UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

PROYECTO FIN DE CARRERA

ANÁLISIS Y CARACTERIZACIÓN DE SERIES TEMPORALES FINANCIERAS

Alfredo Serrano Quejido

Junio 2014

ANÁLISIS Y CARACTERIZACIÓN DE SERIES TEMPORALES FINANCIERAS

AUTOR: Alfredo Serrano Quejido

TUTOR: Álvaro Diéguez Sánchez-Largo

PONENTE: Joaquín González Rodríguez

Área de Tratamiento de Voz y Señales
Dpto. de Tecnología Electrónica y Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio 2014

Resumen

El objetivo del proyecto es el del estudio de las series temporales financieras, abordando temas que van desde la caracterización, predicción, visualización y generación sintética. En primer lugar, se ha realizado un análisis de las características más relevantes de las series temporales financieras. Posteriormente, se ha estudiado el estado del arte de los métodos de predicción actuales como son las medias móviles, así como el desarrollo de un nuevo método consistente en la aplicación primero de un algoritmo de reducción de dimensionalidad mediante PCA con el posterior cálculo de medias móviles, con el objetivo de mejorar los resultados ofrecidos por el anterior.

Se ha implementado, además, una interfaz gráfica de usuario con la que poder visualizar, entre otras funcionalidades, las series financieras de la base de datos disponible.

Finalmente, de las conclusiones extraídas en predicción, se estudiará el efecto que tiene su aplicación en la generación de series sintéticas, mediante el generador que se dispone en el grupo ATVS, con la finalidad de ver si se consigue obtener una mayor verosimilitud de las características de las series sintéticas generadas ahora con respecto a las características que presentan las series financieras reales.

Abstract

The project aims to study the financial time series, addressing issues ranging from the characterization, prediction, visualization and synthetic generation. First, there has been an analysis of the most relevant features of the financial time series. Subsequently, we have studied the current state of the art of prediction methods such as moving averages and the development of a new method which involves the application of a first dimensionality reduction algorithm using PCA with the subsequent calculation of moving averages with the aim to improve the results offered by the former.

In addition, we have implemented a graphic user interface with which to view, among other features, the financial time series of the available database.

Finally, from the conclusions drawn in prediction, the effect of its application in the generation of synthetic series will be studied using the generator provided in ATVS group, in order to see if it is possible to obtain a greater likelihood of the synthetic series features generated now with respect to the characteristics that the real financial series have.

Agradecimientos

En primer lugar, me gustaría agradecer a Joaquín la oportunidad que me dio de poder formar parte de este gran grupo que es el ATVS, y también por su ayuda durante todo este tiempo con mi PFC.

Por supuesto, agradecerle a Álvaro todo su esfuerzo y dedicación que ha tenido conmigo desde el primer día hasta el último, implicándose incluso estando ya fuera. ¡Gracias por todo Álvaro, eres un crack!

Y como no, agradecerle también a Javi Franco toda su ayuda durante esta última etapa, al cual deberíamos dedicarle todos los proyectandos un lugar especial en nuestros agradecimientos por todas las dudas que siempre nos ha ayudado a resolver.

También, quería agradecerles al resto de mis compañeros del ATVS estos buenos momentos vividos durante mi estancia allí, la cual sin duda se me ha hecho mucho más amena gracias a ellos: Fer, Ali, Ester, Marta, Pedro, Ram, Dani, Rubén.

Me gustaría también dedicarle un hueco en esta página a mi compañero y amigo Iván, con el que he compartido tantas y tantas horas dentro y fuera de la EPS desde que entramos hace ya unos años, y al que estoy seguro que echaré de menos cuando esto se acabe.

Pero todo esto no hubiera sido posible sin mis padres, Alfredo y Maite, a los cuales les debo todo lo que soy. Gracias también a ti, Laura, mi hermana mayor, que siempre que has podido me has echado una mano.

Y no me olvido de ti, Arantxa, por todo tu apoyo y todo lo que hemos pasado y pasaremos juntos.

Bueno, aunque parezca mentira, esto ya llega a su fin. Ha sido una etapa dura, pero también bonita, y será con esta última con la que me quede.

Alfredo Serrano Quejido

Madrid, Junio de 2014

Índice general

Índice de figuras

Índice de figuras

1. Introducción

1.1. Motivación del proyecto

Un objetivo de la ingeniería financiera es el análisis de las series temporales financieras. El conocimiento y análisis de las series económicas existentes se está haciendo cada vez más importante. Es cada vez más relevante ser capaces de valorar posibles escenarios económicos en muchos ámbitos. Para poder realizar análisis, estimaciones y predicciones más precisas, se hace necesario el uso de modelos econométricos. En consecuencia, es necesario el desarrollo de una nueva área tecnológica y de investigación conocido como computación inteligente para ingeniería financiera.

El origen de esta disciplina comenzó ya en los años cincuenta del siglo XX. Se investigó acerca del problema de optimización de portfolio[1] o cartera de inversiones. A pesar de que el estudio de los sistemas financieros ha sido sujeto de estudio durante muchos años, el aumento de la potencia computacional ha abierto un abanico de posibilidades de estudio muy importante. En los últimos 20 años, el campo de la ingeniería financiera se ha expandido a prácticamente todas las áreas de las finanzas y su demanda ha crecido considerablemente.

Este nuevo enfoque permite aplicar las prestaciones computacionales de un ordenador, un sector en auge y que tiene grandes perspectivas de crecimiento y desarrollo. De esta manera podemos ofrecer por un lado, un análisis rápido y eficaz de las series temporales y por otro, tomar decisiones guiadas por una función de coste. Las series temporales financieras están claramente influenciadas por el comportamiento humano, añadiendo un componente extra de pseudo-aleatoriedad que hace de este proyecto un reto interesante y con perspectivas de investigación a largo plazo.

1.2. Objetivos y enfoque

Este proyecto se centra en el estudio y desarrollo de una metodología general motivada por el problema del análisis de las series temporales y modelización de características. Para alcanzar este objetivo ha sido necesario profundizar en el análisis de otro de los ámbitos de mayor investigación en la actualidad dentro de las series temporales: análisis de conjuntos de series temporales de alta dimensionalidad.

[1] Es el conjunto de activos financieros en los cuales se invierte. Se considera el siguiente ejemplo de un portfolio de dos activos, donde se invierte el 100% de nuestro capital en este caso entre el activo A y el B:

- Porcentaje de inversión en A y en B respectivamente: X_A , X_B
- Retorno del activo A y B al final del período de inversión respectivamente: R_A, R_B
- Retorno del portfolio al final del período de inversión: R_P
- Capital inicial de inversión: V_i
- Capital después de la inversión: V_f
- $X_A + X_B = 1$
- $R_P = X_A \cdot R_A + X_B \cdot R_B$
- $V_f = V_i \cdot (1 + R_P)$

Donde el objetivo final es conseguir que: $V_i < V_f$

La combinación de estas dos líneas de trabajo, estrechamente relacionadas, puede dar respuesta a diversos problemas prácticos: i) análisis de las propiedades más relevantes de las series financieras de tal manera que posteriormente se sea capaz de aportar información útil y que ayude en la toma de decisiones en el ámbito financiero. ii) mejora del rendimiento a través de nuevos algoritmos de análisis y modelado de series temporales basados en la proyección de las series a un espacio ortogonal en el cual se pueda realizar un análisis particular de cada serie y posteriormente proyectar al espacio original.

El presente proyecto estará centrado tanto en la mejora del rendimiento a través de nuevos algoritmos de análisis y modelado de series temporales basados en la proyección de las series a un espacio ortogonal, como en el análisis de las características que se pueden observar gráficamente las cuales nos pueden aportar más información.

También, se analizará el resultado de usar determinados parámetros en predicción así como la búsqueda de una configuración óptima de ellos la cual aporte unos mejores resultados. Además, se estudiará el resultado de aplicar las configuraciones de parámetros que mejor resultado han obtenido en predicción, en el generador de series sintéticas desarrollado previamente por el grupo ATVS, con el objetivo de analizar si esta configuración mejora o no la verosimilitud de las características de las series sintéticas generadas con respecto a las características que presentan las series reales comparando con la configuración de parámetros usada anteriormente en el generador.

1.3. Introducción al ámbito financiero

El mercado representa el comportamiento general de todos los activos, es decir, una estimación de la economía entera [4]. Puede ser representado por el portfolio del mercado y a veces es medido por índices concretos como el IBEX 35 en España o el S&P 500 en Estados Unidos.

El índice IBEX 35 (Índice Bursátil Español) es el principal índice bursátil de referencia de la bolsa española elaborado por Bolsas y Mercados Españoles (BME). Está formado por las 35 empresas con más liquidez que cotizan en el Sistema de Interconexión Bursátil Electrónico (SIBE) en las cuatro Bolsas Españolas (Madrid, Barcelona, Bilbao y Valencia). No todas las empresas que lo forman tienen el mismo peso. La entrada o salida de valores del índice es decisión de un grupo de expertos denominado Comité Asesor Técnico (CAT).

Por otro lado, el índice Standard & Poor's 500 (Standard & Poor's 500 Index) también conocido como S&P 500 es uno de los índices bursátiles más importantes de Estados Unidos. Al S&P 500 se le considera el índice más representativo de la situación real del mercado. Este índice bursátil se compone de las 500 empresas más grandes de Estados Unidos.

