

## Análisis de datos hidrológicos y procesos de memoria larga

Jorge Graneri

Tesis de Maestría en Ingeniería Matemática

Facultad de Ingeniería - Universidad de la República

2014

Director Académico: Prof. Dr. Franco Robledo Amoza

Director de Tesis: Prof. Dr. Juan Kalemkerian

Tribunal:

Prof. Dr. Gonzalo Perera (Presidente)

Prof. Ing. Enrique Cabaña

Prof. Dr. Rafael Terra

Prof. Dr. José Cataldo

# Análisis de datos hidrológicos y procesos de memoria larga

Jorge Graneri

A mi madre, a la memoria de mi padre. Sin ellos nada hubiera sido posible... A la memoria del "Flaco" Pérez 

Si (como afirma el griego en el Cratilo) el nombre es arquetipo de la cosa, en las letras de 'rosa' está la rosa y todo el Nilo en la palabra 'Nilo'...

Jorge Luis Borges: El Golem

# Índice general

1.	Una	Aproximación naïf al problema	<b>17</b>			
	1.1.	Introducción	19			
	1.2.	Primera aproximación	23			
		1.2.1. Modificaciones aplicadas al modelo autorregresivo:	27			
	1.3.	Segunda aproximación	29			
	1.4.	Algunas consideraciones adicionales al primer tratamiento del problema	31			
		1.4.1. Modelos markovianos	31			
		1.4.2. Vecinos más cercanos	32			
	1.5.	Conclusiones parciales	32			
2.	cesos de memoria larga	<b>35</b>				
	2.1.	Introducción	37			
	2.2.	El estadístico $R$	37			
	2.3.	Un cálculo à l'ancienne	39			
	2.4.	El Estadístico $R/S$	45			
	2.5.	Procesos Autosimilares y Movimiento Browniano Fraccionario	47			
	2.6.	El concepto de memoria larga	50			
	2.7.	Estimación de $H$	52			
		2.7.1. Estimación a partir del estadístico $R/S$	52			
		2.7.2. Método de la varianza agregada	55			
		2.7.3. Método del Periodograma	56			
	2.8.	Análisis de algunas series de datos tomadas de la naturaleza				
		2.8.1. Aportes complexivos de los ríos Uruguay y Negro	57			
		2.8.2. Manchas solares	59			
		2.8.3. Anillos de crecimiento de los árboles	61			
		2.8.4. Velocidad del viento	63			
3.	Modelación con procesos Farima					
	3.1.	Consideraciones previas	69			
	3.2.	Procesos Farima	70			
	3.3.	Procesamiento de la serie de datos hidrológicos	74			
4.	. Conclusiones					
5.	. Apéndices					

ÍNDICE GENERAL

10

## Agradecimientos

Muchas son las personas a las que tengo que agradecer en esta instancia. La lista sería interminable y es por ello que me limitaré a agradecer a quienes, de una forma u otra, han tenido una incidencia directa en este trabajo. Advierto que abusaré de las palabras: amigo, apoyo y generosidad, es una suerte necesitarlas en tal grado en estas líneas.

Agradezco en primer lugar a mis amigos Franco Robledo y Gonzalo Perera. Desde mi ingreso al IMERL he recibido de ellos todo el apoyo del mundo, en particular en esta última instancia en la que el mismo fue decisivo. Es bueno saber que en el mundo académico, que de tanto en tanto muestra su cara hostil, uno cuenta con su confianza y su cordial amistad. ("cordial" del latín cor, cordis (corazón)).

También he contado con el apoyo de mis amigos Raúl Ures, Jana Rodríguez Hertz y Marcelo Lanzilotta, ello ha sido muy importante en instancias de la vida académicas que para mí han sido difíciles y son muy reconfortantes el respaldo y la amistad que me brindan.

Mi amigo Marco Scavino ha sido de gran importancia en este proceso, su claridad y generosidad al comenzar este trabajo y su apoyo a mi proyecto de maestría han sido muy importantes para mí.

Durante el desarrollo del trabajo, tanto en el proyecto, como en la redacción de esta tesis, he tenido el privilegio de contar con docentes y colegas del IMFIA que mucho me han ayudado: Rafael Terra, Álvaro Díaz, Gabriel Pisciottano y Gabriel Cazes, a quienes mucho agradezco.

Agradezco en particular a José Cataldo y Valeria Durañona del IMFIA, que me han ayudado a conseguir los datos de vientos del Puerto de Montevideo y la autorización formal a trabajar con ellos. Agradezco muy en particular a Valeria por las molestias que se ha tomado a este respecto.

A mi amiga Joseline Cortazzo por las muchas consultas bibliográficas en las que me ha ayudado ahora y siempre y a mis amigas Maryori Guillemet, Jimena Rodríguez y Sandra Fleitas por su amistad y su ayuda en todos estos años.

Mi amiga Elisa Rocha, que recientamente presentó un excelente trabajo monográfico sobre estimación en procesos ARMA y con quien he trabajando en tantos cursos ha sido importante en este proceso, siempre aprecio sus puntos de vista y sus sugerencias. En esta instancia, agradezco en particular la ayuda proporcionada en el uso del programa R. Mis queridos amigos Gustavo Guerberoff y Álvaro Moraes siempre me han acompañado tanto en la academia como en la vida y me siento muy afortunado de poder contar con ellos. Agradezco en particular a Álvaro por su ayuda en diversos aspectos informáticos que han resultado de mucha importancia. Agradezco también a mis queridos amigos Alejandro Presto y Daniel Tejerina.

Mi amigo y director de tesis Juan Kalemkerian ha hecho un paciente y diligente trabajo de orientación, ha revisado innumerables versiones de los capítulos, ha respondido mis dudas transmitiéndome seguridad y calma y ha soportado sesiones de trabajo en las que invadí su casa, su computadora y su refrigerador. Su apoyo a lo largo de tantos años y en particular en esta instancia es un gran privilegio que mucho agradezco.

Finalmente, Enrique Cabaña, a quien considero mi padre académico, me ha acompañado desde siempre y en particular en esta instancia con la generosidad y la dedicación de un maestro... un gran maestro. Él es el culpable de que a pesar de los momentos de flaqueza que a veces sobrevienen en la vida académica yo haya optado por seguir el camino. Es un privilegio el poder considerarlo además de todo, un amigo.

# Introducción

Mi primer encuentro con los datos hidrológicos de los ríos Negro y Uruguay fue hace 18 años. Yo, que en aquel entonces era un joven estudiante (ahora soy un viejo estudiante), trabajaba en el Laboratorio de Estadística del Centro de Matemática junto al querido Gonzalo Pérez Iribarren: el "Flaco" Pérez, para los amigos. Luego de un estudio exploratorio bastante intenso de los datos se hizo evidente el hecho de que los promedios móviles de los aportes de esos ríos, al igual que los del Paraná y el Paraguay mostraban un incremento luego de la década del 60. Dicha tendencia puede verse en el siguiente gráfico.



Figura 1 Los promedios móviles de 30 años de los ríos Negro, Uruguay, Paraguay y Paraná calculados en dos períodos: 1902-1940 y 1970-1993.

Ese hecho y otros que fueron producto del trabajo de un equipo interdisciplinario que ya en esos años mostraba la colaboración entre el IMFIA y el Laboratorio que años después pasó a ser el LPE, fue documentado en el artículo de Genta, Pérez-Iribarren y Mechoso (1998) que se cita en la bibliografía. Recuerdo que José Luis Genta, Gabriel Pisciottano, Álvaro Díaz, Gabriel Cazes y yo invadíamos la pequeña oficina de Gonzalo en el segundo piso del local de la calle Eduardo Acevedo, no faltaban el mate ni la disposición a trabajar y conversar de forma amena y cordial.

Unos cuantos años pasaron hasta que en octubre de 2010 tuve la suerte de volver a los temas hidrológicos, esta vez dentro del contexto del Proyecto ANII: "Mejoras en la simulación de aportes a las represas hidroeléctricas para su incorporación a modelos de planificación energética", dirigido por Rafael Terra. Dicho proyecto se extendió del 2010 al 2012. El trabajo del equipo del LPE integrado por: Marco Scavino, Juan Kalemkerian, Pablo Romero y yo consisitió en la modelación estadística de los aportes con el objetivo de simular series sintéticas que reprodujeran las propiedades estadísticas de los aportes de los ríos Uruguay y Negro. El trabajo fue arduo; desde el principio Marco Scavino propuso trabajar con procesos Farima. Desde el principio se dio naturalmente una división de trabajo que implicó que yo hiciera un estudio exploratorio intenso de las series en la predicción a corto plazo de aportes y en la generación de series sintéticas que usara las herramientas más inmediatas (la "artillería ligera": modelos autorregresivos, modelos markovianos simples, vecinos más cercanos) mientras Juan trabajaba los modelos Farima, de mayor complejidad.

Los modelos autorregresivos (que también son usados por el simulador de la plataforma SimSEE -ver Chaer 2006 y 2008) y los markovianos (utilizados por el programa EDF de UTE) son buenos predictores a corto plazo; mientras que otras propiedades de la serie, como la de memoria larga (que surge del interés de modelar los eventos extremos en períodos relativamente largos de tiempo) requieren otro tipo de modelos. Los movimientos brownianos fraccionarios propuestos por Mandelbrot en la década del 60 tienen dicha propiedad; pero los procesos Farima tienen la ventaja (al tener más parámetros) de captar tanto las características a corto plazo de los aportes (gracias a los parámetros del proceso ARMA asociado al Farima, algo que aumenta un poco el nivel de complejidad con respecto a los modelos más elementales), como así también las propiedades de memoria larga (el coeficiente de Hurst está presente en el parámetro d de diferenciación fraccionaria del modelo).

- Este trabajo de tesis consta de tres capítulos:
  - En el primero se describe el trabajo con los modelos más sencillos. Dicho trabajo fue muy intenso y requirió mucho esfuerzo de programación y tiempo de cómputo. Sin embargo, he preferido no bombardear al lector con los detalles de ese primer estudio exploratorio, para privilegiar la discusión detallada de lo hecho con los modelos Farima, que se presenta en el capítulo 3.
  - En el segundo capítulo se discute el concepto de memoria larga, su origen en los trabajos de Hurst con las series históricas del río Nilo, la discusión matemática posterior, en la que se destaca el trabajo de Mandelbrot y la estimación, a modo de ejemplo, del coeficiente de Hurst en cuatro series de datos tomados de la naturaleza.
  - En el tercer capítulo se discute la aplicación de los modelos Farima para la generación de series sintéticas. Dicho trabajo se basa en el de Juan Kalemkerian (ver Kalemkerian (2012)) si bien tiene diferencias en algunos aspectos del algoritmo.

Para terminar, debo decir que ha sido muy motivador sumergirme (nunca mejor empleado el término) en este trabajo, que no termina con esta tesis, sino que recién parece estar comenzando...



El joven Gonzalo Pérez Iribarren.

ÍNDICE GENERAL

16

Capítulo 1

# Una Aproximación naïf al problema

#### 1.1. Introducción

En este capítulo se describe mi trabajo, dentro del grupo del LPE en el Proyecto ANII: "Mejoras en la simulación de aportes a las represas hidroeléctricas para su incorporación a modelos de planificación energética" (proyecto conjunto del IMFIA y el LPE, que se desarrolló entre fines del 2010 y fines del 2012). Se trata de un intenso trabajo exploratorio sobre la modelación de aportes hidrológicos. La intención de este capítulo es describir en líneas generales lo que se hizo; para luego retomar la discusión sobre este problema concreto, con más precisión en el capítulo 3.

El trabajo que se describe en este capítulo tiene como objetivo modelar los datos hidrológicos disponibles: la serie de aportes complexivos de los ríos Uruguay y Negro, medidos en los lugares donde están ubicadas las represas de Salto Grande, Rincón del Bonete y Palmar entre 1909 y 2009. Dichos aportes resultan de una ponderación de los aportes hidrológicos registrados en Salto Grande, Rincón del Bonete y Palmar de acuerdo al rendimiento energético de los mismos. En concreto se tiene que el aporte complexivo en un tiempo t viene dado por la fímula:

$$X_t = \frac{115}{1010} X_t^{(s)} + \frac{618}{1010} X_t^{(b)} + \frac{277}{1010} X_t^{(p)},$$

donde  $X_t^{(s)},\ X_t^{(b)}$  y  $X_t^{(p)}$  son los aportes en el tiempo t de Salto Grande, Bonete y Palmar, respectivamente.



Figura 2 Represa de Salto Grande.



**Figura 3** Represa de Rincón del Bonete.



**Figura 4** Represa de Palmar.

#### 1.1. INTRODUCCIÓN

La figura siguiente muestra dicha serie.



Figura 5 Serie de aportes complexivos de los ríos Uruguay y Negro.

las siguientes figuras muestran respectivamente los promedios y los desvíos según la semana de la serie de datos.



Figura 6 Promedios según la semana de la serie de aportes complexivos.



Figura 7 Desvíos según la semana de la serie de aportes complexivos.

Finalmente, la siguiente figura muestra la serie de datos desestacionalizada, es decir la serie de datos a la que se ha restado el promedio según la semana y que en cada semana ha sido dividida por el desvío según la semana.



Figura 8

Serie de aportes complexivos desestacionalizados de los ríos Uruguay y Negro.

A modo exploratorio se considerará la serie de datos estadísticos de El Niño (índice N3.4) a los efectos de evaluar su influencia en la predicción a corto plazo de los aportes complexivos y en la generación de series sintéticas.

El objetivo en una primera instancia (el trabajo hecho en este capítulo) es modelar la serie de aportes tanto a corto plazo como en la generación de series sintéticas usando las herramientas y los modelos más sencillos. A tales efectos se toman como base los modelos simuladores y predictores de los modelos de optimización usados para despacho y planificación energética.

#### 1.2. PRIMERA APROXIMACIÓN

En particular, el modelo EDF usado por UTE es un modelo markoviano cuyo espacio de 5 estados se define a partir de un promedio móvil de 12 semanas (índice ESHY). Por su parte, el modelo CEGH, de la plataforma SimSEE también utilizada por UTE (ver Chaer 2008) es un modelo autorregresivo multivariado, que no utiliza los aportes complexivos sino un proceso multivariado de dimensión 3 (los aportes en Bonete, Palmar y Salto respectivamente).

En esta primera aproximación al problema se han considerado tanto modelos autorregresivos (al igual que el CEGH) que en alguno de los intentos han sido hibridados con modelos markovianos (en forma similar al modelo de simulación del EDF) a los efectos de reproducir estadísticamente mínimos coherentes con los mínimos históricos observados.

Teniendo en cuenta el carácter exploratorio de esta primera etapa, que entre otras cosas ha servido para evaluar la dificultad del problema, los modelos no serán descritos en detalle, a diferencia de la aproximación mucho más elaborada que se presenta en el capítulo 3. Pero sí es justo dejar constancia de que lo descrito en esta sección es producto del trabajo de muchos meses y de múltiples intentos de abordar un problema que por su naturaleza es complejo y escurridizo.

#### 1.2. Primera aproximación

Se partió de la muestra histórica de 100 años de los aportes complexivos de Salto, Bonete y Palmar (sin desestacionalizar). En el primer caso se considera, para el aporte complexivo  $X_i$ , un modelo dado por la fórmula:

$$X_{i} = \sum_{j=1}^{4} \alpha_{i-j} \cdot X_{i-j} + \sum_{j=1}^{40} \beta_{i-j} \cdot Y_{i-j} + \varepsilon_{j}.$$

donde  $X_j$  representa el aporte en la semana j,  $Y_j$  representa la anomalía de El Niño en la semana j y  $\varepsilon_j$  es un error cuya distribución fue estimada a partir de los errores en la parte de la serie histórica reservada para tales fines.

