pruebas de hipótesis resultaron gaussianas en su mayoría con un nivel de significancia del 95 %.

En el caso de simulación de procesos gaussianos a partir de la densidad espectral, se pudo comprobar que dado un espectro calculado a partir de un registro de olas, ya sea esta gaussiana o no, se obtiene un proceso simulado gaussiano al usar la técnica propuesta por Grigoriu (2009 [10]). Las pruebas de hipótesis dan una aceptación de gaussianidad superior al 60 % de las simulaciones utilizando un nivel de significancia del 95 %.

Apéndice A

Estimación de densidades para muestras independientes

Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria independiente y sea $\varphi_{\mu,\sigma^2}(x)$ su función de densidad.
Sea K una función de núcleo satisfaciendo las siguientes condiciones:

1. $K \in L^1$
2. $K \geq 0$
3. K es tal que $\begin{cases} h_n \to 0 \\ nh_n \to \infty \end{cases}$.

Consideremos la siguiente función de densidad empírica

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - x_k}{h}\right).$$

Estudiaremos primero el caso de una muestra independiente e idénticamente distribuida. Bajo la hipótesis $H_0 : f = \varphi_{\mu,\sigma^2}$, calcularemos el estadístico de prueba.
Sea $(\hat{f}_n(x) - \varphi_{\mu,\sigma^2}(x))$ y consideremos

$$c_n \sum_{i=1}^{m} (\hat{f}_n(x) - \varphi_{\mu,\sigma^2}(x))^2.$$

¿Cuál es la distribución?

Para ello, necesitamos realizar algunos cálculos previos. Empecemos calculando $\mathbb{E}[\hat{f}_n(x)]$

$$\begin{aligned}\mathbb{E}[\hat{f}_n(x)] &= \frac{1}{hn} n \int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right) \varphi_{\mu,\sigma^2}(y)dy = \int_{-\infty}^{\infty} K(u)\varphi_{\mu,\sigma^2}(x-hu)du \\ &= \varphi_{\mu,\sigma^2}(x) + \int_{-\infty}^{\infty} K(u)\varphi'_{\nu,\sigma^2}(x)(-hu)du + \int_{-\infty}^{\infty} \varphi''_{\mu,\sigma^2}(\theta(x,u,h))\frac{h^2u^2}{2}du\end{aligned} \tag{A.1}$$

donde

$$\varphi'_{\mu,\sigma^2}(u) = -\frac{1}{\sqrt{2\pi}\sigma^3}(u-\mu)e^{-\frac{(u-\mu)^2}{2\sigma^2}}.$$

Si K es una función par, el segundo término de la ecuación anterior es cero, luego

$$\mathbb{E}[\hat{f}_n(x)] = \varphi_{\mu,\sigma^2}(x) + \frac{h^2}{2}\int_{-\infty}^{\infty} K(u)\varphi''_{\mu,\sigma^2}(\theta(u,x,h))u^2du. \tag{A.2}$$

De acá, el sesgo es

$$\left[\mathbb{E}[\hat{f}_n(x)] - \varphi_{\mu,\sigma^2}(x)\right] = \frac{h^2}{2}\int_{-\infty}^{\infty} K(u)\varphi''_{\mu,\sigma^2}(\theta(u,x,h))u^2du \tag{A.3}$$

dividiendo por h^2

$$\frac{1}{h^2}\left[\mathbb{E}[\hat{f}_n(x)] - \varphi_{\mu,\sigma^2}(x)\right] = \frac{1}{2}\int_{-\infty}^{\infty} K(u)\varphi''_{\mu,\sigma^2}(\theta(u,x,h))u^2du. \tag{A.4}$$

Tomando límite en n, se tiene

$$\lim_{n\to\infty}\frac{1}{h^2}\left[\mathbb{E}[\hat{f}_n(x)] - \varphi_{\mu,\sigma^2}(x)\right] = \frac{1}{2}\left[\int_{-\infty}^{\infty} K(u)u^2du\right]\varphi''_{\mu,\sigma^2} := K_{12}\frac{1}{2}\varphi''_{\mu,\sigma^2}. \tag{A.5}$$

Si se tiene el sesgo al cuadrado, entonces el límite es

$$\lim_{n\to\infty}\frac{1}{h^2}\left[\mathbb{E}[\hat{f}_n(x)] - \varphi_{\mu,\sigma^2}(x)\right]^2 = \frac{1}{4}K_{12}^2(\varphi''_{\mu,\sigma^2}(x))^2. \tag{A.6}$$

Calculemos ahora la varianza de $\hat{f}_n(x)$

$$\text{var}\hat{f}_n(x) = \mathbb{E}\left[\hat{f}_n(x) - \mathbb{E}\hat{f}_n(x)\right]^2. \tag{A.7}$$

Como $\hat{f}_n(x) = \frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{x-x_k}{h}\right)$.

Luego

$$\hat{f}_n(x) - \mathbb{E}[\hat{f}_n(x)] = \frac{1}{hn}\sum_{k=1}^{n} Y_k(x)$$

donde $Y_k(x) = K(\frac{x-y_k}{h}) - \mathbb{E}\left[K(\frac{x-y_k}{h})\right]$.