Los mercados financieros son turbulentos. Tienen dependencias a corto y a largo plazo que afectan a su comportamiento en el futuro a corto y a largo plazo. Sin embargo, existen herramientas que permiten trabajar con los datos, analizarlos y, en cierto modo, usarlos en proyectos de inversión, debido a las similitudes de comportamiento entre los distintos mercados.

Los acontecimientos políticos, económicos y sociales, tanto locales como internacionales, están íntimamente relacionados con el valor de determinados activos, como el tipo de cambio entre dos divisas, acciones bursátiles o materias primas [8]. La determinación de cuáles de estos acontecimientos afectan al precio de un determinado activo y el modo en que lo hacen es lo que se

conoce como análisis fundamental. Este análisis fundamental cuenta con cierta dificultad, la cual puede provenir de que un mismo acontecimiento puede afectar de forma diferente según el activo considerado pero también según el contexto socioeconómico en el que se desarrolle el acontecimiento. No obstante, existen ciertas reglas formales, generadas sobre todo por la experiencia, que pueden servir de guía para realizar un análisis fundamental.

Los eventos o acontecimientos sobre los que realizar este análisis pueden dividirse entre planificados y no planificados, como pueden ser crisis gubernamentales, cambios políticos, desastres naturales, decisiones de bancos centrales, publicación de indicadores económicos… Los eventos planificados, como la publicación de indicadores económicos por ejemplo, se pueden seguir más fácilmente, debido a su publicación en el calendario económico con antelación. La planificación en el calendario de estos eventos permite prever sus efectos, ya que suelen actuar en la misma dirección ante resultados similares. Por ejemplo, los cambios en la tasa de desempleo de un país tiene un efecto que, salvo algunas excepciones, será siempre similar: un aumento de la tasa de desempleo tendrá un efecto negativo sobre el valor de una divisa y viceversa. Sin embargo, los eventos no planificados, como posibles desastres naturales, pueden suponer cambios inesperados en el rumbo de la cotización de los activos.

Otro hecho destacable es que el tiempo es una dimensión fundamental. Grandes pérdidas y ganancias son hechas en breves períodos de tiempo. Estos eventos diversos (informes de inflación, anuncios gubernamentales, etc) ayudan a esta concentración. Las características, por tanto, evolucionan con el tiempo, no pueden considerarse invariantes. Igualmente, existen tendencias de comportamiento. Muchas teorías financieras obvian que dichas tendencias generan en su desarrollo lo que comúnmente se llaman burbujas.

Aunque predecir los precios ha sido hasta el momento una tarea infructuosa, sí que se están realizando investigaciones novedosas respecto a la evaluación de riesgos. No son predicciones deterministas, pero sí que es viable dar estimaciones. Se puede decir que el valor de cualquier activo financiero presenta un valor intrínseco, generalmente considerado como su valor real y que generalmente no se corresponde con el precio de mercado. En teoría, el precio de mercado tiende a moverse hacia este valor real. En base a esta asunción, se intenta desvelar si el valor real de un determinado activo es mayor o menor al valor actual del mercado y, por tanto, prever hacia dónde se movería su cotización.

2. Series temporales

Una serie temporal es una secuencia ordenada de valores, donde a cada valor del eje horizontal le corresponde un valor del eje vertical.

Este eje horizontal representa siempre el tiempo, de tal manera, que a cada fecha (momento temporal) le corresponde un determinado valor de una variable (o ninguno).

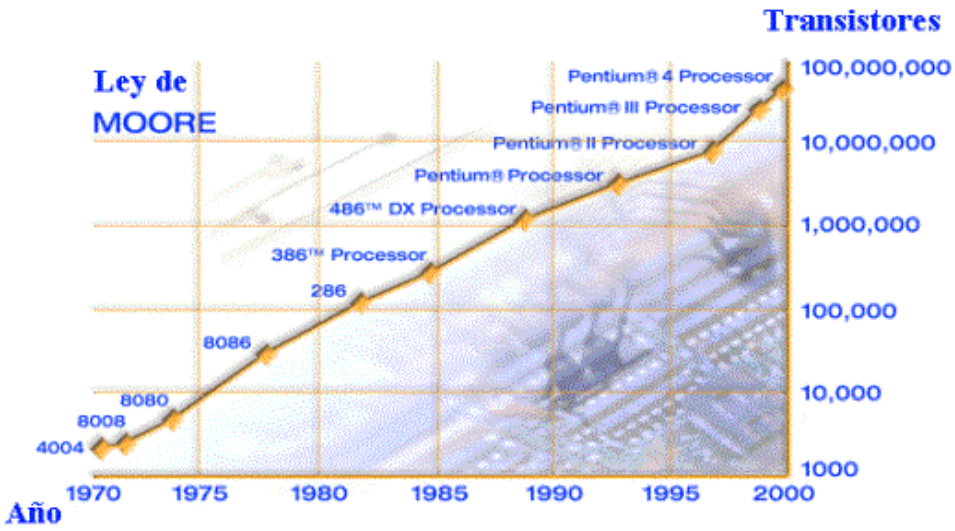

Figura 2.1: Ley de MOORE.

En la gráfica representada, se puede observar la famosa ley de MOORE, la cual afirma que cada 18 meses el número de transistores utilizados en un circuito integrado se duplica.
Como ejemplo de serie temporal también se puede ver una señal de voz, donde la amplitud de la señal varía a lo largo de un eje horizontal que representa el tiempo.

Las acciones de bolsa son un ejemplo de serie temporal, donde para cada t se toman valores, y(t).

Existen series temporales continuas y discretas (con herramientas de estudio y procesamiento diferentes entre sí). En este PFC se trabaja con series temporales discretas.

2.1. Series temporales financieras

Las series financieras se pueden encontrar representadas de diversas formas, pero en el ámbito financiero estas son las dos fundamentales: serie de precios y serie diferencia de precios. Esta última conocida también como la serie de retornos.

2.2. Serie de precios

Una forma de representar estas series es en la forma de serie de precios. Esta serie de precios representa el precio que alcanza una determinada serie para una fecha en concreto. De tal manera que si se disponen de datos para un conjunto de fechas determinado, se podrá ver la evolución de los precios de esa serie en ese periodo de tiempo.

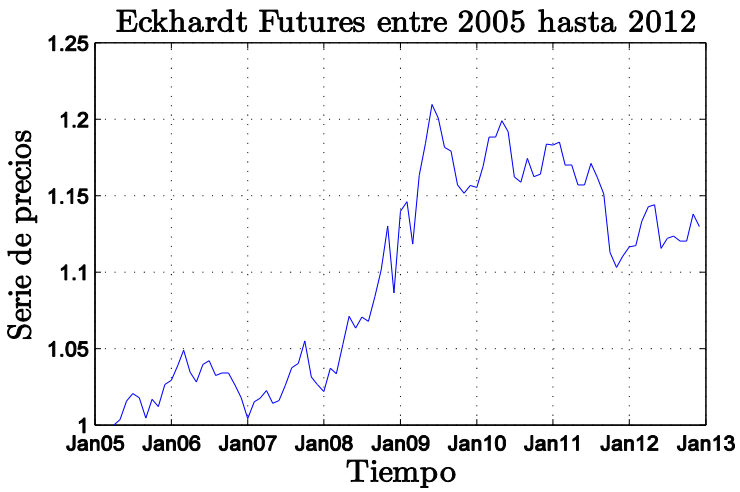

Figura 2.2: Ejemplo de serie de precios.

Las series de precios, como se puede observar en la figura 2.2, presentan tendencia. Se puede ver claramente si tiene tendencia positiva, es decir, varias subidas encadenadas de los precios de la serie, o por el contrario, tendencia negativa, si los precios de la serie comienzan continuamente a encadenar valores cada vez más bajos.

Por tanto, la serie de precios presenta un valor medio y varianza no constantes a lo largo del tiempo, es decir, no es estacionaria en el tiempo.

2.3. Serie diferencia de precios

La otra forma de ver representadas las series financieras es en la forma serie diferencia de precios. También conocida como los retornos de la serie. Esta serie representa el porcentaje de crecimiento de una serie en cada intervalo de tiempo.

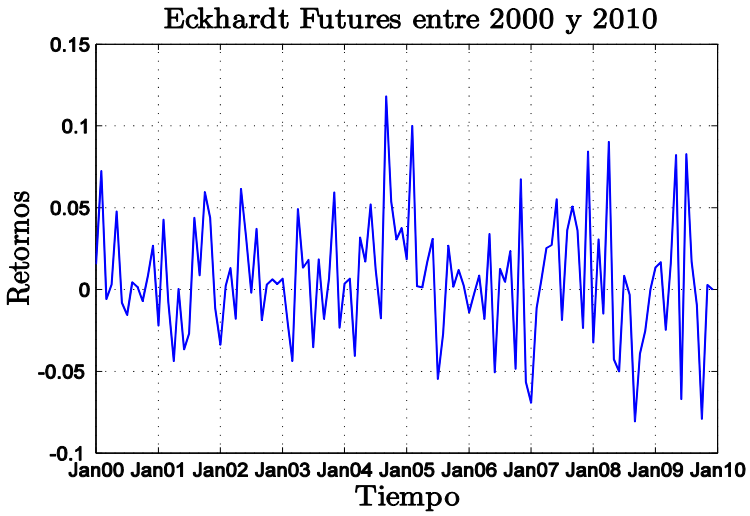

Figura 2.3: Ejemplo de una serie diferencia de precios.