De modo que tenemos un modelo lineal que responde al siguiente esquema:



Figura 9 Modelo utilizado.

En un principio se partió del modelo más general:

$$X_{i} = \sum_{j=1}^{k} \psi_{j} \cdot i^{j} + \sum_{j=1}^{n} \alpha_{i-j} \cdot X_{i-j} + \sum_{j=1}^{m} \beta_{i-j} \cdot Y_{i-j} + \varepsilon_{j}$$

Tempranamente se descartó la parte correspondiente a la línea de tiempo  $\sum_{j=1}^{k} \psi_j . i^j$  y teniendo en cuenta criterios de minimización del error cometido por distintos modelos (en los que se variaba los índices  $n \ge m$ ) se llegó a la conclusión de que eran óptimos los valores  $n = 4 \ge m = 40$ , hecho que arroja como resultado el modelo concreto presentado en la primera fórmula.



#### Figura 10

Predicción a corto plazo. En el gráfico se muestra el resultado de la predicción a corto plazo usando los datos de las cuatro semanas previas y las anomalías de El Niño de las 40 semanas previas. El modelo fue ajustado con los 81 primeros años de la serie y puesto a prueba, como se muestra en este dibujo, con los últimos 10 años. La línea azul indica los datos reales y la roja las predicciones.

Tratamiento de errores

Si se aplica a los errores observados  $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \dots, \hat{\varepsilon}_n$  la prueba de correlación de rangos de Spearman, a los efectos de estudiar su aletoriedad, se obtiene un valor del estadístico de S = 0,0027 que tiene un p-valor de 0,4615; hecho que, además de aceptar la hipótesis de aleatoriedad, muestra claramente que no hay dependencia de los errores con respecto al tiempo.



Figura 11 Nube de errores del modelo autorregresivo según la semana del año.

Si se aplica a los errores observados la prueba de normalidad de Lilliefors, basada en el estadístico:

$$KSL = \sup_{x \in \mathbb{R}} \left| F_n(x) - \Phi(\frac{x - \bar{\varepsilon}}{s_n}) \right|,$$

(donde  $\Phi$  es la función de distribución normal típica y  $\bar{\varepsilon}$  y  $s_n$  son los estimadores habituales de la media y el desvío de la distribución de la muestra de errores, respectivamente y  $F_n$  la función de distribución empírica de la serie estandarizada  $\frac{\hat{\varepsilon}_i - \bar{\varepsilon}}{s_n}$ ) se obtiene un valor del estadístico de S = 0, 1209. Si se compara este valor con el valor crítico al nivel 0,05, que es 0,0248 se concluye que los datos no son normales.

La densidad de esa muestra, que como vimos no es gaussiana, se estima mediante la convolución con núcleos gaussianos (este método es bastante general y por supuesto, no es necesario que la muestra sea gaussiana, en cuyo caso, el problema de la estimación de la densidad sería trivialmente resoluble) obteniéndose la función:

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{nk} \varphi\left(\frac{x - \hat{\varepsilon}_i}{k}\right)$$

donde  $\varphi$  es la función de densidad normal típica y  $k = \frac{1}{\log(\log(n))}$ . Esta densidad nos servirá para generar el ruido de nuestras series sintéticas.





Comparación de la función de distribución empírica de la muestra de errores observados estandarizada (azul) con la función de distribución normal estándar (rojo). En la gráfica se aprecia que la muestra no es normal, si lo fuera (para una muestra de ese tamaño) las curvas deberían estar prácticamente superpuestas.

#### 1.2.1. Modificaciones aplicadas al modelo autorregresivo:

Varias fueron las modificaciones aplicadas al modelo autorregresivo descrito, hechas con el objetivo de mejorar el ajuste de diversos aspectos (en particular de los cocientes de percentiles con los que se estudian las sequías). Entre ellas debemos citar, sin entrar en los detalles de cada modelo:

- 1. Estimación de errores según la semana del año e implementación en la simulación: se parte la muestra de errores en 52 muestras, según la semana, para una estimación más precisa de los errores estadísticos cometidos por el modelo.
- 2. Acotación del error con respecto a la predicción: se observó en algunos casos (y eso es un hecho inevitable asociado a éste y a otros modelos) un error, generado aletoriamente que arrojaba resultados incoherentes con las variables que se quería generar (p. ej. datos negativos). Esos casos fueron corregidos con distintos criterios estadísticos, que no mejoraron los resultados (en lo referente a las curvas de intensidad).
- 3. Iteración de la simulación, nueva estimación de los errores en cada iteración. Estas iteraciones tenían como objetivo mejorar la calidad de los errores estadísticos.
- 4. Corrección de las colas simulando en las franjas inferiores. Teniendo en cuenta que las colas inferiores de la series generadas eran de crucial importancia para el estudio y que, debido a los pocos datos disponibles (los mínimos históricos son pocos) se trató de mejorar la predicción simulando aparte y con diversos criterios, las colas inferiores. En particular se realizó una corrección de las colas según la comparación de las distribuciones empíricas de los datos simulados y los de la serie histórica.



#### Figura 13

Una posible transformación de los datos generados a partir de las distribuciones empíricas: modelo lineal (azul), serie histórica (rojo). Si  $\{Y_i\}$  es la muestra generada (cuya FDE  $F_n^Y$  está graficada en azul) se considera el valor  $F_n^Y(Y_i)$  (el punto en el que la vertical trazada en la abscisa  $Y_i$  corta la línea azul y se la asigna el valor  $F_n^{-1}(F_n^Y(Y_i))$ , que equivale geométricamente a trazar una horizontal desde el punto mencionado hasta cortar la línea roja, de la función de distribución empírica de la muestra  $\{X_i\}$  y luego proyectar sobre el eje de las abscisas.

Es de hacer notar que, cada uno de los puntos mencionados implicó la realización de decenas de simulaciones considerablemente costosas desde el punto de vista computacional.

La mayoría de las modificaciones mencionadas no resultaron en una mejora a la calidad de los programas generadores. Sin embargo, el último punto mencionado en la descripción anterior, es decir, una corrección de colas inferiores y superiores, dio lugar a una mejora en los resultados de predicción. Se muestra en la siguiente figura, la gráfica de estudio de sequías correspondiente a ese modelo.





Curvas de intensidad de sequías. Las bandas de confianza (0,25-0,75 y 0,10-0,90) para 1000 series sintéticas generadas según el algoritmo precedente muestran respectivamente los cocientes del mínimo, y los percentiles 5 %, 10 % y 25 % con la mediana, para distintos períodos de tiempo.

#### 1.3. Segunda aproximación

En una segunda aproximación se divide el recorrido de la serie histórica en franjas de percentiles y se permite, en una primera instancia avanzar al modelo lineal en ese recorrido, corrigiendo cada valor predicho por el modelo según la franja en la que se encuentra.



Figura 15 Franjas de percentiles de la serie histórica (en este caso se representan los percentiles: 2, 3, 5, 8, 15, 25, 55 y 75).

$\alpha$	$p_{\alpha}$				
0,02	75,07				
0,03	84,25				
0,05	99, 55				
0,08	122, 59				
0, 15	176, 58				
0, 25	277, 42				
0,55	782, 28				
0,75	1276, 15				

#### tabla de percentiles de la serie histórica, en la que se basa la división en franjas.



#### Figura 16

Corrección en el modelo híbrido: se hace funcionar el programa según el modelo autorregresivo; pero el último valor se corrige teniendo en cuenta en qué franja cae el dato correspondiente.

El objetivo de esta aproximación es mejorar la predicción general, mejorando la predicción dentro de cada franja. Una vez que el modelo lineal descrito en la parte anterior genera un dato, el mismo es remuestreado según la franja en la que se encuentre. Esto puede hacerse de diversas maneras, entre las que citamos:

#### Posibles correcciones dentro del modelo híbrido:

- 1. Remuestreando los datos de la serie histórica en las distintas franjas.
- 2. Estimando la densidad en cada franja y simulando datos con cada una de las densidades.
- 3. Transformando según la comparación de las funciones de distribución empírica.

4. Transformando según la comparación de las funciones de distribución empírica y comprimiendo las franjas inferiores.

Las correcciones planteadas en las puntos 1, 2 y 4 fueron implementadas en decenas de simulaciones algunas de las cuales resultaron en una mejora. Se muestra a continuación el mejor resultado obtenido siguiendo la especificación del punto 4.



Figura 17 Representación de los cocientes de percentiles para el estudio de las sequías con el modelo híbrido 1.

#### 1.4. Algunas consideraciones adicionales al primer tratamiento del problema

#### 1.4.1. Modelos markovianos

Es lícito preguntarse acerca de la dificultad del problema: ¿por qué es tan difícil reproducir el comportamiento de los mínimos? En primer lugar: porque hay pocos y los sintetizadores mencionados se basan en la información disponible. Dentro de un enfoque naïf ¿existe una mejor manera de usar la información disponible? ¿Es posible mejorar la aproximación con un modelo markoviano? La tabla siguiente desalienta esa idea:

$\alpha$	$p_{11}$ (hist.)	$p_{11}$ (lineal)	$p_{11}$ (híbrido)
0,01	0,0067	0,0425	0,0868
0,05	0.0323	0.0721	0.1261
0,10	0.0696	0.0977	0.1798

tabla de probabilidades de transición para la primera franja en cada uno de los modelos

El hecho de que las probabilidades de transición en cuestión sean mayores en los dos modelos propuestos que en la serie histórica desalienta la idea de que el modelado de los mínimos se arregle aumentando estas probabiliades y pone en evidencia que hay una estructura que se está escapando a los modelos.

#### 1.4.2. Vecinos más cercanos

Entre los programas considerados y redactados en un primer momento se encuentra el de vecinos más cercanos. El programa consiste en encontrar (tomando una cierta ventana de tiempo como referencia) las trayectorias más parecidas a la observada, digamos en las semanas previas a la predicción y usar los valores de esa serie para predecir (por ejemplo, promediando, tomando la mediana, etc.) el valor correspondiente. Sin embargo, si ponermos el énfasis en la predicción de los mínimos (primera curva de intensidad ) no conviene considerar un programa de vecinos más cercanos ya que, por muy bueno que sea, en lo referente a los mínimos no se va a obtener nada peor que los mínimos históricos.

#### 1.5. Conclusiones parciales

Como conclusiones de esta primera aproximación al problema podemos mencionar que la simulación de los mínimos de series sintéticas de datos hidrológicos es un porblema difícil. La complejidad del problema es consecuencia, por un lado de la poca información estadística que se tiene sobre los valores extremos (si bien las series abarcan más de un siglo de registros, dicho período no presenta suficientes eventos extremos, en este caso mínimos, para entender su comportamiento estadístico). Por otra lado, los modelos elementales que se usan: modelos autorregresivos, modelos markovianos y modelos híbridos entre unos y otros, subestiman la complejidad del comportamiento estadístico de las series de aportes. Habrá que recurrir a modelos más complejos. En una aproximación más profunda al problema que se presentará en el capítulo 3, se considerarán modelos ARMA y sus transformados por integración fraccionaria: los modelos FARIMA. Pero antes de ellos nos ocuparemos de un aspecto importante de la modelación de series de datos hidrológicos, que surgió cuando Harold E. Hurst se dio cuenta, trabajando a principios del siglo XX con la serie de aportes históricos del río Nilo, de que el tratamiento que se estaba haciendo de los mismos hasta entonces, también subestimaba la complejidad del problema. En la sección siguiente trataremos el concepto de memoria larga en series de datos estadísticos.



Ese río, que atravesaba por primera vez, y que iba a ser mi horizonte y mi hogar durante diez años, viene del norte, de la selva, y va a morir en el mar que el pobre capitán llamó dulce. Ellos lo llaman padre de ríos. Y es verdad que, mientras viene bajando, engendra ríos a su paso, ríos que van multiplicándose en las proximidades de la desembocadura, que se separan a determinada altura del lecho principal, corren unas leguas paralelos a él, y vuelven a reunírsele un poco más abajo, ríos que a su vez engendran ríos que engendran otros a su vez, con esa tendencia a la multiplicación infinita que frenan a duras penas las barrancas comidas por el agua; río de muchas orillas, a causa de las islas sombrías y pantanosas que las forman. Los hombres que habitan en las inmediaciones tienen el color del barro de la costa, como si también ellos hubiesen sido engendrados por el río, lo que haría decir al padre Quesada años más tarde, cuando oiría mis descripciones, que yo había vivido durante diez años, sin darme cuenta, en la vecindad del paraíso, que en la carne de esos hombres había todavía vestigios del barro del primero, que esos hombres eran sin duda la descendencia putativa de Adán.

Juan José Saer: El entenado.

Capítulo 2

Procesos de memoria larga
# 2.1. Introducción

El objetivo de esta sección es considerar el problema de la modelación de aportes hidrológicos desde un punto de vista diferente y más elaborado que se basa esencialmente en los trabajos de Harold Edwin Hurst, quien llegó a El Cairo en 1906 siendo un joven físico, con la intención de quedarse unos meses para estudiar determinados aspectos hidrológicos del río Nilo. Hurst se terminó quedanndo en Egipto 62 años durante los cuales estudió con detalle el río Nilo, que recorrió en varias ocasiones en toda su extensión. Un problema que lo obsedía, entre otros, estaba en su forma estadística ligado a los fenómenos naturales que Mandelbrot años después llamó El Efecto de Noé, a saber: que las precipitaciones extremas pueden ser realmente extremas y El*Efecto de José*, a saber: que los tiempos extremos (de sequía y de fertilidad, de escasez de precipitaciones y de precipitaciones fuertes) podrían ser inusualmente largos. Hurst partió del problema concreto de construir un embalse de dimensiones tales que compensara la irregularidad del flujo de un río (en este caso el Nilo), con un flujo continuo por un período largo y predeterminado (en este caso 100 años). Estudiando ese problema Hurst se dio cuenta de que los modelos estadísticos considerados hasta ese momento para los aportes de los ríos subestimaban sensiblemente la complejidad de los fluctuaciones hidrológicas. Hurst propuso entonces un modelo más adecuado que involucraba un coeficiente de autosimilitud (el hoy llamado coeficiente de Hurst: H) íntimamente ligado al fenómeno estadístico de la memoria larga. Además de trabajar con datos de ríos, Hurst mostró que un comportamiento similar al anterior se daba en los registros de otros fenómenos naturales como precipitaciones, temperatura, presión, grosor de los anillos de ciertos árboles, grosor de las capas de barro en el lago Saki en Crimea, manchas solares e incluso fenómenos bursátiles, como el valor de rescate de ciertas acciones.

# **2.2.** El estadístico R

Tal como se mencionó, Hurst estaba interesado en el problema de la construcción de una represa. En concreto la pregunta es, cuán alta debería ser la misma para romper (con su flujo continuo) la irregularidad de los aportes por un período largo y determinado, digamos 100 años. Planteado in abstracto, la pregunta es: ¿ cuál debería ser la capacidad del embalse  $R(\delta)$  para asegurar una preformance ideal durante  $\delta$  años?, en donde por "performance ideal" se pretende que:

- 1. El flujo sea uniforme.
- 2. El embalse termine el período tan lleno como lo empezó.
- 3. La represa nunca se desborde.
- 4. La capacidad sea la mínima compatible con 1, 2 y 3.