Luego

$$\begin{aligned}\mathbb{E}\left[\hat{f}_n(x) - \varphi_{\mu,\sigma^2}(x)\right]^2 &= \frac{1}{(hn)^2}\text{var}\left(\sum_{k=1}^{n} Y_k(x)\right) = \frac{1}{h^2n^2}\sum_{k=1}^{n}\text{var}(Y_k(x)) \\ &= \frac{1}{nh^2}\text{var}(Y_k(x)).\end{aligned} \tag{A.8}$$

Ahora bien

$$\text{var}(Y_k(x)) = \mathbb{E}\left[K^2\left(\frac{x-y_k}{h}\right)\right] - \left[\mathbb{E}\left[K\left(\frac{x-y_k}{h}\right)\right]\right]^2.$$

$$\mathbb{E}\left[K\left(\frac{x-y}{h}\right)\right] = \int_{-\infty}^{\infty} K\left(\frac{x-y_k}{h}\right)\varphi_{\mu,\sigma^2}(y)dy = h\int_{-\infty}^{\infty} K(u)\varphi_{\mu,\sigma^2}(x-hu)du \cong h\varphi_{\mu,\sigma^2}(x)$$

y

$$\mathbb{E}\left[K^2\left(\frac{x-y}{h}\right)\right] = \int_{-\infty}^{\infty} K^2\left(\frac{x-y_k}{h}\right)\varphi_{\mu,\sigma^2}(y)dy = h\int_{-\infty}^{\infty} K^2(u)\varphi_{\mu,\sigma^2}(x-hu)du.$$

Por lo tanto

$$\begin{aligned}\mathbb{E}\left[\hat{f}_n(x) - \varphi_{\mu,\sigma^2}(x)\right]^2 &= \frac{1}{nh^2}\left[h\int_{-\infty}^{\infty} K^2(u)\varphi_{\mu,\sigma^2}(x-hu)du - h^2\left[\int_{-\infty}^{\infty} K(u)\varphi_{\mu,\sigma^2}(x-hu)du\right]^2\right] \\ &= \frac{1}{nh}\left[\int_{-\infty}^{\infty} K^2(u)\varphi_{\mu,\sigma^2}(x-hu)du - h\left[\int_{-\infty}^{\infty} K(u)\varphi_{\mu,\sigma^2}(x-hu)du\right]^2\right].\end{aligned}$$

Así, obtenemos

$$\mathbb{E}\left[\hat{f}_n(x) - \varphi_{\mu,\sigma^2}(x)\right]^2 = \underbrace{\frac{1}{nh}\left[\int_{-\infty}^{\infty} K^2(u)\varphi_{\mu,\sigma^2}(x-hu)du\right]}_{\downarrow \int_{-\infty}^{\infty} K^2(u)du\varphi_{\mu,\sigma^2}(x)} + \underbrace{\left[\mathbb{E}[\hat{f}_n(x)] - \varphi_{\mu,\sigma^2}(x)\right]^2 h^4}_{\downarrow \frac{K_{12}^2}{4}(\varphi''_{\mu,\sigma^2}(x))^2}$$

siendo $K_{12} = \int_{-\infty}^{\infty} K(u)u^2du$. Para balancear ambos términos, necesitamos que $\frac{1}{nh} = h^4$, de donde $h = n^{-1/5}$, luego

$$\mathbb{E}\left[\hat{f}_n(x) - \varphi_{\mu,\sigma^2}(x)\right]^2 = n^{-4/5}\left[\int_{-\infty}^{\infty} K^2(u)\varphi_{\mu,\sigma^2}(x-n^{-1/5}u)du\right] + \frac{K_{12}^2}{4}(\varphi''_{\mu,\sigma^2}(x))^2 n^{-4/5}.$$

Luego

$$n^{4/5}\mathbb{E}\left[\hat{f}_n(x) - \varphi_{\mu,\sigma^2}(x)\right]^2 \underset{n\to\infty}{\to} \overbrace{\left[\int_{-\infty}^{\infty} K^2(u)du\right]\varphi_{\mu,\sigma^2}(x) + \frac{(\int_{-\infty}^{\infty} K(u)u^2du)^2}{4}(\varphi''_{\mu,\sigma^2}(x))^2}^{(I)}.$$

Entonces,

$$\mathbb{E}\left[\hat{f}_n(x) - \varphi_{\mu,\sigma^2}(x)\right]^2 \to (I).$$

Más aún

$$\sum_{i=1}^{m} n^{4/5}\mathbb{E}\left[\hat{f}_n(x_i) - \varphi_{\mu,\sigma^2}(x_i)\right]^2 \to \|K\|_2^2\sum_{i=1}^{m}\varphi_{\mu,\sigma^2}(x_i) + \frac{K_{12}^2}{4}\sum_{i=1}^{m}(\varphi''_{\mu,\sigma^2}(x_i))^2.$$

Se puede demostrar por medio del Teorema de Linderberg que

$$\begin{aligned} n^{2/5}(\hat{f}_n(x) - \mathbb{E}[\hat{f}_n(x)]) &\rightarrow N(0, \sigma^2(x)) \\ &= N(0, \|K\|^2 \varphi_{\mu,\sigma^2}(x)). \end{aligned}$$