En la ingeniería financiera no se trabaja normalmente con la serie de precios, sino con la serie diferencia de precios. Existen dos formas distintas de calcular estos retornos [1] (serie diferencia de precios), siendo el resultado con cada una de ellas muy parecido (dado que son valores muy cercanos a 0):

$$\boldsymbol{r_t} = \ln(1 + R_t) = \ln(\frac{P_t}{P_{t-1}}) \tag{2.1}$$

$$e^{\boldsymbol{r_t}} = (1 + R_t) = (\frac{P_t}{P_{t-1}}) \Rightarrow \boldsymbol{R_t} = (\frac{P_t - P_{t-1}}{P_{t-1}}) \tag{2.2}$$

$$\boldsymbol{r_t} \approx \boldsymbol{R_t} \tag{2.3}$$

$$\ln\left(\frac{P_t}{P_{t-1}}\right) \approx (\frac{P_t - P_{t-1}}{P_{t-1}}) \tag{2.4}$$

$\mathbf{r_t}$: continuously compounded return (serie diferencia de precios logarítmica)
$\mathbf{R_t}$: simple return (serie diferencia de precios proporcional)

$$1 + R_t[k] = \frac{P_t}{P_{t-k}} = \frac{P_t}{P_{t-1}} \times \frac{P_{t-1}}{P_{t-2}} \times \frac{P_{t-2}}{P_{t-3}} \times \ldots \times \frac{P_{t-k+1}}{P_{t-k}}$$

$$= (1 + R_t) \times (1 + R_{t-1}) \times (1 + R_{t-2}) \times \ldots \times (1 + R_{t-k+1})$$

$$= {}^{k-1}_{j=0}\Pi(1 + R_{t-j}) \tag{2.5}$$

Con $\mathbf{r_t}$ nos referimos a los retornos calculados con la forma logarítmica (ecuación 2.1), y con Pt a los precios de la serie.

Una de las formas de calcular la serie diferencia de precios es tomar el logaritmo neperiano de la división entre dos precios de la serie para distintos días, pudiendo ser estos precios de la serie de días consecutivos o no.

Si los precios son de días consecutivos se tendrán entonces los retornos diarios, sin embargo, si se toman los precios con una separación de 5 días entre ellos, resultarán los retornos semanales. Si esta separación entre precios es de 20 días, se calcularán los retornos mensuales.

Esta forma logarítmica de calcular los retornos presenta alguna ventaja respecto a $\mathbf{R_t}$ (los retornos calculados con la forma proporcional, ecuación 2.2), por ejemplo, permite calcular los retornos semanales como suma de los retornos de los 5 días de la semana (es aditiva):

$$\begin{aligned} r_t[k] &= \ln(1 + R_t[k]) = \ln[(1 + R_t) \times (1 + R_{t-1}) \times \ldots \times (1 + R_{t-k+1})] \\ &= \ln(1 + R_t) + \ln(1 + R_{t-1}) + \cdots + \ln(1 + R_{t-k+1}) \\ &= r_t + r_{t-1} + \cdots + r_{t-k+1} \end{aligned} \qquad (2.6)$$

Retornos semanales:

$$r_t[5] = r_t + r_{t-1} + r_{t-2} + r_{t-3} + r_{t-4} \qquad (2.7)$$

Esta forma de calcular la serie diferencia de precios logarítmica se realiza automáticamente con una función en Matlab incluida en el toolbox de "Econometrics", por lo que su utilización en este trabajo nos ha resultado más cómoda.

Con respecto al cálculo de los retornos semanales usando la forma proporcional, el procedimiento consistiría en aplicar esta frecuencia semanal en la ecuación 2.2:

$$\boldsymbol{R_{t[5]}} = \left(\frac{P_t - P_{t-5}}{P_{t-5}}\right) \qquad (2.8)$$

Siendo el resultado de las ecuaciones 2.7 y 2.8 muy parecido.

A diferencia con la serie de precios, la serie diferencia de precios no presenta tendencia. La diferencia que se puede observar entre ambas es que en ésta la media de los retornos se encuentra siempre en torno a cero, y la varianza, aunque tampoco permanece constante, es algo más similar a lo largo del tiempo, pudiéndola definir como pseudoestacionaria.

2.4. Base de datos

En esta base se dispone de tres tipos de series: series de renta fija, series de renta variable y series de renta mixta:

La renta fija son los bonos que son contratos de deuda, en la cual, alguien cede dinero a alguien a cambio de devoluciones prefijadas (a plazos, sólo al final,...) con una tasa fijada de antemano a la operación. Sin embargo, la renta variable son activos cuyo precio fluctúa con el tiempo. Es decir, se compra una parte de una empresa por ejemplo en forma de acciones y se posee; y si a esa empresa le va bien, los activos se revalorizan, pero si le va mal, los activos pierden valor. Por último está la renta mixta, siendo ésta combinación entre renta variable y fija.

Los activos de renta fija suelen ofrecer un retorno más estable a la hora de invertir, pero normalmente con una rentabilidad total menor a la que puede ofrecer un activo de renta variable, el cual no asegura un retorno fijo pero puede ofrecer retornos más altos.

En la base de datos con las que se realizará este PFC se tienen:

1. Series de renta fija: se disponen datos de 3065 días para 64 series de este tipo.
2. Series de renta mixta: se disponen datos de 3065 días para 16 series de este tipo.
3. Series de renta variable: se disponen datos de 3065 días para 198 series de este tipo.

Tenemos en torno a 13 años de datos para todas estas series, en días que van desde el 30 de julio de 1999 hasta el 17 de enero de 2013.

La base de datos de la que se dispone para realizar este estudio es reducida en tamaño si se compara con otras posibles bases de datos disponibles en otros ámbitos, como por ejemplo en voz. Es necesario por tanto gestionar de la mejor manera posible los escasos datos que se disponen.

Es importante destacar que todas estas series pertenecen a datos reales que se han producido en la realidad.

En la gráfica que se muestra a continuación se representan todas las series, en forma serie de precios, comenzando todas en 1, después de haber aplicado una transformación a las series de precios originales:

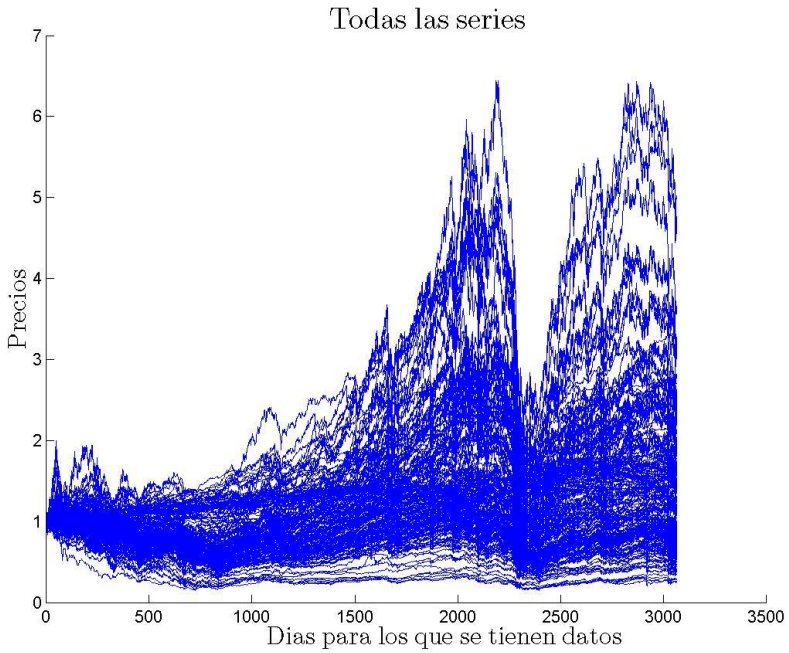

Figura 2.4: Representación de todas las series disponibles en la base de datos en forma de serie de precios normalizada.

Esta transformación se realizó con el objetivo de ver ahora todas las series conjuntamente desde el mismo punto de partida de manera que se aprecien mejor visualmente las subidas y las bajadas de las series, que es lo que nos interesa realmente, y no tanto el visualizar las series con sus precios exactos en el mercado.

La serie diferencia de precios es la misma en ambos casos, sin embargo, este formato permite visualizar mejor todas las series en un rango similar.

Como se puede apreciar en la figura 2.4 al representar todas las series a la vez, es que no se puede distinguir con claridad una serie de otra, ni los distintos movimientos que va realizando la serie a lo largo de los días, ni sus propiedades, características, correlaciones, etc.

Surge entonces la necesidad de poder visualizar las series con una mayor claridad.

2.5. Caracterización

En esta sección se presentan algunas de las características que poseen las series temporales financieras con las que estamos tratando [2]. Este apartado constituye una parte importante del proyecto debido a que aporta una visión más amplia acerca de las características de los datos con los que se trabaja.

En este apartado se detallarán las siguientes características:

- Volatilidad
- Media
- Varianza incondicional

- Skewness
- Kurtosis
- Exponente de Hurst

Este conocimiento de las características que presentan las series financieras puede ser útil a la hora de analizar posibles escenarios que se produzcan en el ámbito real, es decir, en el mercado, lo cual puede arrojar información de lo que está sucediendo, o bien, de lo que puede suceder en un futuro cercano.