Consideremos los aportes acumulados de un río de un período  $\delta$ , representados en la siguiente figura, en la que la línea recta representa un flujo acumulado igual a la media de los aportes.



figura 18

Si ahora consideramos la diferencia entre los aportes acumulados y el buscado flujo continuo acumulado dado por la media de los aportes.



figura 19

No será difícil ver que el volumen mínimo del embalse buscado es la diferencia entre el máximo M y el mínimo m: ya que el tomar esa diferencia nos asegura que partimos de la cantidad de agua necesaria para que el flujo sea continuo e igual a la media, esa dos propiedades del flujo nos aseguran que el volumen del embalse será el mismo al principio y al fin del período y, si los cálculos se hacen en un lapso suficientemente largo, podemos estar (desde un punto de vista estadístico) confiados de que la represa no se desbordará.

Estudiando las propiedades de dicho estadístico es que Hurst se topa con su coeficiente y con la propiedad de memoria larga.

## 2.3. UN CÁLCULO À L'ANCIENNE



figura 20 Harold Edwin Hurst

# 2.3. Un cálculo à l'ancienne

En su artículo de 1951 Hurst hace el cálculo del valor esperado de la diferencia entre el máximo y el mínimo de un paseo al azar. Si asumimos, como se hacía hasta entonces, independencia de los caudales de un año a otro, y teniendo en cuenta que es posible aproximar a un proceso gaussiano de incrementos independientes por paseos al azar, este cálculo nos ayudaría a saber qué se espera de la diferencia de fluctuaciones y en última instancia a conocer el valor esperado del estadístico R.

Hurst parte de un paseo al azar que parte de cero y llega a cero al final del período en cuestión. En su artículo el habla de N tiradas de 2m monedas. Llamando N = nm nosotros vamos a considerar un paseo al azar de 2n pasos que parte del origen y llega al origen, o sea la suma acumulada de variables aleatorias independientes que toman los valores -1 y 1 con igual probabilidad



figura 21

Un paseo al azar de veinte pasos que parte del cero y vuelve al cero.

Hurst divide las trayectorias en tres clases:

- Aquellas en las que el valor del proceso es mayor o igual que cero. En cuyo caso el estadístico buscado es el máximo del proceso: M, lo que él llama la máxima diferencia entre ganancias y pérdidas (G L, G del inglés "gain" y L, del inglés "loss").
- Aquellas en las que el valor del proceso es menor o igual que cero. En cuyo caso el estadístico buscado es el valor absoluto del mínimo del proceso: |m|, lo que él llama la máxima diferencia entre pérdidas y ganancias (L G)
- Aquellas en las que hay valores positivos y negativos. En las que el estadístico buscado es el máximo más el valor absoluto del mínimo:

M + |m| (en la terminología del artículo: la máxima diferencia entre ganancias y pérdidas más la máxima diferencia entre pérdidas y ganancias: máx(G - L) + máx(L - G)).

Siguiendo la notación del artículo, el valor esperado del estadístico en cuestión se puede escribir como:

$$E(R) = \sum_{i} \frac{M_i + |m_i|}{\binom{2n}{n}}$$

En las trayectorias del primer grupo  $m_i = 0$ , en las del segundo grupo  $M_i = 0$ . El número total de trayectorias que parten de 0 y llegan a 0 es  $\binom{2n}{n}$ , o sea la cantidad de formas distintas de ordenar *n* elementos iguales a 1 y *n* elementos iguales a -1. Por la simetría probabilística de los paseos se tiene que

$$\sum_{i} \frac{M_i + |m_i|}{\binom{2n}{n}} = 2\sum_{i} \frac{M_i}{\binom{2n}{n}}.$$

Para este último cálculo será útil tener en cuenta la cantidad de trayectorias que superan una cierta barrera h > 0, a la que llamaremos  $n_h$ . Por el principio reflexión,

por cada trayectoria que alcanza la barrera hy termina en 0, hay una trayectoria especular que termina en 2h.



figura 22 Un paseo al azar de veinte pasos que parte del origen y vuelve al origen y su trayectoria especular con respecto a la barrera h = 2.

De modo que el número de paseos de 2n pasos que parten de 0 y llegan a 0 y además alcanzan la barrera h es igual a la cantidad de paseos al azar de 2n pasos que terminan en 2h, que en consecuencia tienen n + h pasos hacia arriba y n - h pasos hacia abajo y que por lo tanto son  $\binom{2n}{n+h}$ . Por otra parte, la cantidad de trayectorias en que el máximo es exactamente igual a h, a la que llamaremos  $N_h$  cumple:

$$N_h = n_h - n_{h+1} = \begin{pmatrix} 2n \\ n+h \end{pmatrix} - \begin{pmatrix} 2n \\ n+h+1 \end{pmatrix}.$$

Teniendo esto en cuenta, podemos expresar el valor esperado de la siguiente forma:

$$E(R) = 2\sum_{h=1}^{2n} h \frac{N_h}{\left(\begin{array}{c}2n\\n\end{array}\right)} = 2\sum_{h=1}^{2n} h \frac{\left(\begin{array}{c}2n\\n+h\end{array}\right) - \left(\begin{array}{c}2n\\n+h+1\end{array}\right)}{\left(\begin{array}{c}2n\\n\end{array}\right)}.$$

No es difícil ver que se cumple la igualdad

$$\sum_{h=1}^{2n} h\left( \begin{pmatrix} 2n \\ n+h \end{pmatrix} - \begin{pmatrix} 2n \\ n+h+1 \end{pmatrix} \right) = \sum_{h=1}^{2n} \begin{pmatrix} 2n \\ n+h \end{pmatrix}.$$

Por otra parte, teniendo en cuenta que se cumple

$$2^{2n} = (1+1)^{2n} = \sum_{h=0}^{2n} \begin{pmatrix} 2n \\ h \end{pmatrix},$$

es fácil ver que:

$$E(R) = 2\frac{\sum_{h=1}^{2n} \binom{2n}{n+h}}{\binom{2n}{n}} = 2\frac{1}{2}\frac{\binom{2^{2n}-\binom{2n}{n}}{\binom{2n}{n}}}{\binom{2n}{n}} = \frac{2^{2n}}{\binom{2n}{n}} - 1.$$

Teniendo en cuenta ahora la aproximación de Stirling

$$n! \approx \sqrt{2n\pi} \left(\frac{n}{e}\right)^n$$

tenemos

$$\begin{pmatrix} 2n \\ n \end{pmatrix} = \frac{(2n)!}{n! n!} \approx \frac{\sqrt{4n\pi} \left(\frac{2n}{e}\right)^{2n}}{2n\pi \left(\frac{n}{e}\right)^{2n}} = \frac{2^{2n}}{\sqrt{n\pi}}$$
con lo que nos queda

$$E(R) \approx \sqrt{n \pi} = \sqrt{N m \pi}.$$

Hasta aquí se ha hecho el cálculo aproximado del valor esperado del estadístico R en el caso de un paseo que vuelve al origen en 2n pasos. Hurst hace algo ligeramente diferente en las simulaciones, para aproximarse a los datos reales: trabaja con N bloques de 2m tiradas (variables independientes simétricas que toman con la misma probabilidad los valores -1 y 1), de modo que le interesa escribir el valor esperado del estadístico en función de la varianza en cada bloque. Para cada una de esas variables simétricas, se consideran las diferencias entre la cantidad de veces que se observa el 1 y la cantidad de veces que se observa el -1. El desvío del número de unos:  $\sigma_H$  vale:

$$\sigma_H = \sqrt{2m(1/2)(1/2)} = \sqrt{m/2}.$$

Por lo tanto, el desvío de las diferencias  $\sigma$ valdrá:

$$\sigma = 2 \sigma_h = \sqrt{2m},$$

de donde finalmente escribimos

$$E(R) \approx \sigma \sqrt{N \pi/2} = 1.25 \sigma \sqrt{N}.$$

Siguiendo el espíritu del artículo de Hurst, se muestran a continuación los resultados de una simulación de 10 replicaciones de N = 1000 bloques de tamaño 2m = 10, en cada una de ellas se da un estimador  $s_{2m}$  del desvío  $\sigma = \sqrt{10} \approx 3, 16$ , se calcula el estadístico R y el cociente  $R/s_{2m}$ . En la última fila se muestran los promedios de dichas simulaciones.

$s_{2m}$	R	$R/s_m$
3.27	131	1.27
3.03	102.2	1.07
3.08	81.9	0.84
3.28	142.7	1.38
3.19	147.4	1.46
3.22	93.2	0.92
3.12	113	1.14
3.23	136	1.33
3.10	147.3	1.50
3.17	127.3	1.27
3.17	122.2	1.22

42

## 2.3. UN CÁLCULO À L'ANCIENNE

Se muestran a continuación los resultados de una simulación mayor, en la que la cantidad de bloques (2m = 10) se hace variar desde N = 100 hasta N = 5000 (de cien en cien) y para cada uno de esos tamaños se hacen 100 replicaciones. La figura muestra los promedios de los valores del estadístico R en cada caso.



figura 23

En la siguiente figura se grafica en cada caso log(R) contra log(N). Según las consideraciones anteriores, la nube de puntos debería aproximarse por una recta de pendiente 1/2.



figura 24

#### CAPÍTULO 2. PROCESOS DE MEMORIA LARGA

Estos son los cálculos asumiendo el modelo de "tirada de monedas" de Bachelier. La consecuencia fundamental es que el rango varía según la raíz cuadrada del período: por ejemplo, si queremos llevar una presa para un período de 25 años en una presa para un período de 100 años, basta con duplicar su altura. Este tipo de cálculos se tuvieron en cuenta, por ejemplo en la construcción de la primera presa de Asuán en el Nilo en el año 1902 de una altura inadecuada, que fue elevada posteriormente en 1907 y en 1929. Cuando en las crecidas de 1946 la presa corrió el reisgo de desbordarse, el gobierno egipcio decidió la construcción de una segunda presa, de mayor altura, a unos kilómetros de distancia de la primera.



#### figura 25

La presa baja de Asuán: la figura de la derecha fue tomada en el año 1903, poco después de su construcción. La de la izquierda corresponde a la primera elevación, realizada en 1907.

En su trabajo sobre los distintos fenómenos naturales que consideró, Hurst mostró que el exponente a tener en cuenta no era 1/2 sino un cierto H (la notación H en homenaje a Hurst fue propuesta años después por Mandelbrot) que variaba entre 0,69 y 0,80 y que en general, en el tipo de procesos que aquí importan es mayor que 1/2.

44



#### figura 26

Gráfico del artículo de Hurst que muestra los ajustes lineales en los distintas series de tiempo estudiadas por él en su trabajo fundamental.

# **2.4.** El Estadístico R/S

Siguiendo las ideas de la sección anterior definiremos, dado un proceso  $X_1, X_2, \ldots, X_n$ el estadístico R/S

$$R/S(X_1, X_2, \dots, X_n) = \frac{\max_{0 \le i \le n} (S_i - \frac{i}{n} S_n) - \min_{0 \le i \le n} (S_i - \frac{i}{n} S_n)}{\sigma_n}$$

donde  $S_i = \sum_{k=1}^i X_k, S_0 = 0$  y

$$\sigma_n = \left(\frac{1}{n} \sum_{i=1}^n (X_i - \frac{i}{n} S_n))^2\right)^{1/2}$$

¿Qué se puede decir sobre el comportamiento asintótico de este estadístico? Consideremos el espacio D([0,1]) de las funciones cadlag en [0,1], es decir, el espacio de las funciones que toman valores en [0,1], son continuas por derecha y tienen límite finito por la izquierda con la topología inducida por la distancia de Skorohod.

$$d(f,g) = \inf_{\lambda \in \Lambda} \left\{ \sup_{0 \le u \le 1} |f(u) - g(\lambda(u))| \right\},\$$

siendo A la familia de las funciones continuas estrictamente crecientes y sobreyectivas de [0, 1] en [0, 1]. Se cumple que con esta métrica D([0, 1]) es separable y completo. La función

$$f(x)(t) = \sup_{0 \le t \le 1} (x(t) - tx(1)) - \inf_{0 \le t \le 1} (x(t) - tx(1))$$

siendo x una función  $\{x(t)/0 \leq t \leq 1\}$  perteneciente <br/>aD([0,1]),es un funcional continuo.

Dada una sucesión  $X_1, X_2, \ldots, X_n$  equidistribuidas con media  $\mu$  y varianza finita  $\sigma^2$ , consideremos por otro lado la versión D([0, 1]) del proceso de sumas parciales:

$$S^{(n)}(t) = S_{[nt]} - [nt]\mu, \ 0 \le t \le 1$$

En estas condiciones (y condiciones aun más generales) el Principio de Invarianza de Donsker nos asegura que  $\frac{S^{(n)}}{\sqrt{n\sigma}}$  converge en distribución a un proceso de Wiener  $\{W_t\}_{0 \le t \le 1}$  (ver Billingsley 1999).

**Definición.** Un proceso  $\{W_t, t \in \mathbb{R}\}$  se dice un *Movimiento Browniano o Proceso de Wiener* si cumple las siguientes propiedades:

- 1.  $P\{W_0 = 0\} = 1$ .
- 2.  $W_t \sim N(0, t^2), t \in I\!\!R$ .
- 3. Para  $s_1 < t_1 < s_2 < t_2$  las variables  $W_{t_2} W_{s_2}$  y  $W_{t_1} W_{s_1}$  son independientes.

Por otra parte, es fácil ver que:

$$f(S^{(n)}) = \sigma_n R/S(X_1, X_2, \dots, X_n)$$

Por lo tanto, si se verifican las hipótesis del Principio de Invarianza de Donsker, se tendrá que la distribución límite de  $\frac{R/S(X_1, X_2, \dots, X_n)}{\sqrt{n}}$  será igual a la distribución límite de

$$\frac{\sigma}{\sigma_n} \left( \sup_{0 \le t \le 1} (W_t - tW_1) - \inf_{0 \le t \le 1} (W_t - tW_1) \right).$$

En condiciones muy generales se cumple que  $\sigma_n$  tiende a  $\sigma$  con probabilidad 1; por otra parte el proceso  $\{B_t\}_{0 \le t \le 1}$  definido por:

$$B_t = W_t - tW_1,$$

es el llamado Puente Browniano en [0,1] y se puede entonces afirmar que:

$$\frac{R/S(X_1, X_2, \dots, X_n)}{\sqrt{n}} \Rightarrow \sup_{0 \le t \le 1} B_t - \inf_{0 \le t \le 1} B_t$$

donde  $\Rightarrow$  indica convergencia en distribución. De modo que podemos afirmar que si se cumplen las hipótesis del Principio de Invarianza, el estadístico R/S es asintóticamente del orden de  $\sqrt{n}$ . Las observaciones de Hurst sugerían que esto no se cumplía, hecho que llevó a Mandelbrot a concebir un modelo que no cumpliera las hipótesis del Teorema Funcional del Límite Central. Mandelbrot propone en 1965 un modelo de variables estacionarias y de varianza finita, pero en el que las correlaciones decrecen tan lentamente que el teorema central no se cumple. El modelo propuesto es un Movimiento Browniano Fraccionario, que pertenece a la familia de los procesos autosimilares, a los que dedicaremos la próxima sección

# 2.5. Procesos Autosimilares y Movimiento Browniano Fraccionario

Antes de introducir el concepto de memoria larga, propiedad que tienen los modelos que han sido propuestos por los matemáticos para explicar el fenóneno de Hurst y que explicamos en la siguiente sección, haremos algunas consideraciones sobre procesos autosimilares.