Por lo tanto

$$\frac{n^{2/5}(\hat{f}_n(x) - \mathbb{E}[\hat{f}_n(x))]}{\|K\| \varphi^{1/2}_{\mu,\sigma^2(x)}} \rightarrow N(0,1).$$

Nos interesa

$$\begin{aligned} n^{2/5}(\hat{f}_n(x) - f(x)) &= n^{2/5}(\hat{f}_n(x) - \mathbb{E}[\hat{f}_n(x)]) + n^{2/5}(\mathbb{E}[\hat{f}_n(x)] - f(x)) \\ &\rightarrow N(0, \|K\|^2 \varphi_{\mu,\sigma^2}(x)) + \frac{K_{12}}{2} \varphi''_{\mu,\sigma^2}(x) \end{aligned}$$

lo que nos da

$$n^{2/5}(\hat{f}_n(x) - f(x)) \rightarrow N\left(\frac{K_{12}}{2} \varphi''_{\mu,\sigma^2}(x), \|K\|^2 \varphi_{\mu,\sigma^2}(x)\right). \tag{A.9}$$

Centrando, se tiene

$$n^{2/5}(\hat{f}_n(x) - f(x)) - \frac{K_{12}}{2} \varphi''_{\mu,\sigma^2}(x) \rightarrow N\left(0, \|K\|^2 \varphi_{\mu,\sigma^2}(x)\right) \tag{A.10}$$

y además

$$\frac{1}{\|K\| \varphi^{1/2}_{\mu,\sigma^2}(x)} \left[n^{2/5}(\hat{f}_n(x) - f(x)) - \frac{K_{12}}{2} \varphi''_{\mu,\sigma^2}(x)\right] \rightarrow N(0,1). \tag{A.11}$$

Se puede demostrar que para $y_j \neq y_i$ se tiene independencia asintótica de los límites correspondientes.

$$\sum_{i=1}^{m} \left(\frac{1}{\|K\| \varphi^{1/2}_{\mu,\sigma^2}(x)} \left[n^{2/5}(\hat{f}_n(x) - f(x)) - \frac{K_{12}}{2} \varphi''_{\mu,\sigma^2}(x)\right] \right)^2 \underset{d}{\rightsquigarrow} \chi^2_m. \tag{A.12}$$

Observaciones:

1. $h = n^{-1/5}$ ventana optima, pues equilibra sesgo y varianza.

2. Si $h = n^{-\alpha}$ para $\alpha > 1/5$, se tiene que el término de sesgo $\frac{K_{12}}{2} \varphi''_{\mu,\sigma^2}(x)$ desaparece.

Apéndice B

Representación de la densidad gaussiana bivariada en la base de Hermite

Sea $(X, Y) \in \Re$ un par de variables aleatorias, con medias $\mathbb{E}[X] = 0, \mathbb{E}[Y] = 0$ y matriz de covarianza $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.
Como $\mathbb{E}[X] = 0, \mathbb{E}[Y] = 0$ se tiene que $cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] = \rho$.
Considere los polinomios de Hermite $H_n(x)$

$$H_n(x) = (-1)^n \frac{d^n}{dx^n} \left(e^{-x^2/2} \right) e^{x^2/2}. \tag{B.1}$$

Tenemos

$$\langle H_n, H_m \rangle_{L^2(\varphi)} = \delta_{n,m} n!$$

de donde $\|H_n\|^2 = n!$. Además constituyen una base de $L^2(\varphi)$.
Sea $f \in L^2(\varphi)$, entonces

$$\int_{-\infty}^{\infty} f^2(x)\varphi(x)dx < \infty.$$

Además

$$f(x) = \sum_{n=0}^{\infty} \hat{f}_n H_n(x).$$

Para calcular $\hat{f}_n$, hacemos

$$\langle f, H_m \rangle_{L^2(\varphi)} = \sum_{n=0}^{\infty} \hat{f}_n \langle H_n, H_m \rangle_{L^2(\varphi)} = \hat{f}_m m! \tag{B.2}$$

Entonces

$$\hat{f}_m = \frac{1}{m!}\langle f, H_m\rangle_{L\varphi^2(\varphi)} = \frac{1}{m!}\int_{-\infty}^{\infty} f(x)H_m(x)\varphi(x)dx. \tag{B.3}$$

De lo anterior, se tiene

$$\begin{aligned}\hat{f}_0 &= \int_{-\infty}^{\infty} f(x)\varphi(x)dx = \mathbb{E}[f(X)] = 0\\ \hat{f}_n &= \frac{1}{n!}\int_{-\infty}^{\infty} f(x)H_n(x)\varphi(x)dx = \frac{1}{n!}\mathbb{E}[f(X)H_n(X)].\end{aligned}$$

Si X e Y son Gaussianas con correlación ρ. Por la fórmula de Mehler

$$\mathbb{E}[H_n(X)H_m(Y)] = \delta_{n,m}n!\rho^n. \tag{B.4}$$

Sean $f \in L^2(\varphi), g \in L^2(\varphi)$ tal que $\mathbb{E}[f(X)] = 0, \mathbb{E}[g(Y)] = 0$. Entonces

$$\begin{aligned}cov\,[f(X), f(Y)] &= \mathbb{E}[f(X)f(Y)]\\ &= \sum_{n=1}^{\infty}\hat{f}_n\hat{g}_m\mathbb{E}[H_n(X)H_m(Y)]\\ &= \sum_{n=1}^{\infty}\hat{f}_n\hat{g}_n n!\rho^n.\end{aligned}$$

Ahora, sean $(X, Y) \sim \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ un vector aleatorio Gaussiano con matriz de covarianza la indicada.