A continuación se pasan a detallar las características referenciadas anteriormente.

2.5.1. Volatilidad

Se define como la varianza condicional de los retornos de una serie. La volatilidad es una medida de la frecuencia e intensidad de los cambios del precio de un activo o de un tipo definido como la desviación estándar de dicho cambio en un horizonte temporal específico, es decir, varianza condicional. Se usa con frecuencia para cuantificar el riesgo de la inversión [9].

La volatilidad es uno de los medidores más importantes del riesgo. Sin embargo, suele ocurrir que a más alta volatilidad, más retorno. Otro reto sería ser capaces de prever grandes aumentos de la volatilidad acompañados de un desplome de los retornos ocurridos con frecuencia.

Entre las características de la volatilidad [3] se encuentran las siguientes:

- Aglomeración de la volatilidad: aparece agrupada por períodos tanto de alta como de baja volatilidad, cualidad que hace que las series financieras no se consideren estacionarias sino pseudoestacionarias.

- Reversión a la media: existe un nivel normal de volatilidad al cual esta retorna eventualmente. Por tanto, los pronósticos a largo plazo convergerán todos al nivel normal de la volatilidad, independientemente del tiempo.

- Otra característica importante es que la volatilidad no se comporta igual en períodos de retornos negativos que en períodos de retornos positivos.

- Influencia de variables exógenas, es decir, existe cierta información en otras series que afectan en el comportamiento de ésta.

- Existen comovimientos en las volatilidades de distintas series. Cuando las volatilidades se mueven en una dirección, volatilidades de otros activos también lo hacen. El análisis de múltiples series simultáneamente es de gran interés y presenta un abanico más grande de posibilidades que las series unidimensionales.

Para obtener la siguiente gráfica se ha calculado previamente la volatilidad media de todos los activos de renta variable que se disponen en la base de datos (curva de color negro). Además, en la gráfica también se pueden observar las volatilidades de dos series elegidas por el autor:

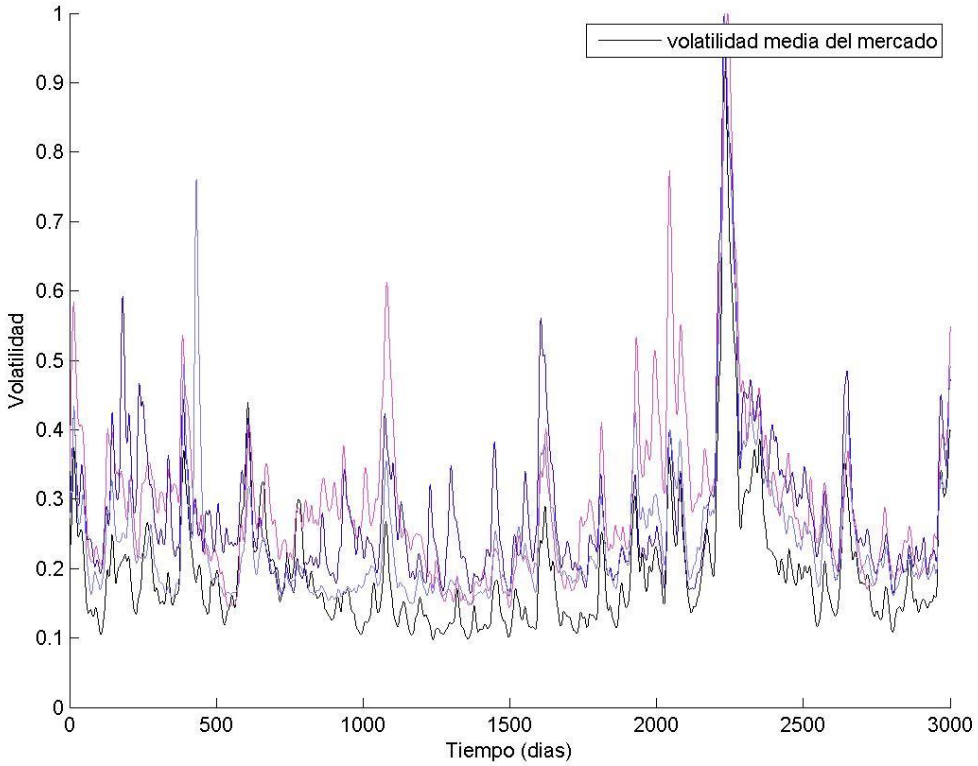

Figura 2.5: Representación de la volatilidad media del mercado y de la volatilidad de dos series elegidas por el autor.

En la figura 2.5 se pueden apreciar algunas de las características mencionadas anteriormente, tales como la aglomeración de las volatilidades en períodos de alta o baja volatilidad, coincidiendo también entre las distintas series estos períodos de tiempo. Esto último corrobora el hecho de que existen comovimientos en las volatilidades de distintas series: cuando las volatilidades de una serie se mueven en una dirección, volatilidades de otras series también lo hacen.

2.5.2. Media

La primera característica que se muestra es la media, o también conocida como momento de primer orden.

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i \qquad (2.9)$$

Si se suman todos los retornos de la serie para las fechas de las que se disponen datos, y posteriormente se divide entre el número de días, se obtendrá un valor en torno a cero. Esto es, la media de todos los retornos de una serie estará en torno a cero. A continuación se muestra el histograma de los retornos de una serie de renta variable:

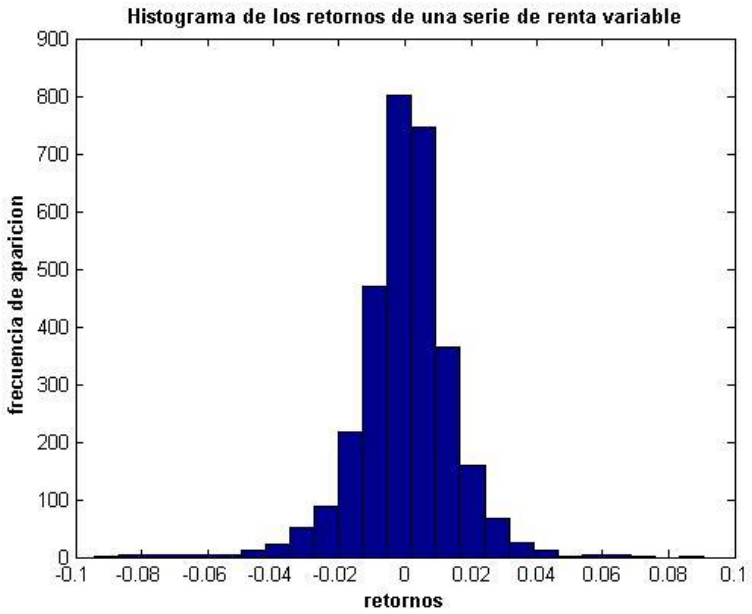

Figura 2.6: Histograma de los retornos de una serie de renta variable.

Si se observa el histograma de la figura 2.6, los valores de retornos que más se repiten están en torno a 0, y valores muy cercanos a éste, por lo tanto, al calcular la media de los retornos de estas series, se tendrá un valor que efectivamente estará próximo a cero. Se puede ver como esta distribución se parece notablemente a la distribución normal. Sin embargo, como se comprobará más adelante, las colas (los valores extremos de la distribución) son mayores que en una distribución normal. Muchas de las teorías de la ingeniería financiera se basan en el supuesto parecido, lo que puede llevar a grandes errores de cálculo.

En general, en los activos se producen más pequeñas subidas que bajadas, pero cuando se producen bajadas, éstas suelen ser más grandes. Aunque es difícil apreciarlo gráficamente en el anterior histograma, más adelante trataremos de dar evidencia analítica de esta característica.

2.5.3. Varianza incondicional

La varianza, o momento de segundo orden de una variable aleatoria, es una medida de dispersión definida como la esperanza del cuadrado de la desviación de dicha variable respecto a su media.

$$Var(X) = E\left[(X - \mu)^2\right] \qquad (2.10)$$

$$= E[X^2] - \mu^2$$

Para ilustrar esta característica se muestra debajo la diferencia de precios diaria de una serie temporal financiera:

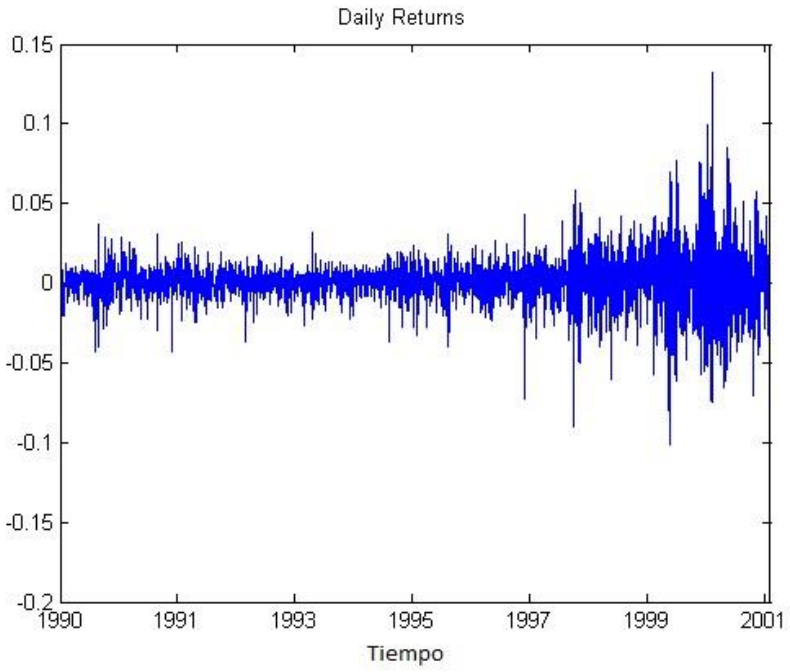

Figura 2.7: Serie diferencia de precios diaria.