**Definición.** Sea  $\{X_t\}$  un procesos estacástico de tiempo continuo. El proceso se dice *autosimilar* con parámetro de autosimilitud H si, dada c > 0 se cumple que la distribución de la variable  $X_{ct}$  es igual a la distribución de la variable  $c^H X_t$ . para todo t.

El resultado fundamental que citaremos es el teorema de Lamperti, en el que los procesos límite de sumas parciales normalizadas son autosimilares y que todo proceso autosimilar puede ser pensado como el límite de procesos de sumas parciales normalizadas (ver Lamperti 1962).

#### Teorema.

1. Sea  $\{Y_t\}$  un proceso estocástico tal  $P\{Y_1 \neq 0\} > 0$  y además se cumple que  $\{Y_t\}$  es el límite en distribución de sumas parciales normalizadas de otro proceso:

$$a_n^{-1} S_{nt} = a_n^{-1} \sum_{i=1}^{[nt]} X_i \xrightarrow{\mathcal{D}} Y_t$$

donde:

- [nt] indica parte entera de nt
- $X_1, X_2, \cdots$  es una sucesión estacionaria de variables aleatorias.
- $a_1, a_2, \cdots$  es una sucesión de constantes positivas tal que  $a_n \to +\infty$

Entinces  $\exists H > 0$  tal que:

$$\lim_{n} \frac{a_{nu}}{a_n} = u^H, \forall u > 0$$

y  $\{Y_t\}$  es *autosimilar* de parámetro H y tiene incrementos estacionarios.

2. Todos los procesos autosimilares de incrementos estacionarios se obtienen de la misma manera.

Definimos, ahora sí el proceso propuesto por Mandelbrot [Mandelbrot 1965, 1968].

**Definición.** Un proceso  $\{W_t^H, t \in \mathbb{R}\}$  se dice un *Movimiento Browniano Fraccionario* si cumple las siguientes propiedades:

1.  $P\{W_0^H = 0\} = 1.$ 2.  $W_t^H \sim N(0, \sigma^2), \ 0 \le t \le 1.$ 3.  $E((W_t^H - W_s^H)^2) = \sigma^2 |t - s|^{2H.}$ 

Los incrementos de este proceso, definen el llamado Ruido Gaussiano Fraccionario

$$X_t = W_t^H - W_{t-1}^H$$

De las propiedades de  $\{W_t^H\}$  deducimos que  $\{X_t\}$  es un proceso gaussiano y estacionario. En el caso particular H = 1/2 se obtiene el proceso de Wiener.

Mandelbrot y Van Ness dan la siguiente representación del movimiento browniano fraccionario en términos de una integral estocástica con respecto al proceso de Wiener en  $I\!\!R$ :

$$W_t^H = \frac{1}{C(H)} \int_{-\infty}^{+\infty} \left[ \left( (t-s)^+ \right)^{H-1/2} - \left( (-s)^+ \right)^{H-1/2} \right] dW_s$$

donde

$$C(H) = \left(\int_0^{+\infty} \left((1+s)^{H-1/2} - s^{H-1/2}\right)^2 \, ds + \frac{1}{2H}\right)^{1/2}.$$

para una demostración ver por ejemplo Nualart 2005. Mostramos a continuación, sólo a modo de ejemplo la realización de un movimiento browniano fraccionario con H = 0, 7.



figura 27 Movimiento browniano fraccionario con H = 0, 7.

Mostramos a continuación la forma de los integrandos de Mandelbrot y Van Ness:

$$K(t,s) = \left( (t-s)^+ \right)^{H-1/2} - \left( (-s)^+ \right)^{H-1/2}$$

48



figura 28 Los integrandos K(t,s) para  $t = 1, s \in [0,1]$  y (a) H = 0,5, (b) H = 0,6, (c) H = 0,7, (d) H = 0,85.

Veamos que ocurre con las covarianzas de este proceso:

$$\gamma_{k} = cov(X_{i}, X_{i+k}) = cov(X_{1}, X_{k+1}) = E(X_{1}X_{k+1}) = \frac{1}{2}E\left((\sum_{j=1}^{k+1} X_{j})^{2} + (\sum_{j=2}^{k} X_{j})^{2} - (\sum_{j=1}^{k} X_{j})^{2} - (\sum_{j=2}^{k+1} X_{j})^{2}\right) = \frac{1}{2}E\left((W_{k+1}^{H} - W_{0}^{H})^{2} + (W_{k}^{H} - W_{1}^{H})^{2} - (W_{k}^{H} - W_{0}^{H})^{2} - (W_{k+1}^{H} - W_{1}^{H})^{2}\right) = \frac{\sigma^{2}}{2}((k+1)^{2H} + (k-1)^{2H} - 2k^{2H}).$$

Si k < 0 se tiene  $\gamma_k = \gamma_{-k}$ . Las correlaciones vienen dadas por la expresión

$$\rho_k = \frac{1}{2}((k+1)^{2H} + (k-1)^{2H} - 2k^{2H}).$$

Si k < 0 se tiene  $\rho_k = \rho_{-k}$ . Observemos en primer lugar que

$$\rho_k = \frac{1}{2} k^{2H} g(k^{-1}),$$

donde

$$g(x) = (1+x)^{2H} + (1-x)^{2H} - 2$$

Si desarrollamos esta función cerca del origen (a los efectos de buscar un equivalente para  $\rho_k)$  tenemos:

$$g'(x) = 2H(1+x)^{2H-1} - 2H(1-x)^{2H-1}$$

que se anula en 0, y

$$g''(x) = 2H(2H-1)(1+x)^{2H-2} + 2H(2H-1)(1-x)^{2H-2}$$

Teniendo en cuenta que g''(0) = 4H(2H - 1) y haciendo el desarrollo de Taylor de g, tenemos:

$$\rho_k \sim H(2H-1)k^{2H-2}.$$

Si 1/2 < H < 1, tal como Mandelbrot propuso, las correlaciones decrecen tan lentamente a 0 que se tiene:

$$\sum_{k=-\infty}^{+\infty} \rho_k = +\infty.$$

Este hecho es el que se toma en general como definición de memoria larga.

# 2.6. El concepto de memoria larga

Tal como se muestra en la sección anterior, uno de los caminos emprendidos para explicar el fenómenos de Hurst es el de la memoria larga, definida, como habitualmente se hace en términos de los momentos de segundo orden, en particular de la convergencia de la serie de las autocorrelaciones. Hagamos en primer lugar algunas consideraciones sobre la varianza del proceso de sumas, que echarán un poco de luz sobre ese concepto: sea un proceso  $X_1, X_2, \ldots, X_n$  estacionario, centrado y con varianza finita  $E(X_i^2) = \sigma^2 < +\infty$ . Sean las covarianzas

$$\gamma_k = cov(X_i, X_{i+k}) = cov(X_1, X_{k+1}),$$

y las correlaciones

$$p_k = cov(X_i, X_{i+k}) / \sigma^2 = cov(X_1, X_{k+1}) / \sigma^2$$

La varianza del proceso de sumas parciales  $S_n = X_1 + X_2 + \ldots + X_n$  viene dada por la fórmula n - n n - n

$$V(S_n) = \sum_{i=1}^n \sum_{j=1}^n cov(X_i, X_j) = \sigma^2 \sum_{i=1}^n \sum_{j=1}^n \rho_{|i-j|} = \sigma^2 \left( \sum_{i=1}^n \rho_0 + \sum_{i=1}^n \sum_{j\neq i}^n \rho_{|i-j|} \right) = \sigma^2 \left( n + 2 \sum_{i=1}^{n-1} (n-i)\rho_i \right),$$

de modo que, si la serie de las correlaciones converge:

$$\sum_{k=0}^{+\infty} \rho_k < +\infty,$$

podemos afirmar que

$$\lim_{n \to +\infty} \frac{V(S_n)}{n} = \sigma^2 \left( 1 + 2\sum_{i=1}^{+\infty} \rho_i \right).$$

Se puede ver que en este caso, bajo las hipótesis apropiadas (condiciones de mixing) se obtiene un principio de invarianza clásico (a menos de la varianza en el proceso de Wiener límite) que hace que el estadístico R/S se comporte como en el caso i.i.d. Por lo tanto, más allá de la definición que se adopte, para explicar el fenómeno de Hurst, la no convergencia de la serie de autocorrelaciones es un hecho que estará presente,

50

hecho en el que en un momento intenso de la discusión de este fenómeno, a mediados de la década del 60 Mandelbrot hizo hincapié, no sólo en su forma matemática, sino en su bella metáfora bíblica sobre el *fenómeno de José*.

Otras condiciones acerca de la memoria larga involucran el concepto de Ergodicidad y Mixing.

**Definición.** Sea  $\{\dots, X_{-1}, X_0, X_1, X_2, \dots, X_n \dots\}$  un proceso estocástico estacionario en el espacio de probabilidad  $(\Omega, \mathcal{F}, P)$  y sea T el operador de traslación (shift, en inglés) tal que:

$$T(\cdots, x_{-1}, x_0, x_1, x_2, \ldots) = (\cdots, x_0, x_1, x_2, x_3 \ldots)$$

es decir, tal que  $(T(x))_i = x_{i+1}$ .

El proceso se dice *Ergódico* si la transformación de shift no tiene conjuntos invariantes probabilísticamente no triviales. Es decir que si  $A \in \mathcal{F}$  cumple:

$$A\Delta T^{-n}(A) = \emptyset, \forall n$$

(donde  $\Delta$  es la diferencia simétrica entre los dos conjuntos) entonces, necesariamente se cumple que P(A) = 0 o bien que P(A) = 1.

**Definición.** Una transformación T en un espacio de probabilidad  $(\Omega, \mathcal{F}, P)$  es *Mixing* si para todo par de conjuntos  $A, B \in \mathcal{F}$  se cumple, cuando  $n \to +\infty$ :

$$P(A \cap T^{-n}(B)) \longrightarrow P(A)P(B).$$

Es decir que si iteramos la transformación  $T^{-1}$  y hacemos tender el número de iteraciones a infinito, el transformado de *B* tiende a ser independiente de *A*. Un proceso  $\{X_t\}_{t\in\mathbb{Z}}$  se dice *Mixing* si el operador de shift es *Mixing* en  $(\Omega, \mathcal{F}, P)$ .

Se tiene el siguiente lema.

**Lema.** Si un proceso  $\{X\}$  en  $(\Omega, \mathcal{F}, P)$  es *Mixing*, entonces es *ergódico*.

Dem. Sea  $A \in \mathcal{F}$  invariante por traslaciones, o sea que se cumple: T(A) = A. Por un lado tenemos:

$$P(A \cap T^{-n}(A)) = P(A).$$

Por otro lado, por la propiedad de mixing:

$$P(A \cap T^{-n}(A)) \longrightarrow P(A)^2.$$

Por lo tanto se cumple la ecuación: P(A) = P(A).<sup>2</sup> y en consecuencia se cumple que P(A) = 0 o bien que P(A) = 1, lo que prueba la ergodicidad del proceso.  $\diamond$ 

La no ergodicidad del proceso implica la existencia de una partición del espacio en dos conjuntos invariantes por traslaciones y probabilísticamente no triviales. Sea A uno de estos conjuntos, la invarianza por traslaciones se podría heurísticamente interpretar diciendo que el proceso guarda una memoria infinita de A. Parece entonces haber un vínculo entre la memoria larga y la no ergodicidad. La pregunta es si se puede extender la línea divisoria, es decir, si es razonable hacer depender la definición de memoria corta de una propiedad más fuerte que la ergodicidad, por ejemplo la propiedad de mixing. A ese respecto vale la pena citar el siguiente resultado:

**Teorema.** Sea X un proceso gaussiano estacionario a valores reales, con medida espectral F en  $(-\pi, \pi]$ . Sean  $\gamma_k, k \ge 0$  las covarianzas y  $\rho_k, k \ge 0$  las correlaciones. Entonces:

- El proceso es ergódico sí y sólo sí F no tiene átomos.
- El proceso es mixing sí y sólo sí lím<sub> $k\to+\infty$ </sub>  $\rho_k = 0$ .

(para una demostración ver, por ejemplo, Aaronson 1997) De modo que el concepto de mixing no parece ser apropiado para definir a partir de él la memoria corta (i.e. decir que un proceso tiene memoria corta si es mixing) para procesos gaussianos estacionarios, ya que una convergencia demasiado lenta de las correlaciones echaría por tierra con ese intento.

Cerramos esta sección con la definición de memoria larga que utilizaremos a partir de ahora:

**Definición.** Sea  $\{X_t\}_{t\in I}$  donde I puede ser el conjunto de los naturales  $I\!N$  o el conjunto de los enteros  $\mathbb{Z}$  (podría concebirse un contexto más general, pero a los efectos del problema que dio origen a estas consideraciones esto es suficiente) definido en un espacio de probabilidad ( $\Omega, \mathcal{F}, P$ ). Se dice que el proceso tiene la propiedad de *memoria larga* si se cumple:

$$\sum_{k=-\infty}^{+\infty} \rho_k = +\infty,$$

donde  $\{\rho_k\}$  indican las autocorrelaciones de orden k del proceso.

# **2.7.** Estimación de H

Hay varios métodos para estimar el coeficiente de Hurst, en esta sección veremos sólo tres, que son a mi juicio los más intuitivos. Para una discusión más general del tema se recomienda ver el arículo de Taqqu, Teverovsky y Willinger de 1995.

#### **2.7.1.** Estimación a partir del estadístico R/S

Bajo la hipóesis de que el proceso es un ruido gaussiano fracccionario de varianza  $\sigma^2$ . Podemos calcular el valor esperado del estadístico R/S.

$$\begin{split} \min_{0 \le i \le n} (S_i - \frac{i}{n} S_n) &- \min_{0 \le i \le n} (S_i - \frac{i}{n} S_n) = \\ \sigma \left( \max_{0 \le i \le n} (W_i^H - \frac{i}{n} W_n^H) - \min_{0 \le i \le n} (W_i^H - \frac{i}{n} W_n^H) \right) \stackrel{\mathcal{D}}{=} \\ \sigma n^H \left( \max_{0 \le i \le n} (W_{i/n}^H - \frac{i}{n} W_1^H) - \min_{0 \le i \le n} (W_{i/n}^H - \frac{i}{n} W_1^H) \right) = \end{split}$$

52

## 2.7. ESTIMACIÓN DE H

donde  $\mathcal{D}_{=}$  indica que la igualdad es es distribución. Por otra parte, por la continuidad de las trayectorias del movimiento browniano fracccionario tenemos que:

$$\begin{split} & \max_{0 \le i \le n} (W_{i/n}^H - \frac{i}{n} W_1^H) - \min_{0 \le i \le n} (W_{i/n}^H - \frac{i}{n} W_1^H) \stackrel{cs}{\longrightarrow} \\ & \sup_{0 \le t \le 1} (W_t^H - t W_1^H) - \min_{0 \le t \le n} (W_t^H - t W_1^H), \end{split}$$

donde
$$\xrightarrow{cs}$$
indica que la convergencia es casi segura o con probabilidad 1. De modo que

$$n^{-H}R/S(X_1, X_2, \ldots, X_n)$$

se comporta asintóticamente como

$$\sup_{0 \le t \le 1} (W_t^H - tW_1^H) - \min_{0 \le t \le n} (W_t^H - tW_1^H).$$

Una forma de proceder en la práctica para estimar H consiste en dividir la muestra de partida  $X_1, X_2, \ldots, X_n$  en bloques de m datos. Habiendo definido previamente el proceso de sumas acumuladas  $S_k = \sum_{i=1}^k X_i$ , con  $S_0 = 0$ . En cada bloque calculamos el rango R(t, m):

$$R(t,m) = \max_{0 \le i \le m} \left( S_{t+i} - S_t - \frac{i}{m} (S_{t+m} - S_t) \right) - \min_{0 \le i \le m} \left( S_{t+i} - S_t - \frac{i}{m} (S_{t+m} - S_t) \right).$$
$$\sigma^2(t,m) = \frac{1}{m} \sum_{i=t+1}^{t+m} \left( X_i - \overline{X}_{t,m} \right)^2.$$

donde  $\overline{X}_{t,m} = \frac{1}{m} \sum_{i=t+1}^{t+m} X_i.$ Se calcula entonces el estadístico

$$R/S(t_i + 1, m) = \frac{R(t_i + 1, m)}{\sigma^2(t_i, m)},$$

siendo  $t_i = \frac{in}{m}$ , para  $i = 0, 1, \dots, [n/m] - 1$ . Variando convenientemente *m* tendremos, para cada *m* un conjunto de [n/m] realizaciones del estadístico R/S, a partir de la cual se tienen dos posibilidades:

- Hacer una regresión lineal con todos los datos del logaritmo del estadístico R/S en función del logaritmo de m.
- Calcular, para cada *m* la media:

$$\overline{R/S}(m) = \frac{1}{[n/m]} \sum_{i=0}^{[n/m]-1} R/S(t_i+1,m)$$

y luego hacer una regresión lineal del logaritmo del estadístico  $\overline{R/S}(m)$  en función del logaritmo de m.



figura 29: Método R/S

Primera de las dos alternativas mencionadas arriba: se muestra la regresión de todos los valores del logaritmo del estadístico R/S en función del logaritmo de m en el caso de una muestra de 10.000 variables gaussianas i.i.d. Se tomaron valores de m entre 100 y 10.000 (con paso 100). El valor de la pendiente es 0.5385 y el verdadero valor de H es 1/2.