Sea

$$\varphi_\rho(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right]$$

su función de densidad conjunta; usando los resultados anteriores

$$\varphi_\rho(x, y) = \frac{e^{-x^2/2}}{2\pi}\sum_{n=0}^{\infty}\frac{\rho^n}{n!}H_n(x)H_n(y)e^{-y^2/2}.$$

Dividiendo entre $\varphi(x)\varphi(y)$ queda

$$\frac{\varphi_\rho(x, y)}{\varphi(x)\varphi(y)} = \sum_{n=0}^{\infty}\frac{\rho^n}{n!}H_n(x)H_n(y).$$

Por otra parte

$$\begin{aligned}\frac{1}{n!m!}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\varphi_\rho(x, y)H_n(x)H_n(y)dxdy &= \frac{1}{n!m!}\mathbb{E}[H_n(x)H_m(y)]\\ &= \frac{n!\rho^n\delta_{n,m}}{(n!)^2} = \frac{1}{n!}\rho^n\delta_{n,m}\\ &= \begin{cases} 0, & \text{si } n \neq m\\ \frac{1}{n!}\rho^n, & \text{si } n = m.\end{cases}\end{aligned}$$

Luego

$$\frac{\varphi(x,y)}{\varphi(x)\varphi(y)} = \sum_{n=0}^{\infty} \frac{1}{n!}\rho^n H_n(x)H_n(y) \tag{B.5}$$

si y sólo si

$$\varphi_\rho(x,y) = \varphi(x)\varphi(y) \sum_{n=0}^{\infty} \frac{1}{n!}\rho^n H_n(x)H_n(y). \tag{B.6}$$

Apéndice C

Definición de dos tipos de mixing y aplicación a procesos gaussianos

Definición C.0.1 *Para cualesquiera dos σ-álgebras $\mathcal{A}$ y $\mathcal{B}$ la expresión siguiente*

$$\alpha(\mathcal{A},\mathcal{B}) = \sup_{(A\in\mathcal{A},B\in\mathcal{B})} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|, \tag{C.1}$$

denota el coeficiente de mezcla fuerte (strong mixing).

Los coeficientes $(\alpha(n))_{(n>0)}$ de la sucesión $(X_i)_{i\in\mathbb{Z}}$ se definen como

$$\alpha(n) = \sup_{k\in\mathbb{Z}} \alpha(\mathcal{F}_{k-n}, \mathcal{G}_k),$$

donde se ha definido $\mathcal{F}_l = \sigma(X_i : i \leq l)$ y $\mathcal{G}_l = \sigma(X_i : i \geq l)$.

Un coeficiente de dependencia débil muy útil para el caso de sucesiones gaussianas es el de ρ-mixing. Sean $\mathbb{L}^2(\mathcal{A})$ y $\mathbb{L}^2(\mathcal{B})$ los espacios de funciones de cuadrado integrables con respecto a las σ-álgebras $\mathcal{A}$ y $\mathcal{B}$ respectivamente entonces definimos

$$\rho(\mathcal{A},\mathcal{B}) = \sup\{|Cov(f,g)| : f \in \mathbb{L}^2(\mathcal{A}), ||f|| \leq 1, g \in \mathbb{L}^2(\mathcal{B}), ||g|| \leq 1\}. \tag{C.2}$$

Se tiene la desigualdad inmediata

$$4\alpha(\mathcal{A},\mathcal{B}) \leq \rho(\mathcal{A},\mathcal{B}). \tag{C.3}$$

Esta desigualdad es consecuencia de la definición pues

$$\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{E}[1_A 1_B] - \mathbb{E}[1_A]\mathbb{E}[1_B],$$

y además para todo conjunto A se tiene $Var(1_A) \leq \frac{1}{4}$. En el caso gaussiano se tiene la desigualdad inversa demostrada por Kolmogorov-Rosanov. Sean X e Y dos variables aleatorias gaussianas centradas y de varianza 1, tales que X es $\mathcal{A}$-medible e Y es $\mathcal{B}$-medible que satisfacen

$$r = \mathbb{E}[XY] \geq \rho(\mathcal{A}, \mathcal{B}) - \varepsilon.$$

Sean $U = X^{-1}[x, \infty)$ y $V = Y^{-1}[x, \infty)$. De esta forma se tiene que

$$\mathbb{P}(U \cap V) = \mathbb{E}\left[1_{[x,\infty)}(X)1_{[x,\infty)}(Y)\right] \quad \text{y} \quad \mathbb{P}(U) = \mathbb{E}\left[1_{[x,\infty)}(X)\right],$$

igualmente para la otra variable. Es fácil ver cuales son los coeficientes en la base de Hermite de la función $1_{[x,\infty)}(y)$.

$$1_{[x,\infty)}(y) = \sum_{n=0}^{\infty} c_n(x)H_n(y) \quad c_n(x) = \frac{1}{n!}H_{n-1}(x)\varphi(x).$$