Como se puede ver en la figura 2.7, la varianza no siempre se mantiene constante en las series de retornos, puede variar en mayor o menor medida dependiendo del periodo en que nos encontremos. Es decir, la varianza no es constante a lo largo del tiempo. En este ejemplo, este hecho se observa claramente si comparamos los datos del año 2000 con respecto a los datos del año 1994.

A menudo, en la literatura se suele emplear el término heteroscedasticidad [5], el cual se utiliza para referirse a la distinta dispersión respecto de la media observada en las series.

Otro fenómeno no menos usado y el cual también se puede apreciar en la figura 2.7 es el conocido como Volatility clustering [3], el cual afirma que frecuentemente las series de retornos se caracterizan por cambios grandes que van seguidos por cambios grandes, y por cambios pequeños que van seguidos por cambios pequeños.

2.5.4. Skewness

La siguiente característica explicada es el Skewness, o momento de tercer orden. Esta característica mide la simetría de una distribución de probabilidad. El Skewness se calcula a partir de la serie diferencia de precios.

$$Skew(x) = E\left[\left(\frac{X - \mu_x}{\sigma_x}\right)^3\right]$$

$$= \frac{E[X^3] - 3\,\mu E[X^2] + 3\mu^2 E[X] - \mu^3}{\sigma^3}$$

$$= \frac{E[X^3] - 3\,\mu\sigma^2 + 2\mu^3}{\sigma^3} \tag{2.11}$$

Distinguimos entre dos posibles valores de Skewness: valores positivos o valores negativos.

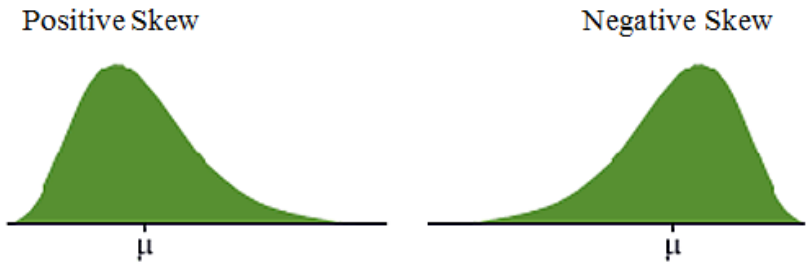

Figura 2.8: Representación de una supuesta distribución con Skewness positivo y otra con Skewness negativo.

Si la mayoría de valores de x están por encima de la media, al calcular el skewness se tendrá un número positivo, es decir, habrá una inclinación en la distribución hacia valores a la derecha de la media. Por tanto, las distribuciones con Skewness positivos presentarán una gran cola derecha.

Sin embargo, si la mayoría de valores de x están por debajo de la media, al calcular el Skewness resultará un número negativo, y por tanto, se tendrá una inclinación en la distribución hacia valores a la izquierda de la media. Resultando así que las distribuciones con Skewness negativos presentarán una gran cola izquierda.

En definitiva, el Skewness nos da una idea del grado en que una distribución de probabilidad se encuentra inclinada hacia un lado u otro de la media, es decir, del grado de simetría de la distribución.

En una distribución simétrica como la normal, los valores por encima y por debajo de la media se cancelan, y cuando se calcula el Skewness se obtiene un valor igual a cero.

A continuación se muestra el histograma del Skewness de las series de renta variable de las que se dispone en la base de datos:

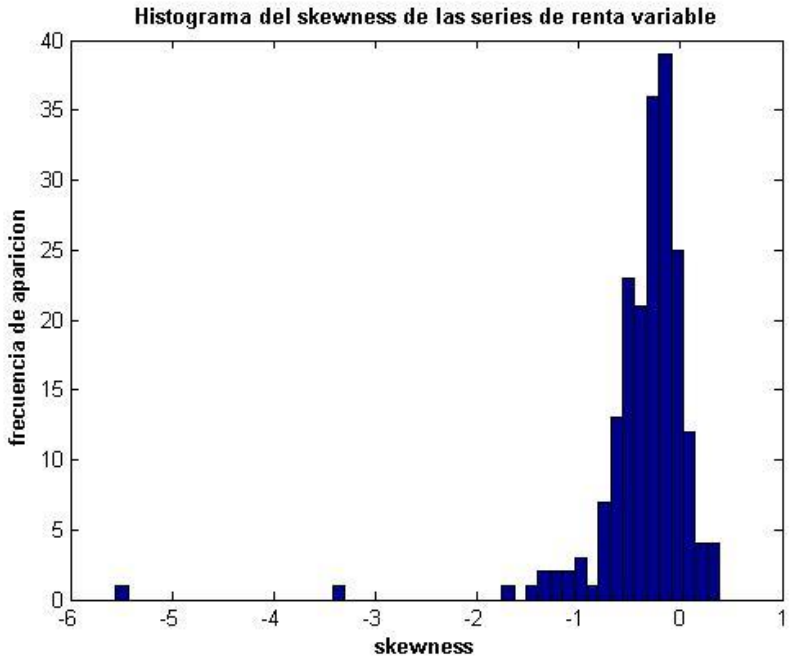

Figura 2.9: Histograma del Skewness de todas las series de renta variable disponibles en la base de datos.

Como se puede observar en el histograma, la gran mayoría de valores está por debajo de cero, con lo que se puede deducir que los retornos negativos "pesan" más que los retornos positivos en este tipo de series.

Por tanto, las series con las que se trabaja en este ámbito suelen presentar un Skewness negativo, cercano a 0. Esto también se puede afirmar si recordamos que el valor medio de los retornos se encuentra en torno a cero, y que los retornos negativos pesan algo más que los positivos. Lo cual coincide con el hecho de que se producen más pequeñas subidas que bajadas, pero las bajadas que se producen son de un valor más elevado o brusco.

2.5.5. Kurtosis

La característica ahora explicada es la Kurtosis, o momento de cuarto orden. Esta característica mide la probabilidad de eventos extremos para una distribución. Trata de estudiar la mayor o menor concentración alrededor de la media y en la zona central de la distribución:

$$Kurt(x) = E\left[\left(\frac{x - \mu_x}{\sigma_x}\right)^4\right] \qquad (2.12)$$

Distinguimos entre tres rangos de valores posibles de Kurtosis:

- Valores **mayores que 3**: si la distribución tiene valores alejados de la media, ya sean por la izquierda o por la derecha, producirán grandes valores de Kurtosis. Las distribuciones con estos valores de Kurtosis presentan unas colas más elevadas o altas que las de la normal.

- Valores **iguales a 3**: este valor es característico de la distribución normal, y esta distribución implica que los eventos extremos no ocurren muy a menudo, son poco probables.

- Valores **menores que 3**: si la distribución no tiene valores alejados de la media, el valor de Kurtosis para estas distribuciones no será muy elevado. Las distribuciones con estos valores de Kurtosis presentan unas colas más delgadas o bajas que las de la normal.

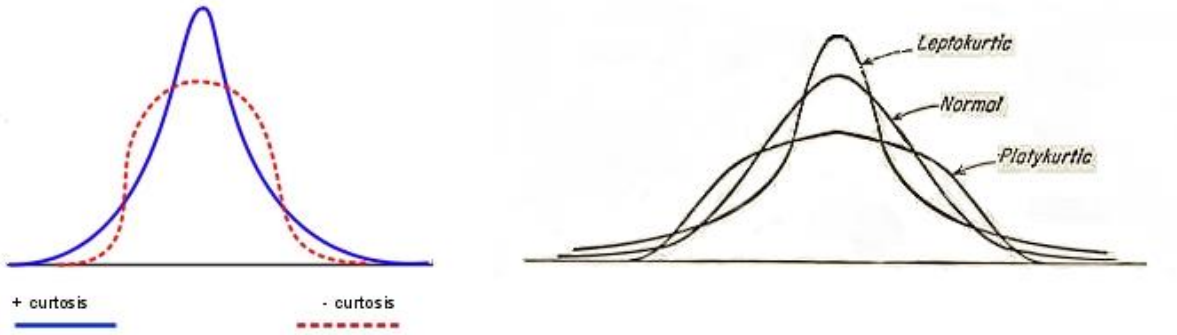

Figura 2.10: Representación de distribuciones con distintos valores de Kurtosis.

A menudo se usa el término "exceso de Kurtosis", que no es más que el valor resultante de restar la Kurtosis de una distribución menos la Kurtosis de una distribución normal, la cual es 3. Si este "exceso de Kurtosis" es positivo, estaremos en el primer caso. Si el "exceso de Kurtosis" es igual a 0, estaremos en el segundo caso (la distribución tiene la misma Kurtosis que la distribución normal), y si el "exceso de Kurtosis" es negativo, estaremos en el tercer caso.