#### figura 30: Método R/S

Segunda de las dos alternativas mencionadas arriba: se muestra la regresión de los valores del logaritmo del estadístico  $\overline{R/S}(m)$  en función del logaritmo de m en el caso de una muestra de 10.000 variables gaussianas i.i.d. (la misma de la figura anterior). Se tomaron valores de m entre 100 y 10.000 (con paso 100). El valor de la pendiente es 0.5361 y el verdadero valor de H es 1/2.

## 2.7.2. Método de la varianza agregada

Bajo la hipóesis de que el proceso es un ruido browniano fracc<br/>cionario de varianza  $\sigma^2$ se tiene que la varianza d<br/>e $S_n$  cumple:

$$V(S_n) = V(X_1 + X_2 + \ldots + X_n) = V(W_n^H) = \sigma^2 n^{2H}.$$

Más en general se cuenta con un resultado muy útil a este respecto, debido a Beran (1989).

**Teorema.** Si  $\{X\}$  es un proceso estacionario con memoria larga se cumple:

$$\lim_{n \to +\infty} \frac{V(\sum_{i=1}^{n} X_i)}{n^{2H}} = \frac{k}{H(2H-1)}$$

donde k es una constante.

En la práctica se divide la muestra de partida  $X_1, X_2, \ldots, X_n$  en bloques de m datos y en cada uno de ellos se calcula la media del bloque  $\overline{X}_{t,m} = \frac{1}{m} \sum_{i=t+1}^{t+m} X_i$ . Luego se ve la varianza de la muestra de las medias de los bloques con respecto a la media de la muestra  $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ . De acuerdo a los resultados mencionados anteriormente, dicha varianza debería ser del orden de  $m^{2H-2}$ :

$$\sigma^{2}(m) = \frac{1}{[n/m] - 1} \sum_{i=0}^{[n/m] - 1} \left( \overline{X}_{t_{i},m} - \overline{X}_{n} \right)^{2}.$$

siendo  $t_i = \frac{in}{m}$ , para  $i = 0, 1, \cdots, [n/m] - 1$ .

Variando convenientemente m y haciendo una regresión lineal del logaritmo de  $\sigma^2(m)$  en función del logaritmo de m, se debería obtener una recta de pendiente cercana a 2H - 2.



figura 31: Método de la varianza agregada

Se muestra la regresión de los valores del logaritmo del estadístico  $\sigma^2(m)$  en función del logaritmo de m en el caso de una muestra de 10.000 variables gaussianas i.i.d. (la misma de las figuras anteriores). Se tomaron valores de m entre 10 y 600 (con paso 10). El valor de la pendiente es -1.07, lo que da un valor de 0.4647 para el estimador de H, cuyo verdadero valor es 1/2.

## 2.7.3. Método del Periodograma

Dado el proceso  $X_1, X_2, \ldots, X_n$  definimos el periodograma como:

$$I_n(\lambda) = \frac{1}{n} \sum_{j=1}^n X_j e^{ij\lambda}.$$

El periodograma es un estimador de  $2\pi f_x$ , donde  $f_X$  es la densidad espectral del proceso (ver Apéndice A). Se puede probar que si el proceso, es de incrementos estacionarios, autosimilar con parámetro H < 1, tiene segundo momento finito y si lím<sub>k→+∞</sub>  $\rho_k = 0$ , o sea; si las correlaciones convergen a 0 con k, entonces la densidad espectral f de  $\{X\}$  se puede expresar, en un entorno del origen, como:

$$f(\lambda) = c|\lambda|^{1-2H} + O(|\lambda|^{(3-2H)\wedge 2}),$$

para cierta constante c y donde O(x) indica que el límite  $\lim_{x\to 0} \frac{O(x)}{x}$  es finito. De modo que si para valores de  $\lambda$  próximos al origen hacemos una regresión lineal del logaritmo de  $I(\lambda)$  con respecto al logarimo del valor absoluto de  $\lambda$ , deberíamos obtener una recta de pendiente cercana a 1 - 2H.





La gráfica muestra valores de  $I_n(\lambda)$  en función de  $\lambda$ , para valores pequeños de  $\lambda$ , para la misma muestra de los ejemplos anteriores. Si tomamos logaritmos en ambos conjuntos de datos y hacemos una regresión lineal obtenemos una pendiente  $1 - 2\hat{H}$ muy cercana a 0, de modo que el estimador  $\hat{H}$  estaría a una distancia despreciable del verdadero valor del parámetro: H = 1/2.

# 2.8. Análisis de algunas series de datos tomadas de la naturaleza

## 2.8.1. Aportes complexivos de los ríos Uruguay y Negro.

Se trabajó en primer lugar con la serie de aportes complexivos de los ríos Uruguay y Negro entre los años 1909 y 2009 que se presenta en el primer capítulo.



figura 33: Método R/S para los datos hidrológicos Primera de las dos alternativas mencionadas arriba para el método R/S: Se muestra la regresión de todos los valores del logaritmo del estadístico R/S en función del logaritmo de m en el caso de la serie correspondiente a los aportes complexivos de los ríos Uruguay y Negro. Se tomaron valores de m entre 100 y 3.000 (con paso 100). El valor de la pendiente es 0.6892.



figura 34: Método de la varianza agregada datos hidrológicos Se muestra la regresión de los valores del logaritmo del estadístico  $\sigma^2(m)$  en función del logaritmo de m en el caso de la serie correspondiente a los aportes complexivos de los ríos Uruguay y Negro. Se tomaron valores de m entre 10 y 600 (con paso 10). El valor de la pendiente es -0.5983, lo que da un valor de 0.7008 para el estimador de H.

#### 2.8. ANÁLISIS DE ALGUNAS SERIES DE DATOS TOMADAS DE LA NATURALEZA59

## 2.8.2. Manchas solares

Las manchas solares son regiones de la superficie del sol que están a una temperatura inferior a la del resto de la superficie, lo que explica su color oscuro. Están fuertemente asociadas a campos magnéticos. Este rasgo del sol fue observado desde la antigüedad; pero recién en el siglo XVIII fueron estudiadas en forma más sistemática. La serie que se estudia en esta sección consta de los registros (promedios mensuales) de manchas solares desde 1749 hasta 2013.



figura 35: Manchas solares

El inicio de la serie se debe a la reconstrucción hecha por el astrónomo suizo Rudolf Wolf a mediados del siglo XIX. Wolf llegó a una fórmula que permite estimar el número de manchas partiendo, entre otras cosas de la cantidad de grupos de manchas. Tiempo antes de Wolf, Samuel Schwabe descubriría, luego de una paciente observación de años, un ciclo en la serie de manchas de unos 11 años. Dicho ciclo fue eliminado de la serie antes de su análisis estadístico.



figura 36: Anomalías de la serie de manchas solares







figura 38: Método de la varianza agregada para la serie de manchas solares

Se muestra la regresión de los valores del logaritmo del estadístico  $\sigma^2(m)$  en función del logaritmo de m en el caso de la serie correspondiente al registro de manchas solares (promedios mensuales) desde 1749 hasta 2013. Se tomaron valores de m entre 10 y 600 (con paso 10). El valor de la pendiente es -0.4693, lo que da un valor de 0.7653 para el estimador de H.

## 2.8.3. Anillos de crecimiento de los árboles

La dendrocronología es la ciencia que estudia la datación de los anillos de crecimiento de los árboles. Uno de los primeros científicos en ocuparse de estas cuestiones fue el astrónomo norteamericano Andrew Douglass, quien fundara el *Laboratory of tree ring research* de la universidad de Arizona. Existe una estrecha relación entre el grosor de dichos anillos y muchos fenómenos naturales, como los aportes hidrológicos, las precipitaciones, los registros de temperatura e incluso (según mostró Douglass a principios del siglo XX) con las manchas solares. Los estudios de series dendrocronológicas han permitido reconstrucciones de series de variables climáticas muy importantes. Se destacan en esta área los nombres de Harold C. Fritts y Edward Cook. Sólo a modo de ejemplo citamos la reconstrucción de 1100 años del fenómeno de el Niño (ENSO) realizada en 2011 por un equipo de trabajo integrado entre otros por Cook.



figura 39: Secciones de árboles mostrando anillos

En este ejemplo se trabajó con registros de Aravena de un árbol de la especie Tsuga mertensiana, entre los años 1715 y 2001 (ver Aravena (B3A19S) en http://www.ncdc.noaa.gov/paleo/study/16500).



figura 40 Serie de anillos de crecimiento de un ejemplar de *Tsuga mertensiana*.



figura 41 Método R/S para la serie de anillos de un árbol

Primera de las dos alternativas mencionadas arriba para el método R/S: se muestra la regresión de todos los valores del logaritmo del estadístico R/S en función del logaritmo de m en el caso de la serie correspondiente al registro del grosor de los anillos de un árbol de la especie *Tsuga mertensiana* entre los años 1715 y 2001. Se tomaron valores de m entre 10 y 280 (con paso 10). El valor de la pendiente es 0.9581.



#### figura 42

Método de la varianza agregada para la serie de anillos de un árbol Se muestra la regresión de los valores del logaritmo del estadístico  $\sigma^2(m)$  en función del logaritmo de m en el caso de la serie correspondiente al registro del grosor de los anillos de un árbol de la especie *Tsuga mertensiana* entre los años 1715 y 2001. Se tomaron valores de m entre 10 y 280 (con paso 10). El valor de la pendiente es -0.1701, lo que da un valor de 0.9150 para el estimador de H.

# 2.8.4. Velocidad del viento

Se presenta a continuación un gráfico de la serie de registros del anemómetro del puerto de Montevideo, correspondiente a 500 registros tomados cada un minuto con promedios de la velocidad de viento, un día de agosto de 2005. Es pertinente observar que el concepto de memoria larga depende del tipo de fenómeno que se está registrando, de la frecuencia con la que se toman los datos de la muestra, de si se trabaja con promedios o no. En el caso del viento el concepto de memoria larga sólo puede hacer referencia a un período de unas horas, como máximo.



figura 43: velocidad del viento: datos del anemómetro del puerto de Montevideo correspondientes a un día de agosto de 2005.



figura 44: Método R/S para la serie de velocidad del viento

Primera de las dos alternativas mencionadas arriba para el método R/S: Se muestra la regresión de todos los valores del logaritmo del estadístico R/S en función del logaritmo de m en el caso de la serie correspondiente al registro de velocidad del viento. Se tomaron valores de m entre 20 y 400 (con paso 20). El valor de la pendiente es 0.9115.



figura 45: Método de la varianza agregada para la serie de velocidad del viento

Se muestra la regresión de los valores del logaritmo del estadístico  $\sigma^2(m)$  en función del logaritmo de m en el caso de la serie de velocidad del viento. Se tomaron valores de m entre 10 y 400 (con paso 10). El valor de la pendiente es -0,3068, lo que da un valor de 0.8466 para el estimador de H.



Entonces Marco Polo habló: -Tu tablero, sire, es una taracea de dos maderas: ébano y arce. La tesela en al que se fija tu mirada luminosa fue tallada en un estrato del tronco que creció durante un año de sequía: ¿ves cómo se disponen las fibras? Aquí se distingue un nudo apenas insinuado: una yema trató de despuntar un día de primavera precoz, pero la helada de la noche la obligó a desistir -el Gran Kan no había advertido hasta ese momento que el extranjero supiera expresarse con tanta fluidez en su lengua, pero no era esto lo que le pasmaba-. Aquí hay un poro más grande: tal vez fue el nido de una larva; no de carcoma, porque apenas nacido hubiera seguido excavando, sino de una oruga que royó las hojas y fue la causa de que se eligiera el árbol para talarlo... Este borde lo talló el ebanista con su gubia para que se adhiriera al cuadrado vecino que sobresalía... La cantidad de cosas que se podían leer en un trocito de madera liso y vacío abismaba a Kublai; Polo le estaba hablando ya de los bosques de ébano, de las balsas de troncos que descienden los ríos, de los atracaderos, de las mujeres en las ventanas...

Italo Calvino: Las ciudades invisibles

Capítulo 3

# Modelación con procesos Farima

# 3.1. Consideraciones previas

Un procedimiento habitual en estadística al trabajar con series de tiempo consiste en derivar el proceso, o considerar el proceso de incrementos de la serie original, a los efectos de verificar ciertas propiedades que por alguna razón pudieran resultarnos deseables. Consideremos la serie que resulta de la discretización de un proceso de Wiener, que no es estacionario (la varianza de  $W_t$  aumenta con t, de hecho  $V(W_t) = t$ ). Si consideramos el proceso de los incrementos:

$$b_t = W_t - W_{t-1}, t = 1, \cdots, n$$

tendremos la propiedad de estacionariedad débil:

- 1.  $E(b_t) = 0, \forall t.$
- 2.  $V(b_t) = 1, \forall t.$
- 3.  $cov(b_t, b_{t+h}) = 0, \forall h > 0.$



figura 46 Trayectoria de un proceso de Wiener.



figura 47 Ruido blanco: proceso derivado del Wiener.

Convendrá introducir la siguiente notación: si llamamos B al operador de traslación (*"shift"* en inglés), cumple:

$$B(W_t) = W_{t-1}$$

podemos escribir el procesos de incrementos como:

$$b_t = W_t - W_{t-1} = W_t - B(W_t) = (1 - B)(W_t), t = 1, \cdots, n$$

De la misma manera, se puede derivar el proceso más de una vez en busca de un modelo más adecuado para explicar las observaciones. Supongamos que partimos de un proceso:  $\{X_i, t = i, \dots, n.\}$ , los sucesivos derivados serán:

$$X_t^{(1)} = X_t - X_{t-1} = X_t - B(X_t) = (1 - B)(X_t),$$
$$X_t^{(2)} = X_t^{(1)} - X_{t-1}^{(1)} = X_t^{(1)} - B(X_t^{(1)}) = (1 - B)(X_t^{(1)}) = (1 - B)^2(X_t)$$

y en general:

$$X_t^{(k+1)} = X_t^{(k)} - X_{t-1}^{(k)} = (1-B)^k (X_t).$$

De este tipo de consideraciones surgen los procesos ARIMA y los procesos FARIMA.