Usando la fórmula de Mehler se tiene

$$|\mathbb{P}(U \cap V) - \mathbb{P}(U)\mathbb{P}(V)| = \varphi^2(x) \sum_{n=1}^{\infty} \frac{1}{n!}H_{n-1}^2(x)r^n = \varphi^2(x) \sum_{n=0}^{\infty} \frac{1}{(n+1)!}H_n^2(x)r^{n+1}. \tag{C.4}$$

De lo que se desprende, si $\varphi_\rho(x, y)$ denota la densidad gaussiana estándar bidimensional con correlación ρ, que

$$\text{(C.4)} = \varphi^2(x) \sum_{n=0}^{\infty} \frac{1}{n!}H_n^2(x) \int_0^r \rho^n d\rho = \int_0^r \varphi_\rho(x, x)d\rho = \int_0^r \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{x^2}{1+\rho}} d\rho. \tag{C.5}$$

De aquí se extrae por la definición del coeficiente α que

$$\alpha(\mathcal{A}, \mathcal{B}) \geq \int_0^{\rho(\mathcal{A},\mathcal{B})-\varepsilon} \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{x^2}{1+\rho}} d\rho. \tag{C.6}$$

Y como ε es arbitrario

$$\alpha(\mathcal{A}, \mathcal{B}) \geq \int_0^{\rho(\mathcal{A},\mathcal{B})} \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{x^2}{1+\rho}} d\rho. \tag{C.7}$$

Tomando $x = 0$ se obtiene

$$\alpha(\mathcal{A}, \mathcal{B}) \geq \frac{1}{2\pi} \arcsin \rho(\mathcal{A}, \mathcal{B}), \tag{C.8}$$

uniendo esta desigualdad con la demostrada más arriba (C.3) se obtiene finalmente

$$4\alpha(\mathcal{A}, \mathcal{B}) \leq \rho(\mathcal{A}, \mathcal{B}) \leq \sin(2\pi\alpha(\mathcal{A}, \mathcal{B})) \leq 2\pi\alpha(\mathcal{A}, \mathcal{B}). \tag{C.9}$$

Para enunciar el teorema siguiente debemos señalar que en el caso gaussiano también son equivalentes la noción de ρ-mixing con la de proceso linealmente regular. Para introducir esta última noción consideremos

$H(0,s)$ el subespacio lineal engendrado por las variables X_i, $0 \leq i \leq s$ y $H(\tau,\infty)$ el subespacio lineal generado por X_i, $i \geq \tau$. Definimos el coeficiente de regularidad lineal como

$$\tilde{\rho}(n) = \sup_{l\geq 0}\{|\mathbb{E}[\eta_1\eta_2]|, \eta_1 \in H(0,l),\ ||\eta_1|| = 1;\ \eta_2 \in H(l+n,\infty),\ ||\eta_2|| = 1\}.$$

El proceso $X = (X_i)_{i\in\mathbb{Z}}$ es completamente regular si $\tilde{\rho}(n) \to 0$ siempre que $n \to \infty$. Para procesos Gaussianos se tiene el siguiente resultado de Kolmogorov-Rosanov: $\rho(n) = \tilde{\rho}(n)$. La demostración de este hecho es de nuevo una aplicación de la fórmula de Mehler.
Existen resultados que establecen una relación entre la mejor aproximación por polinomios trigonométricos de la densidad espectral del proceso y el coeficiente de ρ-mixing. Las desigualdad anterior nos permite usar este resultado para acotar el coeficiente de α-mixing. A continuación enunciaremos uno de tales resultados. (Ver [8]página 58-59.)

Teorema C.0.1 *Sea* $\mathbf{X} = (X_k)_{k\in\mathbb{Z}}$ *un proceso gaussiano de media cero, con densidad espectral*

$$f(\lambda) = \sum_{k\in\mathbb{Z}} \mathbb{E}[X_0X_k]\, e^{ik\lambda} \tag{C.10}$$

tal que $f(\lambda) \geq a > 0$ *para* $\lambda \in [0,2\pi]$. *Entonces el coeficiente de* ρ_X*-mixing del proceso satisface*

$$\rho_X(k) \leq \frac{1}{a}\Delta_k(f), \tag{C.11}$$

donde $\Delta_k(f) = \text{ínf}\{||f-P||_\infty\}$ *es la mejor aproximación uniforme de la función* f *por polinomios trigonométricos de grado* $(k-1)$.

Este resultado y la equivalencia entre las dos nociones de mixing en el caso gaussiano, dan como consecuencia lo siguiente

Corolario C.0.1 *Supongamos que* $\mathbf{X} = (X_k)_{k\in\mathbb{Z}}$ *es un proceso gaussiano estacionario con densidad espectral acotada inferiormente por a y que satisface* $Cov(X_0,X_l) = O(|l|^{-m})$ *para algún* $m > 1$ *entonces* $\alpha_X(k) = O(k^{1-m})$.