A continuación se muestra el histograma de valores de Kurtosis de las series de renta variable de las que se dispone en la base de datos:

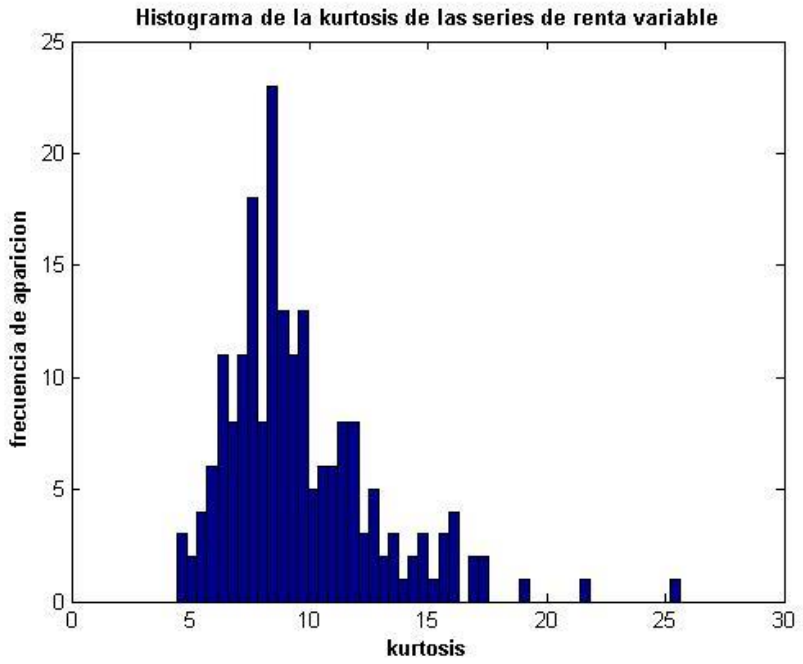

Figura 2.11: Histograma de los valores de Kurtosis de todas las series de renta variable disponibles en la base de datos.

A la vista del histograma de la figura 2.11, se puede concluir que las series de este tipo presentan un valor de Kurtosis mayor que 3, por lo que la probabilidad de eventos extremos, es decir, la probabilidad de que haya retornos alejados del valor medio, es elevada, ocurren con cierta frecuencia.

Los grandes cambios que han ocurrido y que pueden ocurrir quedan reflejados en características como esta. Por ejemplo, si un día se produce una ruptura del mercado, tendremos un retorno que estará en lo lejos de la cola izquierda de la distribución. Y si tenemos un evento extremo positivo, esto es, un día en el que el mercado va hacia arriba, es decir, si se produce una gran subida, tendremos un valor extremo en la cola derecha de la distribución.

2.5.6. Exponente de Hurst

El exponente de Hurst es un medidor de aleatoriedad de las series. Este valor está comprendido entre 0 y 1:

$$0 < H < 1$$

Un valor de Hurst igual a 0.5 (H=0.5) indica que la serie se comporta de manera totalmente aleatoria. Valores extremos como 0 o 1 reflejan que la serie se comporta de manera totalmente determinista, pero significando cosas diferentes según sea uno u otro:

- H > 0.5: correlación positiva entre incrementos. Si un valor de Hurst es mayor que 0.5, implica que si el precio de la serie sube, al siguiente día tendrá más probabilidad de subir, y si baja, es más probable que al día siguiente vuelva a bajar.

- H < 0.5: correlación negativa entre incrementos. Para valores menores que 0.5, es decir, más cercanos a 0, implican que si el precio de la serie ha subido, al día siguiente tendrá más probabilidad de bajar, y al siguiente día de volver a subir.

A continuación se muestra el histograma del exponente de Hurst calculado sobre las series de renta variable:

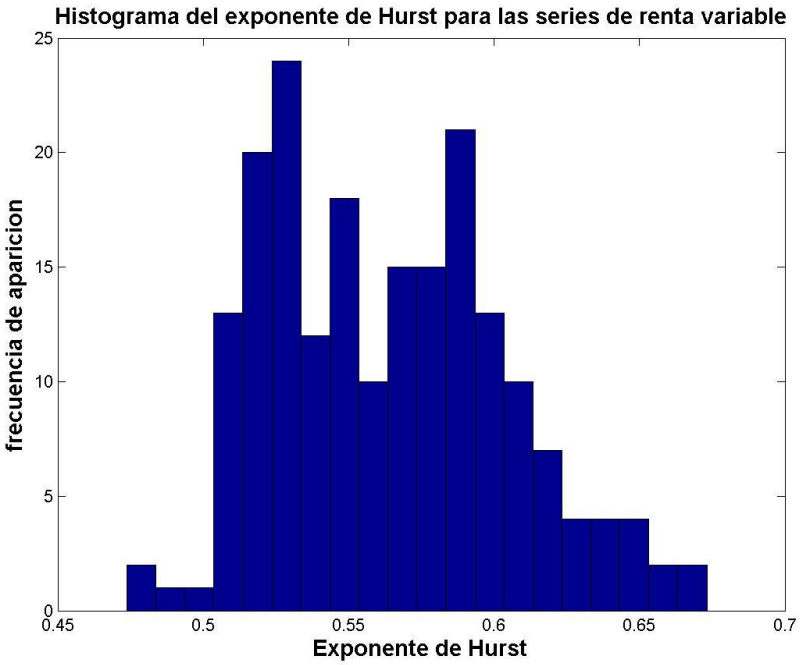

Figura 2.12: Histograma del exponente de Hurst de todas las series de renta variable disponibles en la base de datos.

Como se puede observar en el histograma de la figura 2.12, la mayoría de valores se concentran por encima de H=0.5, con lo que las series de este estudio presentan una correlación positiva entre incrementos, no obstante, son valores que están más próximos a 0.5 que a 1, con lo que se tendrán unas series con un comportamiento cercano a la aleatoriedad.

2.6. Generador de series sintéticas

Para realizar una de las partes de este PFC se ha utilizado un generador de series sintéticas implementado anteriormente por el grupo ATVS. En los siguientes apartados se pasará a contar las causas que previamente motivaron su implementación y una introducción en la cual se describe brevemente su funcionamiento.

2.6.1. Motivación del generador

La base de datos utilizada en este trabajo cuenta con un total de 278 series, con información para 3065 días, dentro de la cual se encuentran series de distintos tipos: renta variable, renta fija, renta mixta y divisas. El generador de series sintéticas desarrollado previamente por el grupo ATVS surge de la necesidad de poder contar con un mayor número de datos, es decir, de días para los que se tenga información de las series temporales financieras, debido a la relativa escasez existente.

Estos nuevos datos generados pueden ser de gran utilidad para probar la bondad de los diversos algoritmos ya implementados, o futuros algoritmos surgidos a partir de una necesaria mejora vista a partir de los resultados obtenidos.

2.6.2. Breve introducción

El generador ha evolucionado a lo largo de sus distintas versiones de manera que ha ido mejorando la generación de series sintéticas en lo que a similitudes con las características de las series reales se refiere.

Para la generación de series sintéticas, el generador parte de las series diferencias de precios reales que se tienen. En este apartado se habla de dos niveles:

- Nivel 1: nivel micro, en él se generan los retornos serie a serie.
- Nivel 2: nivel macro, en él se modela la volatilidad de todo el mercado.

A nivel 2, primero se calcula la volatilidad media de todo el mercado, es decir, de todas las series para todos los días de los que se tiene información.

Una vez calculada, se agrupan períodos de volatilidad similar utilizando un clustering bottom-up. Inicialmente se inicializa el algoritmo con tantos clústeres como semanas tienen las series. De forma iterativa, se van uniendo clústeres contiguos con comportamientos similares. El punto de parada lo establece un número fijo introducido manualmente.

Dependiendo de la longitud en días respecto del total de cada uno de los macroestados generados con el clustering, se asigna a cada macroestado un ancho proporcional a su probabilidad de aparición en un intervalo entre 0 y 1. Posteriormente se genera un número aleatorio, el cual indica el intervalo a seleccionar, correspondiente a cada uno de los macroestados.

A continuación, a nivel 1, se entrena la matriz de datos reales mediante PCA. La generación de series sintéticas se realiza serie a serie a partir de las series ortogonales originadas mediante PCA. El uso de PCA permite en este punto generar las series una a una, siendo estas autoseries generadas incorreladas entre sí, pero siendo capaces de recuperar esa información entre ellas al volver a proyectar las series sintéticas sobre el espacio original de partida.

En este punto, las autoseries generadas se suponen incorreladas entre sí, y se procede a realizar el proceso generativo serie a serie. El período de días que comprende el macroestado seleccionado previamente, es utilizado para seleccionar los datos de las series ortogonales en ese mismo período. Este período es dividido a su vez en segmentos formados por cinco valores, siendo estos valores los retornos de la serie. Después, estos segmentos se ordenan en un banco de segmentos de mayor a menor volatilidad.

Posteriormente, se calcula un autómata de nivel 1 de transición de estados a partir de la volatilidad de la serie ortogonal, con tantos estados diferentes como número de segmentos hayan resultado anteriormente. Finalmente, las series sintéticas se van generando copiando los valores de retornos que contienen el banco de segmentos según vaya determinando el autómata de nivel 1.