# 3.2. Procesos Farima

Los tratamientos primarios que se hacen del problema, en el capítulo 1 utilizan naturalmente los procesos autorregresivos. De hecho el modelo CEGH de la plataforma de simulación SimSEE, utilizada por UTE utiliza este tipo de procesos. Una idea interesante, resulta de preguntarse si alguno de los derivados del procesos de los aportes complexivos de los ríos Uruguay y Negro responden a algún modelo de este tipo, digamos a un modelo más general, como por ejemplo el profusamente utilizado ARMA (proceso autorregresivo de media móvil). Recordamos su definición:

#### 3.2. PROCESOS FARIMA

**Definición** Un proceso  $\{X_t, t \in \mathbb{Z}\}$  se dice ARMA(p, q) si existen  $p, q \in \mathbb{N}$  y coeficientes  $\phi_1, \phi_2, \dots, \phi_p, \theta_1, \theta_2, \dots, \theta_q$  tales que para todo t verifica la igualdad:

$$X_t - \phi_1 \cdot X_{t-1} - \phi_2 \cdot X_{t-2} + \dots - \phi_p \cdot X_{t-p} = \varepsilon_t + \theta_1 \cdot \varepsilon_{t-1} + \theta_2 \cdot \varepsilon_{t-2} + \dots + \theta_q \cdot \varepsilon_{t-q}$$

Donde el proceso  $\{\varepsilon_t\}$  es un ruido blanco; es decir que es una sucesión de variables independientes e idénticamente distribuidas, con media 0 y varianza  $\sigma^2$ . La igualdad anterior se puede escribir en la forma:

$$\Phi(B)(X_t) = \Theta(B)(\varepsilon_t)$$

donde  $\Phi$  es el polinomio

$$\Phi(x) = 1 - \phi_1 x - \phi_2 x^2 + \dots - \phi_p x^p$$

y  $\Theta$  es el polinomio

$$\Theta(x) = 1 + \theta_1 x + \theta_2 x^2 + \dots + \theta_q x^q,$$

y donde  $B^k$  es tomado como la iteración k veces del operador de shift. Se tiene el siguiente teorema, cuya demostración omitimos:

**Teorema.** Si el polinomio  $\Phi$  cumple  $\Phi(z) \neq 0$  para todo z en  $\{z \in \mathbb{C}/|z| = 1\}$ , entonces existe y es única la solución estacionaria a la ecuación:

$$\Phi(B)(X_t) = \Theta(B)(\varepsilon_t).$$

Para una demostración ver, por ejemplo: Rocha 2014.

**Observación.** Al considerar el operador  $(1-B)^k$  vale la escritura formal del Binomio de Newton, es decir

$$(1-B)^k = \sum_{i=0}^k \binom{k}{i} (-1)^i B^i.$$

Teniendo en cuenta que  $\begin{pmatrix} k \\ i \end{pmatrix} = 0$  si i > k podemos escribir:

$$(1-B)^k = \sum_{i=0}^{+\infty} \begin{pmatrix} k \\ i \end{pmatrix} (-1)^i B^i.$$

Recordemos que

$$\left( \begin{array}{c} k \\ i \end{array} \right) = \frac{k!}{i!(k-i)!} = \frac{\Gamma(k+1)}{\Gamma(i+1)\Gamma(k-i+1)}$$

donde la función  $\Gamma$  de Legendre se define como:

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$$

es una extensión del factorial a los números complejos y cumple para los enteros positivos:  $\Gamma(i) = (i - 1)!$ .

Esto permite una extensión del operador diferencial de orden d a valores de d fraccionarios, simplemente siguiendo la escritura formal:

$$(1-B)^d = \sum_{i=0}^{+\infty} \frac{\Gamma(d+1)}{\Gamma(i+1)\Gamma(d-i+1)} \, (-1)^i B^i$$

En este sentido se definen los procesos Farima:

**Definición.** Un proceso  $\{X_t\}$  es un FARIMA(p, d, q), con  $d \in (0, 1)$ ,  $p \ge 0$  y  $q \ge 0$  si su transformado  $\{(1 - B)^d(X_t)\}$  es un ARMA(p, q); es decir si existen polinomios

$$\Phi(x) = 1 - \phi_1 x - \phi_2 x^2 + \dots - \phi_p x^p$$

у

$$\Theta(x) = 1 + \theta_1 x + \theta_2 x^2 + \dots + \theta_a x^q,$$

tales que se cumple la igualdad:

$$\Phi(B)(1-B)^d(X_t) = \Theta(B)(\varepsilon_t).$$

El modelo general fue introducido por Hosking en 1981. Granger y Joyeux estudian en su artículo de 1980 el modelo (0,d,0). Para  $0 \leq |d| < 1/2$ , si se cumple que las raíces de  $\Phi$  y de  $\Theta$  están fuera del círculo unidad, entonces el proceso es estacionario e invertible. Suponemos además que  $\Phi$  y de  $\Theta$  no tienen raíces comunes. Granger y Joyeux aplican este tipo de modelos para el análisis de series de tiempo en economía y McLeod y Hipel (1978) lo aplican al análisis de datos hidrológicos. Previamente Whittle había observado que la modelación mediante procesos ARMA puede requerir un gran número de parámetros, La introducción de estos modelos presenta una alternativa más parsimoniosa en dicho sentido.

Para tener una idea del aspecto de este tipo de procesos dentro del contexto al que los queremos aplicar, consderemos el proceso de sumas acumuladas resultante de un ARMA(2,2),

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$$

donde

 $\begin{array}{rcl} \phi_1 &=& 0,1667,\\ \phi_2 &=& 0,3333,\\ \theta_1 &=& -0,1212,\\ \theta_2 &=& -0,1515,\\ \sigma &=& 2. \end{array}$ 

Una realización de 5000 datos de dicho proceso se muestra en la siguiente figura, subfigura 1; las subfiguras 2,3 y 4 muestran sus antitransformados por el operador  $(1-B)^{-d}$  para d = 0,25, d = 0,40 y d = 0,45 respectivamente, o sea que son procesos FARIMA.

Los dibujos están hechos en la misma escala para acentuar las diferencias entre los procesos. Nótese el aumento del rango de los valores extremos a medida que aumenta el coeficiente d.

En la figura 49 se grafican las correlaciones empíricas de las cuatro series en donde se nota que las mismas son mayores a medida que aumenta d.

72


figura 48 Un proceso ARMA y 3 FARIMA resultantes de transformar el primero según el operador  $(1 - B)^d$  con d = 0, 25, d = 0, 40 y d = 0, 45 respectivamente.



#### figura 49

Correlaciones empíricas de los cuatro procesos de la figura anterior: en negro se ven las correlaciones del ARMA, en rojo la del FARIMA con d = 0, 25, en azul FARIMA con d = 0, 4 y en verde la del FARIMA con d = 0, 45.

#### 3.3. Procesamiento de la serie de datos hidrológicos

Una primera posibilidad en el momento de modelar mediante un proceso FARIMA la serie de datos hidrológicos de los ríos Negro y Uruguay es estimar el coeficiente de Hurst H por alguno de los métodos anteriormente mencionados y luego relacionarlo con el d del proceso FARIMA. Hosking prueba en su artículo de 1981 que

$$\rho_k = O(k^{2d-1})$$

(donde  $\{\rho_k\}$  es la función de autocorrelación de orden k del proceso) es decir que  $\frac{\rho_k}{k^{2d-1}}$  está acotado al tender k a infinito.

Enunciamos a continuación un teorema sobre procesos FARIMA debido a Hosking que incluye dicho resultado y que será importante en el momento de elegir el modelo para los aportes hidrológicos.

**Teorema.** Sea  $\{X_t\}$  un proceso FARIMA(p, d, q). Entonces:

- 1.  $\{X_t\}$  es estacionario si d < 1/2 y la ecuación  $\Phi(x) = 0$  no tiene raíces dentro del círculo unidad.
- 2.  $\{X_t\}$  es invertible si d > -1/2 y la ecuación  $\Theta(x) = 0$  no tiene raíces dentro del círculo unidad.

Si  $\{X_t\}$  es estacionario e invertible, con densidad espectral  $f_X$  y función de correlación  $\{\rho_k\}$ , entonces:

- **3.** Existe y es finito el límite:  $\lim_{x\to 0} x^{2d} f_X(x)$ .
- **4.** Existe y es finito el límite:  $\lim_{k\to\infty} k^{1-2d}\rho_k$ .

(ver la demostración en el apéndice B)

Por otra parte, vimos en el capítulo 2 que para este tipo de procesos se debía cumplir:

$$\rho_k \sim H(2H-1)k^{2H-2}.$$

Igualando, obtenemos que el d del FARIMA subyacente debería ser d=H-1/2.

Sin embargo, las discrepancias que pueden arrojar los distintos métodos de estimación de H parecen sugerir otro camino, que es el que se sigue en el presente trabajo y consiste en probar con distintas transformaciones del proceso original, basadas en distintos valores de d y elegir la más adecuada al aspecto del problema que se quiere enfatizar.

#### Primer paso: transformación gaussiana de los datos

Se procede en primer lugar a llevar los datos a un contexto más amigable (la razón para llevar todo a un espacio gaussiano y al final del tratamiento estadístico volver de ese espacio de la forma en que lo hacemos es evitar los valores negativos que aparecen al generar un proceso Farima. Se ha comprobado que esta forma, que no es nueva, en particular se usa en el trabajo de Chaer (2006 y 2008), es eficaz a este respecto, además de mejorar las curvas de intensidad de sequías, mostradas en el capítulo 1 y

en las que concentraremos nuestro esfuerzo): el contexto gaussiano. Para ello se utiliza la siguiente transformación canónica, basada en la Función de Distribución Empírica. Dado el proceso a aportes  $\{X_t, 1 \le t \le n\}$ . Se aplica la función de distribución empírica a los datos intermedios (o sea a todos los datos con excepción del mínimo y del máximo). Recordemos que la función de distribución empírica asociada a la muestra, se define en los reales y vale:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \le x\}}.$$

Básicamente, esta función la asigna a cada real x un número entre 0 y 1 que es la proporción de observaciones de la muestra que son menores o iguales que x. Esta función la componemos con la función G definida por:

$$G(x) = \begin{cases} \Phi^{-1}(x) & \text{si } x \in (0,1) \\ \Phi^{-1}(1/m) & \text{si } x = 0 \\ \Phi^{-1}(1-1/m) & \text{si } x = 1 \end{cases}$$

Donde  $\Phi^{-1}$  es la inversa de la función de distribución normal típica. La distinción entre los puntos interiores y los extremos se hace sólo a los efecto de evitar problemas en el infinto. Luego de probar con varios valores de m, elegimos m = 2n. Obtenemos entonces el proceso

$$Z_t = G(F_n(X_t)), \ 1 \le t \le n.$$

Luego de esta transformación, los datos tienen un comportamiento aproximadamente normal.



figura 50 Gráfico de barras de los datos luego de la transformación gaussiana.

#### Segundo paso: Desestacionalización de los datos

En nuestro caso estamos trabajando con una serie de datos semanales de 101 años de extensión, correspondientes por lo tanto a n = 5252 semanas. Comenzamos por transformar los datos, restándoles la media:

$$Z'_{t} = Z_{t} - \frac{1}{n} \sum_{t=1}^{n} Z_{t}.$$

A continuación quitamos la tendencia semanal, es decir, se resta a cada dato el promedio a lo largo de esos 101 años de la semana correspondiente del año y se divide por el desvío correspondiente a esa semana:

$$Z_t'' = \frac{Z_t' - \frac{1}{101} \sum_{h=0}^{100} Z_{r(t)+h}'}{s_{r(t)+h}}.$$

donde r(t) es el resto de la división entera de t entre 52 y

$$s_{r(t)+h}^2 = \frac{1}{100} \sum_{h=0}^{100} \left( Z_{r(t)+h}' - \frac{1}{101} \sum_{j=0}^{100} Z_{r(j)+h}' \right)^2$$

es un estimador de la varianza según la semana del año.

#### Tercer paso: Diferenciación fraccionaria de los datos

En este paso partimos de una grilla de puntos  $\{d_1, d_2, \dots, d_K\}$  entre 0 y 1/2 y para cada valor de d aplicamos la transformación que lleva un proceso FARIMA a un ARMA:

$$Y_t = (1 - B)^d (Z_t), \ 1 \le t \le n.$$

Es a este proceso al que intentaremos ajustar un ARMA.

#### Cuarto paso: estimación y ajuste del ARMA

Para cada valor de d se estiman, para una grilla de posibles valores de p y q (en nuestro caso, variando p y q entre 1 y 10, o sea, variando el vector (p,q) en una grilla de  $10 \times 10$ ). Para cada par de valores (p,q) se estiman los polinomios:

$$\Phi_p(x) = 1 - \phi_1 x - \phi_2 x^2 + \dots - \phi_p x^p$$

у

$$\Theta_q(x) = 1 + \theta_1 x + \theta_2 x^2 + \dots + \theta_q x^q,$$

correspondientes al ARMA(p,q) (con  $p \ge q$  fijos de antemano) al que mejor se ajustan los datos. Sean  $\hat{\Phi}_p \ge \hat{\Theta}_q$  dichos estimadores. A continuación se genera un ARMA con esos polinomios  $U_1, U_2, \dots, U_n$  y se estiman las correlaciones mediante las correlaciones empíricas:

$$\hat{\rho}_k(p,q) = \hat{\gamma}_k(p,q) / \hat{\gamma}_0(p,q), \ k = 0, 1, \cdots, h,$$

donde

$$\hat{\gamma}_k(p,q) = \frac{1}{n-k} \sum_{i=1}^{n-k} U_i U_{i+k} - \overline{U}_n^2,$$

siendo

$$\overline{U}_n = \frac{1}{n} \sum_{i=1}^n U_i.$$

A continuación se comparan, para cada par (p,q) estas correlaciones empíricas con las correlaciones empíricas de la serie  $Y_1, Y_2, \dots, Y_n$ , llamémosles  $\{\hat{\rho}_k, k = 0, 1, \dots, h\}$ (en nuestro caso tomamos h = 15), mediante la distancia:

$$\Delta(p,q) = \sum_{k=1}^{h} \left(\hat{\rho}_k(p,q) - \hat{\rho}_k\right)^2,$$

y se eligen los valores de p y q que minimizan esta distancia:

$$(\hat{p} \ \hat{q}) = \arg\min_{p,q} \left\{ \Delta(p,q) \right\}.$$

Tenemos entonces  $\hat{p}$ ,  $\hat{q}$  y los estimadores de los polinomios  $\Phi_{\hat{p}}$  y  $\Theta_{\hat{q}}$ .