Demostración. Dada las desigualdades tenemos que

$$\alpha_X(k) \leq \frac{1}{4}\rho_X(k) \leq \frac{1}{4a}\Delta_k(f). \tag{C.12}$$

Por otra parte tenemos $\Delta_k(f) \leq ||f-S_k||_\infty$ donde S_k es la suma parcial de Fourier de la densidad espectral f. Además se tiene que

$$f(\lambda) - S_k(\lambda) = \sum_{|l|>k} r(l)e^{il\lambda},$$

de donde

$$||f(\lambda) - S_k(\lambda)|| \leq \sum_{|l|>k} |r(l)| \equiv C \sum_{|l|>k} |l|^{-m} = \frac{C}{1-m} k^{-m+1}.$$

□

Ahora podemos establecer la ley fuerte de los grandes números para sucesiones α-mixing. Sabemos que si $X = (X_i)_{i\in\mathbb{N}}$ es una sucesión estacionaria α-mixing entonces si $\mathbb{E}|X_i|^{2+\delta} < \infty$, entonces (ver [8] página 42)

$$\frac{S_n}{n} \to \mathbb{E}[X_i] := \mu \quad \text{c.s.}$$

Dos aplicaciones de este resultado se dan para la estimación de la media y la varianza de un proceso α-mixing, verificando la condición de momentos antes reseñada. Así, sea $X_1, X_2, \ldots, X_n$ una muestra de un tal proceso entonces

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \to \mu. \tag{C.13}$$

Además, si $\sum_{n=1}^{\infty} \alpha^{\frac{\delta}{(2+\delta)}}(n) < \infty$, entonces

$$\sqrt{n}(\overline{X}_n - \mu) \to N(0, \gamma^2). \tag{C.14}$$

La varianza límite se escribe por medio de la fórmula

$$\gamma^2 = \mathbb{E}[X_1 - \mu]^2 + 2\sum_{j=2}^{\infty} Cov(X_1, X_j) = r(0) + 2\sum_{j=1}^{\infty} r(j) = f(0).$$

En la expresión anterior r y f designan la función de covarianza y la densidad espectral de la sucesión X_i.
Otro parámetro de interés para estimar es la varianza de la muestra, esto es, $\sigma^2 = \mathbb{E}[X_i - \mu]^2$. Para estimarla se define

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X}_n - \mu)^2.$$

Bajo la hipótesis de que $\mathbb{E}|X_i|^{4+\delta} < \infty$ y al usar la LFGN se tiene que $\hat{\sigma}_n^2 \to \sigma^2$ c.s. Además bajo la misma condición de momento y la convergencia de la serie de los coeficientes de mixing se verifica que:

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \to N(0, \gamma_1^2), \tag{C.15}$$

donde $\gamma_1^2 = Var((X_1 - \mu)^2) + 2\sum_{i=2}^{\infty} Cov((X_1 - \mu)^2, (X_i - \mu)^2)$. En el caso gaussiano podemos dar una expresión más explícita de este parámetro. Si H_2 es el polinomio de Hermite de grado 2 se tiene

$$\gamma_1^2 = \sigma^4 \left[\mathbb{E}\left[H_2(\frac{X_1 - \mu}{\sigma}) \right]^2 + 2\sum_{i=2}^{\infty} Corr(H_2(\frac{X_1 - \mu}{\sigma}), H_2(\frac{X_i - \mu}{\sigma})) \right].$$

Al usar la fórmula de Mehler se obtiene.

$$\gamma_1^2 = 2[r^2(0) + 2\sum_{i=1}^{\infty} r^2(i)] = 2(f * f(0)) > 0.$$

Consideremos ahora la estimación de la densidad marginal de un proceso α-mixing. Sea K una densidad en la recta y para cada y_l definamos el estimador de núcleo de

$$\hat{f}_n(y_l) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{y_l - X_i}{h}), \tag{C.16}$$

si definimos

$$Z_n(y_l) = \sqrt{nh}(\hat{f}_n(y_l) - \mathbb{E}[\hat{f}_n(y_l)]) \tag{C.17}$$

se tiene el siguiente teorema (ver [1] y [30]).

Teorema C.0.2 *Supongamos que el proceso $X = (X_n)$ es α-mixing y admite una densidad marginal f, además $\alpha(n) = O(n^{-a})$ para algún $a > 1$, y las densidades bivariadas $f_k(x, y)$ de los vectores (X_1, X_k) existen y verifican*

$$\sup_k \sup_{(y,z)\in\mathbb{R}^2} f_k(y, z) < \infty.$$

Entonces

$$(Z_n(y_1), \ldots, Z_n(y_m)) \to N(0, D(y_1, y_2, \ldots, y_m)). \tag{C.18}$$

Donde $D(y_1, y_2, \ldots, y_m)$ es una matriz diagonal tal que $d_{ii} = f(y_i)\int_{\mathbb{R}} K^2(u)du$.

Este resultado permite implementar pruebas de gaussianidad. Se puede demostrar que si $h_n = O(n^{-\beta})$ con $\beta < 1/5$ entonces podemos centrar el estimador con la densidad, obteniendo así que si definimos $Y_n(x_l) = \sqrt{nh}(\hat{f}_n(x_l) - f(x_l))$ se tiene entonces que:

$$\left(\frac{Y_n(x_1)}{\sqrt{d_{11}}}, \ldots, \frac{Y_n(x_m)}{\sqrt{d_{mm}}}\right) \to N(0, I_m).$$

De esta manera el estadístico

$$S_n^2 = \sum_{i=1}^{m} \frac{Y_n^2(x_i)}{d_{ii}} \to \chi_m^2. \tag{C.19}$$

El Teorema C.0.2 vale para la muestra $\{x_i\} \in \mathbb{R}^d, d > 1$ y el enunciado es casi el mismo, ver [1] y [30].