Por último, se proyectan las series sintéticas generadas al espacio original, ya que en el espacio proyectado las series carecen de sentido económico, siendo las proyecciones en el espacio original las que se suponen que poseen las características de una serie financiera real.

A continuación se muestra una gráfica donde se pueden observar las series sintéticas generadas con dicho generador a partir de una serie de precios real normalizada:

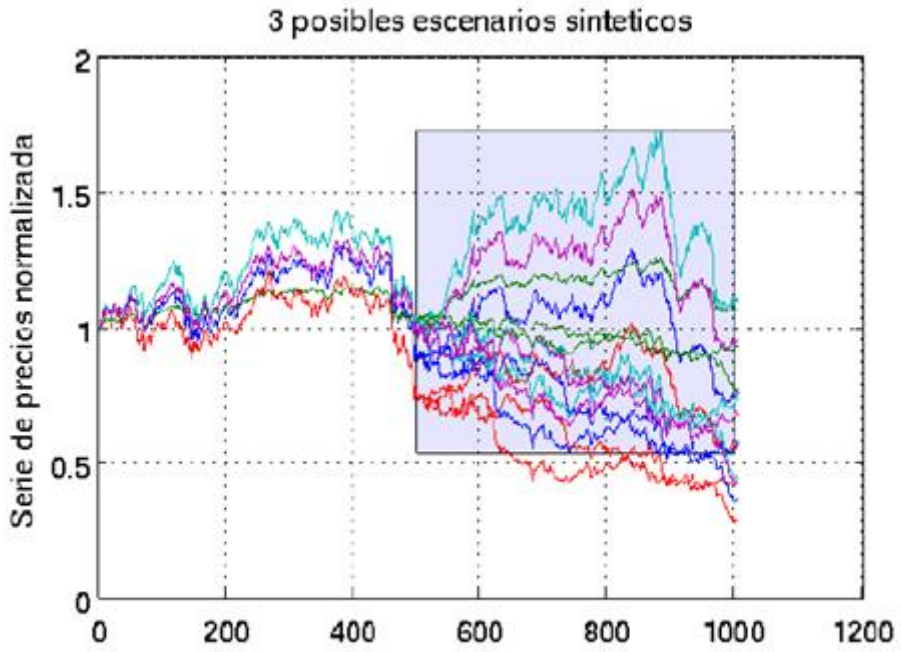

Figura 2.13: Ejemplo de generación de tres escenarios diferentes de series temporales a partir de series temporales reales.

3. Predicción

Cuando se habla de predicción en series temporales financieras se entiende el uso y aplicación de diferentes algoritmos y técnicas basadas en comportamientos pasados de las series con el fin de dar estimaciones de posibles comportamientos futuros de éstas.

Sin embargo estas estimaciones no son infalibles, debido entre otros a la gran cantidad de variables que influyen en el comportamiento real de las series, y algunas de las cuales poseen una componente aleatoria que no es recogida por el modelo, pero pueden aportar una buena aproximación de su posible comportamiento futuro.

3.1. Algoritmos de predicción

Estos métodos de predicción tratan de dar una estimación de posibles valores futuros de las series en base a valores anteriores reales que se han producido en estas mismas series. Para ello se ha usado el algoritmo conocido en la literatura como medias móviles, el cual se pasa a explicar más adelante. Este algoritmo se aplicará a los datos anteriores de las series reales.

En el presente proyecto se han utilizado y comparado diversos algoritmos y métodos de cara a poder predecir de la mejor manera posible valores futuros de las series.

3.1.1. Medias Móviles

Las medias móviles son un promedio aritmético el cual suaviza la serie de precios y determina una línea de la tendencia seguida en los precios.

Se pueden encontrar varios tipos de suavizado [10]:

$$\widehat{s_t} = suavizado(s_t) \qquad (3.1)$$

- Simple: calcula la media aritmética de los precios en un cierto período de tiempo a determinar.
- Ponderado: multiplica la media aritmética por determinados factores, de tal manera que aporta más relevancia a los precios más recientes respecto a los más antiguos.
- Exponencial: da una importancia progresiva a los precios más recientes utilizando un sistema de ponderación o suavizado exponencial. La ponderación para los precios más antiguos decrece exponencialmente.

La siguiente imagen ha sido extraída de http://www.efxto.com/indicadores-mas-usados/medias-moviles/media-movil-exponencial. En ella se pueden apreciar la serie de precios de un activo en color verde para varios días, con datos para tres horas distintas en cada uno de los días. En la gráfica también se muestran dos tipos de suavizados en torno a esta serie de precios, uno de tipo exponencial, en color rojo, y otro de tipo simple, en color azul:

Figura 3.14: Representación de suavizado de tipo exponencial (curva roja) y de tipo simple (curva azul) junto a la serie de precios.

Cabe destacar que los suavizados de tipo exponencial son más eficientes que los suavizados de tipo simple y ponderados a la hora de adaptarse rápidamente a la tendencia de las fluctuaciones en los datos recientes, hecho que puede ser ventajoso o no dependiendo de la longitud del período de la tendencia que se quiera analizar. Si este período a analizar es corto, interesará el uso de un suavizado exponencial, y si es de mayor duración, interesará el uso de un suavizado simple.

Una vez suavizadas las series de precios, se pasan a calcular ahora las medias móviles.

Para calcular las medias móviles se puede establecer un período de días hacia atrás (α) variable:

$$Predicción_t = \frac{\widehat{s_t} - \widehat{s_{t-\alpha}}}{\widehat{s_{t-\alpha}}} \qquad (3.2)$$

Si este período de días es más grande, mayor será la influencia de los datos antiguos. Por el contrario, si se selecciona un período menor, se tendrán en cuenta datos más recientes para la predicción.

Las medias móviles calculadas sobre períodos de tiempo relativamente cortos responden rápidamente a cambios grandes en los precios de una acción, mientras que las calculadas sobre períodos más largos, responden de una manera más lenta a estos cambios bruscos en los precios de la acción.

Los datos reales, en este caso la serie diferencia de precios proporcional, pueden calcularse con una diferencia entre días hacia delante T, con la que posteriormente comparar las predicciones aportadas por las medias móviles:

$$Datos\ reales_t = \left(\frac{s_t - s_{t-T}}{s_{t-T}}\right) \qquad (3.3)$$

En la siguiente imagen se muestra una predicción (en color azul) a partir de una serie de precios original (en color rojo), junto a los datos reales (en color verde) para el mismo período donde se ha calculado la predicción:

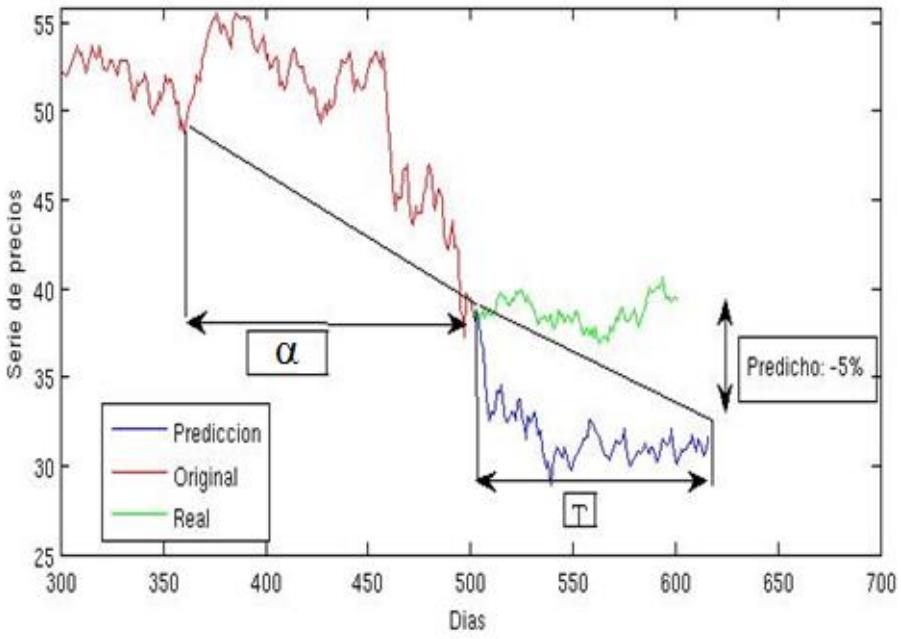

Figura 3.2: Ejemplo de predicción para T días teniendo en cuenta α días anteriores.

En la figura 3.2 se puede observar la diferencia entre los valores predichos para esos T días y los datos reales en ese mismo período.

Por tanto, el uso de las medias móviles nos aporta una visión acerca del comportamiento que podría tener más probabilidad de producirse en la serie en un futuro en base a comportamientos pasados de esa misma serie.

De este modo nos pueden indicar tanto tendencias alcistas en el precio de una determinada acción, como tendencias bajistas en el precio de la misma.

3.1.2. Principal Component Analysis (PCA)

PCA es un algoritmo utilizado para reducir la dimensionalidad de grandes conjuntos de datos en diversos campos científicos. Su aplicación, por ejemplo, en un campo como el de la biometría, queda patente en el reconocimiento facial, pero también tiene cabida en otras áreas como el reconocimiento de locutor, el procesado de imágenes en vídeo, etc. La aplicación de PCA en las series temporales financieras que se verá más adelante supone un hecho de más reciente incorporación.