#### Quinto paso: Estimación de errores

Partiendo de los polinomios  $\Phi_{\hat{p}}$  y  $\Theta_{\hat{q}}$  consideramos el proceso ARMA:

$$\Phi_{\hat{p}}(B)(1-B)^d(Z_t) = \Theta_{\hat{q}}(B)(\hat{\varepsilon}_t).$$

Si el operador  $\Theta_{\hat{q}}$  es invertible, cosa que en nuestro caso se cumple, podemos escribir esta igualdad<br/>de la forma:

$$\hat{\varepsilon}_t = \Theta_{\hat{q}}^{-1}(B)\Phi_{\hat{p}}(B)(1-B)^d(Z_t).$$

Aplicando esta transformación, se obtiene a partir de la serie de datos  $\{Y_t = (1-B)^d(Z_t)\}$  una muestra de errores, que son el ruido del proceso ARMA $(\hat{p}, \hat{q})$  ajustado, de la que las siguientes figuras, una correspondiente al gráfico de barras y la otra a la función de distribución empírica (que se compara con la de la normal) dan una idea.



figura 51 Gráfico de barras de la serie que constituye el ruido del proceso ARMA $(\hat{p}, \hat{q})$ ajustado, para la serie { $Y_t = (1 - B)^d(Z_t)$ } con d = 0,2499.



#### figura 52

Comparación de la función de distribución empírica de la serie que constituye el ruido del proceso ARMA $(\hat{p}, \hat{q})$  ajustado, para la serie  $\{Y_t = (1 - B)^d(Z_t)\}$  con d = 0,2499 con la función de distribución normal, con la misma media y varianza que la serie de errores.

La densidad de esa muestra, que no es gaussiana según se verá a continuación, se estima mediante la convolución con núcleos gaussianos, obteniéndose la función:

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{nk} \varphi\left(\frac{x - \hat{\varepsilon}_i}{k}\right)$$

donde  $\varphi$  es la función de densidad normal típica y  $k = \frac{1}{\log(\log(n))}$ . Esta densidad nos servirá para generar el ruido de nuestros procesos ARMA.

Si bien la figura anterior muestra la similitud entre la función de distribución empírica de la muestra de errores y la función de distribución normal con la misma media y varianza de la muestra de errores, las pruebas de normalidad aplicadas a dicha muestra rechazan la hipótesis de normalidad.

#### Prueba de normalidad de D'Agostino.

En primer lugar se aplicó la prueba de D'Agostino a la muestra de errores  $\{\hat{\varepsilon}_t\}$ , (ver Dgostino 1970) basada en el estadístico:

$$D = \frac{1}{N^2 \sigma_N} \sum_{i=1}^N \left( i - \frac{N+1}{2} \right) \hat{\varepsilon}_i^*$$

donde

$$\sigma_N^2 = \frac{1}{N} \sum_{i=1}^N \left( \hat{\varepsilon}_i - \frac{1}{N} \sum_{j=1}^N \hat{\varepsilon}_j \right)^2$$
$$\hat{\varepsilon}_1^* \le \hat{\varepsilon}_2^* \le \dots \le \hat{\varepsilon}_N^*$$

у

son los estadísticos de orden de la muestra de errores (es decir: las mismas observaciones, pero ordenadas de menor a mayor).

En nuestro caso se obtuvo D = 0,2725. Luego de hacer una simulación con muestras normales del tamaño de la muestra de errores N = 4998, se ve que para el nivel  $\alpha = 0,05$  la región de aceptación de la hipótesis de normalidad es el intervalo [0,2811; 0,2821] por lo que rechazamos dicha hipótesis.

#### Prueba de normalidad de Lilliefors.

Se calcula el estadístico de Lilliefors (ver Lilliefors, 1973)

$$KSL = \sup_{x \in \mathbb{R}} \left| F_N(x) - \Phi(\frac{x - \overline{\varepsilon}}{\sigma_N}) \right|$$

donde

$$\overline{\varepsilon} = \frac{1}{N} \sum_{j=1}^{N} \hat{\varepsilon}_j$$

y  $F_N$  es la función de distribución empírica de la muestra de errores.

Se obtuvo para el estadístico un valor de KSL = 0,0344, que al ser mayor que el valor crítico para  $\alpha = 0,05$ :  $c_{\alpha} = 0,0125$  (la función de distribución empírica de la muestra de errores difiere demasiado de la normal con su misma media y varianza) nos hace, al igual que la prueba anterior, rechazar la hipótesis de normalidad.

#### Sexto paso: Generación del proceso ARMA

A continuación se generan series de datos según la ecuación

$$\Phi_{\hat{p}}(B)(\hat{Y}_t) = \Theta_{\hat{q}}(B)(\varepsilon_t).$$

donde las variables  $\{\varepsilon_t\}$  son independientes con la densidad  $\hat{f}$  estimada en la parte anterior. El proceso es generado definiendo cada variable  $\hat{Y}_t$  a partir de los datos anteriores:

$$\hat{Y}_t = \phi_1 \cdot \hat{Y}_{t-1} + \phi_2 \cdot \hat{Y}_{t-2} + \dots + \phi_p \cdot \hat{Y}_{t-p} + \varepsilon_t + \theta_1 \cdot \varepsilon_{t-1} + \theta_2 \cdot \varepsilon_{t-2} + \dots + \theta_q \cdot \varepsilon_{t-q}$$

Para iniciar el algoritmo se toman las primeras variables  $\hat{Y}_{t-1} = \hat{Y}_{t-2} = \cdots = \hat{Y}_{t-p} = 0$  y se deja transcurrir el número necesario de iteraciones hasta que el proceso de estabilice. En nuestro caso se hacen 1000 replicaciones antes de comenzar a registrar los valores, a partir de entonces se toma una serie de 101 años (5252 semanas). Ese proceso se repite 1000 veces, es decir: se generan 1000 series de ese tamaño.

#### Séptimo paso: Obtención de las series sintéticas

En este paso se deshacen las transformaciones aplicadas al proceso en los primeros dos pasos. En primer lugar, dada una serie de las generadas en el paso anterior, por ejemplo la *j*-ésima:  $\{Y_t^{(j)}\}$  se le aplica la integración fraccionaria a los efectos de obtener:

$$U_t^{(j)} = (1 - B)^{-d} (Y_t^{(j)})$$

donde

$$(1-B)^{-d} = \sum_{i=0}^{+\infty} \frac{\Gamma(-d+1)}{\Gamma(i+1)\Gamma(-d-i+1)} (-1)^i B^i$$

A continuación se multiplica por los desvíos históricos semanales y se agregan las medias semanales:

$$W_t^{(j)} = s_{r(t)+h} U_t^{(j)} + \frac{1}{101} \sum_{h=0}^{100} Z'_{r(t)+h}.$$

donde r(t) es el resto de la división entera de t entre 52 y  $\{Z'_t\}$  y  $s_{r(t)+h}$  son los definidos en el paso 2. A continuación se agrega la media de las  $\{Z_t\}$  del paso 1:

$$Z_t^{(j)} = W_t^{(j)} + \frac{1}{n} \sum_{t=1}^n Z_t.$$

Finalmente, se deshace la transformación gaussiana hecha en el paso 1. Un criterio empírico nos hace volver del espacio gaussiano, usando la siguiente transformación, en lugar de la inversa exacta de la aplicada. Se considera la muestra de estadísticos de  $\{X_t\}$ :

$$X_1^* \le X_2^* \le \cdots X_n^*$$

Se considera la muestra de rangos de la serie de datos  $\{Z_t^{(j)}\}$ :

$$R_1^{(j)} \le R_2^{(j)} \le \cdots R_n^{(j)}$$

donde  $R_t^{(j)}$  es el lugar que ocupa  $Z_t^{(j)}$  si ordenamos la muestra  $\{Z_t^{(j)}\}$  de menor a mayor. Definimos ahora sí la serie sintética  $\{X_t^{(j)}\}$  como:

$$X_t^{(j)} = X_{R_t^{(j)}}^*$$

#### Resultados

Luego de una intensa simulación en la que no sólo se tuvieron en cuenta las correlaciones de los distintos procesos obtenidos de la diferenciación fraccionario de los datos de partida, sino la simulitud de las curvas de intensidad en el caso de las sequías, se obtuvo para d un valor de 0,2499 (es de destacar que el valor obtenido por Hurst para series de datos hidrológicos es H = 0,75, que corresponde a d = 0,25), y se ajustó al transformado un proceso ARMA(5,5) con coeficientes:

$\phi_1$	=	0,9310
$\phi_2$	=	-0,2291
$\phi_3$	=	0,0093
$\phi_4$	=	$0,\!1150$
$\phi_5$	=	-0,0178
$\theta_1$	=	-0,1161
$egin{array}{c}  heta_1 \  heta_2 \end{array}$	=	$-0,1161 \\ -0,0578$
$egin{array}{c}  heta_1 \  heta_2 \  heta_3 \end{array}$	= = =	$-0,1161 \\ -0,0578 \\ 0,0160$
$egin{array}{c}  heta_1 \  heta_2 \  heta_3 \  heta_4 \end{array}$	    	-0,1161 -0,0578 0,0160 -0,0482
$egin{array}{c}  heta_1 \  heta_2 \  heta_3 \  heta_4 \  heta_5 \end{array}$		$\begin{array}{c} -0,1161 \\ -0,0578 \\ 0,0160 \\ -0,0482 \\ -0,0479 \end{array}$

El ruido del proceso  $\{\varepsilon_t\}$ es generado con la densidad estimada según lo descrito en el paso 5.

A continuación se muestran los promedios según la semana, las correlaciones para 96 semanas y las curvas de intensidad de sequías de las series sintéticas generadas.



figura 53 Promedios según la semana de las series sintéticas generadas (azul) comparados con los de la serie histórica de aportes (negro).



figura 54

Correlaciones hasta 96 semanas de las series sintéticas desestacionalizadas (azul) comparados con los de la serie histórica de aportes desestacionalizada (negro).



#### figura 55

Curvas de intensidad de sequías. Las bandas de confianza (0,25-0,75 y 0,10-0,90) para 1000 series sintéticas generadas según el algoritmo precedente muestran respectivamente los cocientes del mínimo, y los percentiles 5 %, 10 % y 25 % con la mediana, para distintos períodos de tiempo.

84

### Capítulo 4

### Conclusiones

Como conclusiones de este trabajo podemos mencionar:

- 1. El comportamiento estadístico de las series de aportes hidrológicos es complejo; en particular los eventos extremos son difíciles de modelar. Sería deseable disponer de series históricas más extendidas en el tiempo que las de los ríos Uruguay y Negro, a los efectos de entender mejor el comportamiento de los extremos.
- 2. En el corto plazo, modelos sencillos, como los autorregresivos, modelos markovianos e híbridos de unos y otros, pueden ser buenos predictores; pero fallan en el momento de modelar fenómenos extremos en períodos largos.
- 3. Dentro de la predicción a corto plazo, series de datos adicionales, como los del fenóneno El Niño Oscilación Sur (índice N 3.4) mejoran la performance (de hecho, en los modelos autoregresivos para el corto plazo el mejor resultado obtenido considera las 40 semanas previas de El Niño).
- 4. En el largo plazo, los movimientos brownianos fraccionarios, propuestos por Mandelbrot pueden ser de utilidad. Pero dentro de ese contexto parace más adecuado considerar procesos Farima (propuestos por Hosking, Granger y Joyeux) en los que los parámetros del proceso ARMA asociado explican mejor el comportamiento a corto plazo, mientras que el parámetro fraccional d ayuda a explicar las propiedades de memoria larga. De una forma bastante parsimoniosa estos modelos han mejorado la performance de la curva extrema de intensidad de sequías (asociada al mínimo).
- 5. Dentro de este contexto, la propiedad de memoria larga, descubierta por Hurst, parece ser una característica esencial de este tipo de procesos y de otras series de datos proporcionadas por la naturaleza.
- 6. La estimación de parámetros en este tipo de modelos es una cuestión delicada. Por un lado, los diferentes métodos (de los cuales aquí sólo se presentan tres), pueden mostrar sensibles diferencias entre ellos; de la misma forma que la estimación de los parámetros del ARMA asociado. Sin embargo, conjuntos de estimadores muy diferentes pueden dar lugar a modelos igualmente buenos.
- 7. Teniendo en cuenta lo anterior, la busca de nuevos métodos de estimación puede traer mejoras a este respecto.
- 8. En la discusión del trabajo futuro y en el mismo sentido que el comentario anterior, los modelos SARFIMA (modelos Farima estacionales) que se podrían aplicar a los datos sin la desestacionalización podrían resultar interesantes.

9. De la misma forma, se podrían considerar los procesos continuos del tipo Ornstein-Uhlenbeck, propuestos por Arratia, Cabaña y Cabaña (comunicación personal). Dichos procesos tienen la forma

$$X_{\kappa,\sigma}(t) = \sigma \prod_{j=1}^{p} \int_{-\infty}^{t} e^{-\kappa_j(t-s)} dW(s),$$

donde W es un proceso de Wiener. En este caso, el proceso  $\{X_{\kappa,\sigma}\}$  es una versión más parsimoniosa del ARMA(p, p - 1) con ruido blanco gaussiano. Si se introdujera en este tipo de procesos la memoria larga, podrían aplicarse a la modelación de aportes hidrológicos.

## Capítulo 5

# Apéndices

88

## Apéndice A: Densidad Espectral

**Definición.** Sea  $\{X_t\}$  es un proceso estacionario y centrado con función de autcovarianza  $\gamma(\cdot)$  tal que  $\sum_{k=-\infty}^{+\infty} |\gamma(k)| < +\infty$ . Se define la *Densidad Espectral del proceso*  $\{X_t\}$  como:

$$f_X(\lambda) = \sum_{k=-\infty}^{+\infty} \gamma(k) e^{-ik\lambda},$$

para  $\lambda \in \mathbb{R}$ . O sea que  $f_X$  es la Transformada de Fourier de la sucesión  $\{\gamma(k)\}$ . Observación. Como la función  $e^{-ik}$  tiene período  $2\pi$ , bastará considerar la densidad espectral en el intervalo  $(-\pi, \pi]$ .

#### Propiedades de la Densidad Espectral.

- 1. La función es par:  $f_X(\lambda) = f_X(-\lambda), \forall \lambda$ .
- 2.  $f_X(\lambda) \ge 0, \forall \lambda \in (-\pi, \pi].$
- 3.  $\gamma(k) = \int_{-\pi}^{\pi} f_X(\lambda) e^{-ik\lambda} d\lambda, \ \forall k.$

La definición puede hacerse más general para el caso en el que la sucesión de autocovarianzas  $\{\gamma(k)\}$  no es absolutamente sumable. En ese caso tenemos:

**Definición.** Sea  $\{X_t\}$  es un proceso estacionario y centrado con función de autcovarianza  $\gamma(\cdot)$ . Decimos que  $f_X$  es la *Densidad Espectral del proceso*  $\{X_t\}$  si se cumple:

- 1.  $f_X(\lambda) \ge 0, \forall \lambda \in (-\pi, \pi].$
- 2.  $\gamma(k) = \int_{-\pi}^{\pi} f_X(\lambda) e^{-ik\lambda} d\lambda, \ \forall k.$

Las densidades espectrales de un proceso son esencialmente únicas. Si tuviéramos dos densidades espectrales  $f \ge g$  para el proceso  $\{X_t\}$ , tendríamos; por la segunda propiedad que

$$\gamma(k) = \int_{-\pi}^{\pi} f(\lambda) \, e^{-ik\lambda} \, d\lambda = \int_{-\pi}^{\pi} g(\lambda) \, e^{-ik\lambda} \, d\lambda, \; \forall k,$$

lo que implica que sus desarrollos de Fourier coinciden y por lo tanto son la misma función.

Íntimamente ligado de la función de densidad espectral  $f_X$  del proceso está el *Periodograma*, que es una suerte de versión muestral de la función  $2\pi f_X(\dot{)}$ .