Apéndice D

Convergencia de la densidad empírica a la densidad gaussiana con parámetros estimados

Sea

$$\sqrt{nh}\left(\frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi}\sigma}-\frac{e^{-x^2/2s^2}}{\sqrt{2\pi}s}\right). \tag{D.1}$$

Sabemos que $s^2-\sigma^2 \to 0$ c.s. No hay que olvidar que s^2 es una variable aleatoria. Definamos la función $g(\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{x^2}{2\sigma^2}}$. Entonces

$$(\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{x^2}{2\sigma^2}}-\frac{1}{s\sqrt{2\pi}}\,e^{-\frac{x^2}{2s^2}})=(g(\sigma^2)-g(s^2)).$$

De esta forma

$$\begin{aligned}(g(\sigma^2)-g(s^2)) &= \frac{1}{\sigma\sqrt{2\pi}}(e^{-\frac{x^2}{2\sigma^2}}-e^{-\frac{x^2}{2s^2}})+\frac{e^{-\frac{x^2}{2s^2}}}{\sqrt{2\pi}}(\frac{s-\sigma}{\sigma s})\\ &= \frac{1}{\sigma\sqrt{2\pi}}(e^{-\frac{x^2}{2\sigma^2}}-e^{-\frac{x^2}{2s^2}})+\frac{e^{-\frac{x^2}{2s^2}}}{\sqrt{2\pi}}(\frac{s^2-\sigma^2}{\sigma s(s+\sigma)}).\end{aligned}$$

El primer término tiende a cero por la continuidad de la función $e^{-\frac{x^2}{2\sigma^2}}$ con respecto a la variable σ^2 y al hecho de que $\sigma^2-s^2\to 0$ c.s. Para mostrar que el segundo también tiende a cero basta ver que como $\sigma^2>0$ entonces el denominador converge c.s. a $2\sigma^3$ y el numerador tiende a cero c.s. Concluimos entonces que $g(\sigma^2)-g(s^2)\to 0$

c.s. Pero además como $\sqrt{n}(\sigma^2 - s^2)$ verifica un TCL, entonces está acotada en probabilidad y por consiguiente $\sqrt{nh}(\sigma^2 - s^2) \rightarrow 0$ en probabilidad. Como consecuencia usando este resultado y el desarrollo de Taylor de primer orden de $e^{-\frac{x^2}{2s^2}}$ alrededor de σ^2 se tiene que

$$\sqrt{nh}(g(\sigma^2) - g(s^2)) \rightarrow 0, \tag{D.2}$$

en probabilidad. Entonces

$$\sqrt{nh}(f_n(x) - \varphi_{0,s^2}) = \sqrt{nh}(f_n(x) - \varphi_{0,\sigma^2}) + \sqrt{nh}(g(\sigma^2) - g(s^2)). \tag{D.3}$$

Pero el primer término tiende en distribución a $N(0, \sigma^2(x))$ y el segundo a cero en probabilidad lo cual implica que

$$\sqrt{nh}(f_n(x) - \varphi_{0,s^2}) \xrightarrow{d} N(0, \sigma^2(x)). \tag{D.4}$$

Hemos hecho la demostración en el caso de la ventana h suboptimal, para la ventana óptima la misma demostración vale.

Bibliografía

[1] Ango Nze, Doukhan, P.(1996) *Non-parametric Minimax estimation in a weakly dependent framework I: Quadratic properties.* Math. Meth. Statist. 5 (4), 404-423.

[2] Athanassoulis, GA., Vranas, PB. y Soukissian, TH., (1992). *A new model for leng-term stochastic analysis and prediction. I: Theoretical Background*, Journal of Ship Research, 36(1), pp. 1-16.

[3] Basseville, M. y Nikiforov, N. (1993). *The Detection of Abrupt Change-Theory and Applications*. Englewood Cliffs, NJ: Prentice Hall.

[4] Brockwell, P.J., y Davis, R.A., (1996).*Introduction to Time Series and Forecasting* Springer-Verlag, New York, New York, p.24.

[5] Brodsky, B. and Darkhovsky, B. (1993) *Nonparametric Methods in Change-Point Problems*. Dordrecht: Kluwer Academic Pub.

[6] Charbonnier, S. (2005). *On line extraction of temporal episodes from ICU high-frecuency data: A visual support for signal interpretation*, Computer Methods and Programs in Biomedicine. 78, pp. 115-132.

[7] Coifman, R. y Wickerhauser, M. (1992). *Entropy Based Algorithms for Best Basis Selection*, IEEE Trans. on Information Theory. Vol. 32, pp. 941-981.

[8] Doukhan P. (1995) *Mixing. Properties and Examples.* Lecture Notes in Statistics 85. Springer-Verlag. New York.

[9] Forristall, G. (1998) *On the Statistical Distribution of Wave Heights in a Storm*. J. Geophys. Res. Vol. 83, pp. 2353-2358.

[10] Grigoriu, M. (2009) *Existence and construction of translation models for stationary non-Gaussian processes*. Probabilistic Engineering Mechanics, 24, pp. 545-551.

[11] Guedes Soares, C., Cherneva, Z. y Antão, EM. (2004). *Abnormal Waves During Hurricane Camille*, J. Geophys. Res. Vol. 109, C08008.

[12] Hernández C., J.B., Ortega, J. y Smith, G.H. (2009) *Estudio de los Espectros de Energía usando la Transformada de Hilbert-Huang para la Segmentación de Tormentas dado por el Algoritmo SLEX*. `http://www.cimat.mx:88/~jortega/Publicaciones/HOS1v3.pdf`.

[13] Huang, N.E, y Shen, S. (2005) *Hilbert-Huang Transform and its Applications*. World Scientific Publishing Co. Pte. Ltd.

[14] Huang, H., Ombao, H. y Stoffer, D. (2004) *Discrimination and Classification of Nonstationary Time Series using the SLEX Model*. Journal of the American Statistical Association. Volumen 99, pp. 763-774.

[15] Keogh, E., Chu, S., Hart, D., y Pazzani, M. (2001) *An online algorithm for segmenting time series*, The IEEE International Conference on Data Mining (ICDM)

[16] Labeyrie, J. (1990). *Stationary and transient states for random seas.* Marine Structures, 3-1, pp.43-58.

[17] Lavielle, M. (1998). *Optimal Segmentation of Random Processes,* IEEE Trans. Signal Proc. Vol. 46, No. 5, pp 1365-1373.

[18] Lavielle, M. (1999). *Detection of Multiple Changes in a Sequence of Dependent Variables,* Stochastic Proc. Appl. Vol. 83, pp. 79-102.

[19] Lavielle, M and Ludeña, C (2000). *The Multiple change-Points Problem for the Spectral Distribution*, Bernoulli. Vol. 65. No. 5, pp 845-869.

[20] Lindgren, G. (2006) *Lectures on Stationary Stochastic Processes*. Centrum Scientiarum Mathematicarum. Lund University

[21] Ochi, M.K. (1998) *Ocean Waves, The Stochastic Approach*. Cambridge University Press. Cambridge, UK. pp. 13-57.

[22] Ombao, H., Raz, J., von Sachs, R. and Guo, W. (2002) *The Slex Model of a Non-Stationary Random Process*. Ann. Inst. Statist. Math. Vol 52, No. 1, pp 1-18.

[23] Ombao, H., Raz, J., Von Sachs, R. y Malow, B. (2001) *Automatic Statistical Analysis of Bivariate Non-Stationary Time Series*. Journal of the American Statistical Association, Volume 96, pp. 543-560.

[24] Ombao, H., Von Sachs, R. y Guo, W. (2005) *SLEX Analysis of Multivariate Non-Stationary Time Series*. Journal of the American Statistical Association, Volumen 100, pp. 519-531.

[25] Ombao, H. y Ringo Ho, M. (2006) *Time-dependent Frequency Domain Principal Components Analysis of Multichannel Non-Stationary Signals*. Computational Statistics & Data Analysis. Elsevier Science. Volumen 50, pp. 2339–2360.

[26] Ombao, H., Raz, J., Strawderman, R. y Von Sachs, R. (2001) *A Simple Generalised Cross Validation Method of Span Selection for Periodogram Smoothing*. Biometrika, Volumen 88. pp. 1186-1192.

[27] Ortega, J. y Hernández C., J.B., (2006) *A Comparison of Two Methods for Spectral Analysis of Waves*. Proceeding of the Sixteenth (2006) International Offshore and Polar Engineering Conference, San Francisco, California, USA. pp 45-52.

[28] Ortega, J. y Smith, G.H. (2007) *Spectral Analysis of Storm Waves Using the Hilbert-Huang Transfrom*. Proceedings of the Seventeenth (2007) International Offshore and Polar Engineering Conference. Vol. 3. 1830-1835.

[29] Repko, A., Van Gelder, P.H.A.J.M., Voortman, H.G. y Vrijling, J.K. (2000) *Bivariate statistical analysis of wave climates*. Proceedings of the 27th International Conference on Coastal Engineering (ICCE 2000). pp. 583-596.

[30] Robinson, P.M. (1983) *Non parametric estimators for time series.* J. Time Ser. Anal. 4 (3), 185-207.

[31] Rosemblantt, M. (1991) *Stochastic Curve Estimation*. NSF-CBMS Regional Conference Series in Probability and Statistics. Vol 3, pp 54-60.

[32] Soukissian, T.H. y Samalekos, P.E. (2006) *Analysis of the Duration and Intensity of Sea States Using Segmentation of Significant Wave Height Time Series.* Proc. ISOPE 2006, Vol. 3, pp. 107-113.

[33] Wickerhauser, M.V. (1994) *Adapted Wavelet Analysis From Theory to Software*. IEEE Press, New York.

[34] (2000) *WAFO a Matlab Toolbox of Random Waves and Loads* Tutorial Version 2.0.02. Lund Institute of Technology, Suecia. pp. 6-27.