Para definir el periodograma, trabajaremos en el espacio vectorial  $C^n$ , donde C indica el espacio de los complejos.

Sea

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in C^n$$

Definamos ahora

$$\omega_k = \frac{2\pi k}{n}, \ k = -[(n-1)/2], \cdots, [n/2]$$

donde  $[\cdot]$  indica *parte entera*. Ahora definimos los vectores

$$e_k = \frac{1}{\sqrt{n}} \begin{pmatrix} e^{i\omega_k} \\ e^{2i\omega_k} \\ \vdots \\ e^{ni\omega_k} \end{pmatrix}, \ k = -[(n-1)/2], \cdots, [n/2].$$

Los vectores  $\{e_1, e_2, \dots, e_n\}$  constituyen una base ortonormal de  $C^n$ . Podemos entonces escribir el vector x en términos de esa base:

$$x = \sum_{k=1}^{n} a_k e_k,$$

donde

$$a_k = x \cdot e_k = \frac{1}{\sqrt{n}} \sum_{j=1}^n x_j e^{-ij\omega_k}.$$

La sucesión  $\{a_1, a_2, \dots, a_n\}$  es la llamada Transformada de Fourier Discreta de  $\{x_1, x_2, \dots, x_n\}$ .

**Definición.** Se llama *Periodograma* de  $\{x_1, x_2, \cdots, x_n\}$  a la función:

$$I_n(\lambda) = \frac{1}{n} \left| \sum_{j=1}^n x_j \, e^{-ij\lambda} \right|^2.$$

La siguiente proposición es importante y permite interpretar al periodograma como un análogo empírico de la función  $2\pi f_X(\dot{)}$ .

**Proposición.** Si  $\{x_1, x_2, \cdots, x_n\}$  son número reales y

$$\omega_k = \frac{2\pi k}{n}, \ k = -[(n-1)/2], \cdots, [n/2]$$

Entonces se cumple que:

$$I_n(\omega_k) = \sum_{|j| < n} \hat{\gamma}(j) e^{-ij\omega_k},$$

donde, para cada  $j, \hat{\gamma}(j)$  es la covarianza muestral de orden j de la serie  $\{x_1, x_2, \dots, x_n\}$ .

## Apéndice B: Teorema de Hosking

**Teorema.** Sea  $\{X_t\}$  un proceso FARIMA(p, d, q). Entonces:

- 1.  $\{X_t\}$  es estacionario si d < 1/2 y la ecuación  $\Phi(x) = 0$  no tiene raíces dentro del círculo unidad.
- 2.  $\{X_t\}$  es invertible (es posible despejar el ruido blanco en función del proceso) si d > -1/2 y la ecuación  $\Theta(x) = 0$  no tiene raíces dentro del círculo unidad.

Si  $\{X_t\}$  es estacionario e invertible, con densidad espectral  $f_X$  y función de autocorrelación  $\{\rho_k\}$ , entonces:

- **3.** Existe y es finito el límite:  $\lim_{x\to 0} x^{2d} f_X(x)$ .
- **4.** Existe y es finito el límite:  $\lim_{k\to\infty} k^{1-2d}\rho_k$ .

#### Demostración

1. Partiendo de la ecuación

$$\Phi(B)(1-B)^d(X_t) = \Theta(B)(\varepsilon_t),$$

en la que asumiremos, sin pérdida de generalidad  $V(\varepsilon) = 1$ , escribimos:

$$X_t = (1-B)^{-d} \Phi^{-1}(B) \Theta(B)(\varepsilon_t) = \psi(\varepsilon_t).$$

donde

$$\psi(x) = (1 - x)^{-d} \Phi^{-1}(x) \Theta(x).$$

Tenemos que la serie de potencias de  $(1-x)^{-d}$  converge para todo x tal que |x| < 1 para d < 1/2. La serie de  $\Phi^{-1}(x)$  converge para  $|x| \leq 1$  cuando las raíces de  $\Phi$  están fuera del círculo unidad y la de  $\Theta(x)$  converge para todo x, ya que  $\Theta$  es un polinomio. Por lo tanto, el desarrollo en series de potencia de  $\psi(x)$  converge para |x| < 1 lo que prueba la estacionariedad de  $\{X_t\}$ .

- 2. La demostración es similar, multiplicando a la izquierda por  $\Theta^{-1}$ , se trabaja entonces con  $\pi(x) = \Theta^{-1}(x)(1-x)^d \Phi(x)$ .
- 3. Se tiene:

$$X_t = \psi(B)(\varepsilon_t).$$

donde $\psi$ viene dada por

$$\psi(x) = (1-x)^{-d} \Phi^{-1}(x) \Theta(x)$$

y su desarrollo en series de potencias es absolutamente convergente. Escrito de otra forma

$$X_t = \sum_{j=1}^{\infty} \psi_j . \varepsilon_{t-j}$$

Se tiene entonces que la densidad espectral de  ${\cal X}$  viene dada por:

$$f_X(x) = |\psi(x)|^2 f_{\varepsilon}(x) = |\psi(x)|^2 \frac{1}{2\pi},$$

donde

$$f_{\varepsilon}(x) = \frac{1}{2\pi}.$$

es la densidad espectral del ruido blanco. Por otra parte se tiene:

$$|\psi(x)|^2 = \frac{\Theta(e^{ix})\Theta(e^{-ix})}{\Phi(e^{ix})\Phi(e^{-ix})}(1-e^{ix})(1-e^{-ix}),$$

cuyo equivalente es

$$\frac{(1-\theta_1-\theta_2+\cdots-\theta_q)^2}{(1-\phi_1-\phi_2+\cdots-\phi_p)^2} \left(2\,sen(\frac{x}{2})\right)^{-2d}$$

cuando  $x \to 0,$ lo que prueba la afirmación **3.** 

**4.** Sea el proceso:  $Y_t = \Theta^{-1}(B)\Phi(B)X_t$ . Se tiene que  $(1-B)^d(Y_t) = \varepsilon_t$ . De modo que  $Y_t$  es un proceso Farima(0,d,0). Por otra parte, definamos el proceso:  $U_t = (1-B)^d(X_t)$ . Tenemos entonces que se cumple:

$$\{X_t\}$$
 ~ Farima $(p, d, q) \Rightarrow f_X(\lambda) = \frac{1}{2\pi} \frac{\Theta(e^{-i\lambda})}{\Phi(e^{-i\lambda})} \left|1 - e^{-i\lambda}\right|^{-2d}$ .

$$\{Y_t\}$$
 ~ Farima $(0, d, 0) \Rightarrow f_Y(\lambda) = \frac{1}{2\pi} \left|1 - e^{-i\lambda}\right|^{-2d}$ .

$$\{U_t\} \sim \operatorname{ARMA}(p,q) \Rightarrow f_U(\lambda) = \frac{1}{2\pi} \frac{\Theta(e^{-i\lambda})}{\Phi(e^{-i\lambda})}.$$

Recordemos además que para cada uno de estos procesos podemos obtener las covarianzas a partir de la densidad espectral:

$$\begin{split} \gamma_k^X &= \int_{-\pi}^{\pi} f_X(\lambda) \, e^{ik\lambda} \, d\lambda. \\ \gamma_k^Y &= \int_{-\pi}^{\pi} f_Y(\lambda) \, e^{ik\lambda} \, d\lambda. \\ \gamma_k^U &= \int_{-\pi}^{\pi} f_U(\lambda) \, e^{ik\lambda} \, d\lambda. \end{split}$$

92

Recordemos el desarrollo de Fourier. Dada una función h de cuadrado integrable en  $[-\pi, \pi]$  se tiene:  $h(\lambda) = \sum_{n=-\infty}^{+\infty} h_n e^{in\lambda}$  donde  $h_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} h(\lambda) e^{-in\lambda} d\lambda$ . Para un producto de funciones de cuadrado integrable h y g tenemos:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} h(\lambda)g(\lambda) e^{-ik\lambda} d\lambda = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{j=-\infty}^{+\infty} h_j e^{ij\lambda} \sum_{n=-\infty}^{+\infty} g_n e^{in\lambda} e^{-ik\lambda} d\lambda =$$

$$\frac{1}{2\pi} \sum_{j=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} h_j g_n \int_{-\pi}^{\pi} e^{i(j+n-k)\lambda} d\lambda = \sum_{j=-\pi}^{\pi} h_j g_{k-j},$$

ya que las integrales de las funciones  $e^{i(j+n-k)\lambda}$ se anulan a menos que j+n-k=0.

Tomando, en la fórmula anterior  $h(\lambda) = f_U(\lambda)$  y  $g(\lambda) = f_Y(\lambda)$  tenemos:

$$\frac{1}{4\pi^2} \gamma_k^X = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} f_X(\lambda) e^{-ik\lambda} d\lambda =$$

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} f_U(\lambda) f_Y(\lambda) e^{-ik\lambda} d\lambda = \frac{1}{4\pi^2} \sum_{j=-\infty}^{+\infty} \gamma_j^U \gamma_{k-j}^Y.$$

Tenemos entonces:

$$\gamma_k^X = \sum_{j=-\infty}^{+\infty} \gamma_j^U \gamma_{k-j}^Y = \sum_{j=-q}^q \gamma_j^U \gamma_{k-j}^Y + \sum_{j=-\infty}^{-q-1} \gamma_j^U \gamma_{k-j}^Y + \sum_{j=q+1}^{+\infty} \gamma_j^U \gamma_{k-j}^Y.$$

Por otro lado se tiene (Hosking 1981):

$$\gamma_k^Y \sim C k^{2d-1},$$

siendo C una constante. Por lo tanto se cumple para  $|j| < \max\{p,q\}$ :

$$\lim_{k \to +\infty} \frac{\gamma_{k-j}^Y}{\gamma_k^Y} = 1.$$

Es posible mostrar (Box, Jenkins & Reinsel 2008) que existen constantes  $\alpha_r$  y  $\delta_r$ , (con  $|\delta_r| < 1$ , para todo r) tal que, para j > q se puede escribir:

$$\gamma_j^U = \sum_{r=1}^p \alpha_r \, \delta_r^{j-q}.$$

Con esta fórmula y un poco de cálculo (que aquí omitimos; pero que se puede consultar en Hosking 1981) es posible acotar:

$$\lim_{k \to +\infty} \left( \sum_{j=-\infty}^{-q-1} \gamma_j^U \, \frac{\gamma_{k-j}^Y}{\gamma_k^Y} + \sum_{j=q+1}^{+\infty} \gamma_j^U \, \frac{\gamma_{k-j}^Y}{\gamma_k^Y} \right) \le 2 \sum_{r=1}^p \alpha_r \frac{\delta_r}{1-\delta_r}.$$

Finalmente tenemos:

$$\lim_{k \to +\infty} \sum_{j=-q}^{q} \gamma_j^U \frac{\gamma_{k-j}^Y}{\gamma_k^Y} = \sum_{j=-q}^{q} \gamma_j^U.$$

Hemos probado entonces que el límite

$$\lim_{k \to +\infty} \frac{\gamma^X_k}{\gamma^Y_k}$$

es finito de lo que se desprende la tesis.  $\diamondsuit$ 

### Referencias

- **Aaronson J.:** An Introduction to Infinite Ergodic Theory, Mathematical Surveys and Monographs, AMS (1997).
- Beran J.: A test of location for data with slowly decaying serial correlations, Biometrika, 76, 261-269 (1989).
- Beran J.: Statistics for Long-Memory Processes, Chapman & Hall/CRC. (1994).
- Billingsley P.: Convergence of probability measures, Wiley (1999).
- Box G., Jenkins G. & Reinsel G.: Time Series Analysis Forecasting and Control, Wiley (2008).
- Brockwell P.J. & Davis R.A.: *Time Series Theory and Methods*, Springer Verlag (1987).
- Brockwell P.J. & Davis R.A.: Introduction to Time Series and Forecasting, Springer Verlag (2002).
- **Cabaña E.:** Contigüidad, pruebas de ajuste y Procesos Empíricos Transformados, X Escuela Venezolana de Matemáticas (1997).
- Chaer R. & Zeballos R.: Modelo Simplificado de Central con Embalse con fines didácticos, IEEE Latin America Transactions, VOL. 4, N. 3 (2006).
- **Chaer R.:** Simulación de sistema de energía eléctrica, Tesis de Maestría en Ingeniería Eléctrica (2008).
- D'Agostino R.: Transformation to normality of the null distribution of g1, Biometrika, 57 (3) (1970): 679-681.
- **Donsker M.D.:** An invariant principle for certain probability limit theorems, 6, Amer. Math. Soc. pp. 110 (1951).
- Genta J.L & Pérez Iribarren G. & Mechoso C.R. .: A recent increasing trend in the streamflow of rivers in southeastern South America, Journal of Climate, Vol. 11, Issue 11, pp. 2858-2862 (1998).
- Granger C.W.J & Joyeux R.: An introduction to long-memory time series models and fractional differencing, Journal of Time Series Analysis Volume 1, Issue 1, pages 1529 (1980).
- Hipel K.W. & McLeod A.I.: Time Series Modelling of Water Resources and Environmental Systems, Elsevier (1994).

Hosking, J.R.M.: Fractional differencing, Biometrika 68(1), 165–176, 17 (1981).

- Hurst H.E.: Long-term storage capacity of reservoirs, Transactions of the American Society of Civil Engineers, Volume 116, 770-799 (1951).
- Kalemkerian J.: Performance de distintos modelos Farima ajustados a la serie de aportes hidrológicos en las represas de Uruguay, Actas del X Clatse, Córdoba (2012).
- Lamperti J.: Semi-stable stochastic processes, Trans. Amer. Math. Soc.: 104, 62-78 (1962).
- Li J. et all: Interdecadal modulation of El Niño amplitude during the past millennium, Nature Climate Change, Vol 1. 114-118 (2011).
- Li W.K. & McLeod A.I. : Fractional time series modelling, Biometrika 73 (1): 217-221 (1986).
- Lilliefors, H.W.: On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown, Journal of American Statistical Association, Vol. 62, No.318, pp. 399-402 (1967).
- Maciel F., Terra R. & Díaz A.: Incorporación de información climática en la simulación de aportes a represas en un modelo del sistema eléctrico, XXV Congreso Latinoamericano de Hidráulica, San José, Costa Rica (2012).
- Mandelbrot B.: Une classe de processus stochastiques homothetiques a soi, application a loi climatologique de H.E. Hurst. Comptes Rendus Acad. Sci. Paris 240: 3274-3277 (1965).
- Mandelbrot B. & Wallis J.R.: Noah, Joseph and operational hydrology, Water Resources Research: 4, 909-918 (1968).
- Mandelbrot B. & Hudson R.L.: The Misbehavior of Markets: A Fractal View of Financial Turbulence, Basic Books Pub. (2006).
- Pérez Iribarren G.: Cadenas de Markov gobernando algunos procesos aplicables a los ríos, Publicaciones Matemáticas del Uruguay, serie de notas y publiaciones técnicas (1999).
- Nualart D.: Fractional Brownian Motion: Stochastic Calculus and Applications, XIV Escuela Latinoamericana de Matemática, Balneario Solís, Uruguay (2005).
- Rocha E.: Consistencia de los estimadores de máxima verosimulitud gaussiana y de mínimos cuadrados en procesos autorregresivos de media móvil, Monografía de Licenciatura, Facultad de Ciencias (2014).
- **Samorodnitsky G.:** Long Range Dependence, Foundations and trends in stochastic systems (2006).
- Taqqu M.S., Teverovsky V., Willinger W.: Estimators for Long-Range Dependence: An Empirical Study, Fractals: 3, 785-798 (1995).