PCA es un algoritmo el cual se engloba dentro de las técnicas de análisis multivariable [6]. Este tipo de análisis basa su funcionamiento en el análisis de los datos teniendo en cuenta las variables en conjunto en lugar de realizar un análisis de las variables de manera individual.

En conjuntos de datos con un gran número de variables, éstas suelen variar conjuntamente y no de manera individual. Esto es debido a que existen variables latentes (variables que no se observan directamente). El análisis multivariable por tanto, elimina la información redundante de un conjunto de variables y las combina en un grupo más reducido de variables.

El análisis de las componentes principales de los datos es un tipo de análisis multivariable, en el cual se busca una base que maximice la varianza entre los datos proyectados en ella. Las nuevas variables generadas son llamadas componentes principales de los datos. Estas componentes son combinaciones lineales de las variables originales, siendo también ortogonales entre ellas.

A continuación se muestra la imagen de una base no ortogonal a la izquierda y una base ortogonalizada a la derecha:

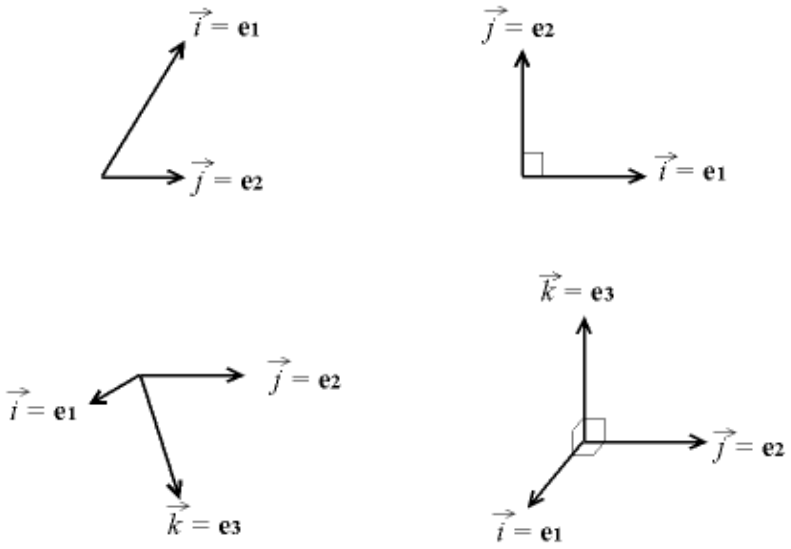

Figura 3.3: Ejemplo de una base no ortogonal a la izquierda y de una base ortogonal a la derecha.

La primera componente principal define un eje en el espacio *n*-dimensional, donde *n* es igual al número de variables existentes en el modelo. La primera variable originada a partir de la proyección de los datos iniciales sobre el eje definido por la primera componente principal, es la que mayor varianza contiene.

Matemáticamente [7], las nuevas *n* variables generadas $\Psi_{\alpha,}$ son combinaciones lineales a partir de las *n* variables originales x_i (donde *u* representa el autovector), cada una con una importancia medida por su varianza, igual a su autovalor, λ_α:

$$\Psi_\alpha = u_1 x_1 + u_2 x_2 + \ldots + u_n x_n \quad (3.4)$$

$$var(\Psi_\alpha) = \lambda_\alpha \quad (3.5)$$

La siguiente componente principal, es la segunda dimensión de más varianza, siendo esta dimensión ortogonal al resto de las dimensiones.

El resto de componentes principales siguen formando ejes, perpendiculares entre ellos, hasta formar dicho espacio *n*-dimensional, igual al número de variables iniciales del modelo.

En la siguiente imagen se muestra un ejemplo de representación de las componentes principales calculadas a partir de un conjunto inicial de datos de dos dimensiones mediante la aplicación de PCA:

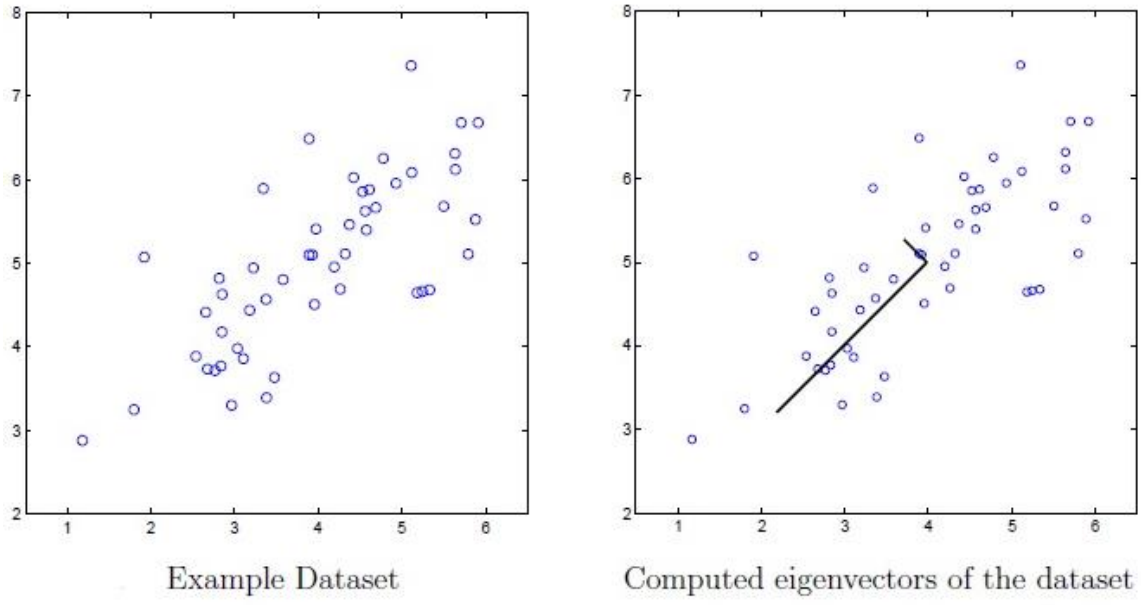

Figura 3.4: Representación a la izquierda de un conjunto de datos de dos dimensiones y a la derecha las componentes principales calculadas mediante PCA del conjunto de datos.

En la imagen de la derecha de la figura 3.4 se puede apreciar como el eje principal es el de mayor varianza de los datos. De esta forma se es capaz de identificar la dimensión de más "importancia" o varianza.

La ventaja de esta disposición de componentes principales reside en que seleccionando ahora las primeras componentes principales, se podrán proyectar el conjunto de datos originales sobre ellas reduciendo así la dimensionalidad inicial y extrayendo la mayor varianza posible de los datos con estas primeras componentes.

Por tanto, el uso de PCA permite reducir la dimensionalidad de un conjunto inicial de datos con la mínima pérdida de información posible. De tal manera que es capaz de devolver las dimensiones con mayor varianza, ordenadas de mayor a menor (varianza como unidad de información retenida), permitiendo seleccionar las más importantes y posteriormente proyectar el conjunto de datos original sobre estas dimensiones, consiguiendo de este modo reducir la dimensionalidad.

A continuación se enumeran los pasos para la implementación del algoritmo [4]:

Sea $x \in R^n$ el vector de características de los datos. El algoritmo realiza una proyección lineal de la forma:

$$y = W \cdot x \tag{3.6}$$

Siendo $y \in R^k \quad k < n.$

1. Se calculan la media y la varianza de los datos *x* y se normaliza el conjunto de datos por estos valores.

2. Se calcula la matriz de correlaciones (de covarianza si no se ha normalizado previamente) de los *x*, C:

$$C = x^T \cdot x \tag{3.7}$$

$$C = \begin{pmatrix} \boldsymbol{\sigma_A^2} & \sigma_{AB} & \sigma_{AC} & \cdots & \sigma_{AN} \\ \sigma_{BA} & \boldsymbol{\sigma_B^2} & \sigma_{BC} & \cdots & \sigma_{BN} \\ \sigma_{CA} & \sigma_{CB} & \boldsymbol{\sigma_C^2} & \cdots & \sigma_{CN} \\ \vdots & \vdots & \vdots & & \\ \sigma_{NA} & \sigma_{NB} & \sigma_{NC} & & \boldsymbol{\sigma_N^2} \end{pmatrix}$$

3. Se diagonaliza la matriz C con el objetivo de calcular los autovalores y autovectores de dicha matriz tal que:

$$C = U \cdot S \cdot U' \tag{3.8}$$

   siendo S una matriz diagonal con los autovalores y U una matriz ortogonal con los respectivos autovectores. Además se ordena en orden decreciente del tamaño de los autovalores.

4. Se seleccionan los k autovalores mayores. La matriz U_k formada por los *k* correspondientes autovectores es la matriz de transformación W.

La selección del valor *k* se puede determinar respecto al porcentaje total de la información que se quiera retener. El porcentaje total de información explicada por los *k* primeros autovalores/autovectores es:

$$I = \frac{\Sigma_j^k \lambda_j}{\Sigma_i^n \lambda_i} \tag{3.9}$$

En la gráfica que se muestra a continuación se puede ver un ejemplo de la información que se retiene para 174 series de nuestra base de datos en función del número de autovalores escogidos y de la frecuencia de los datos:

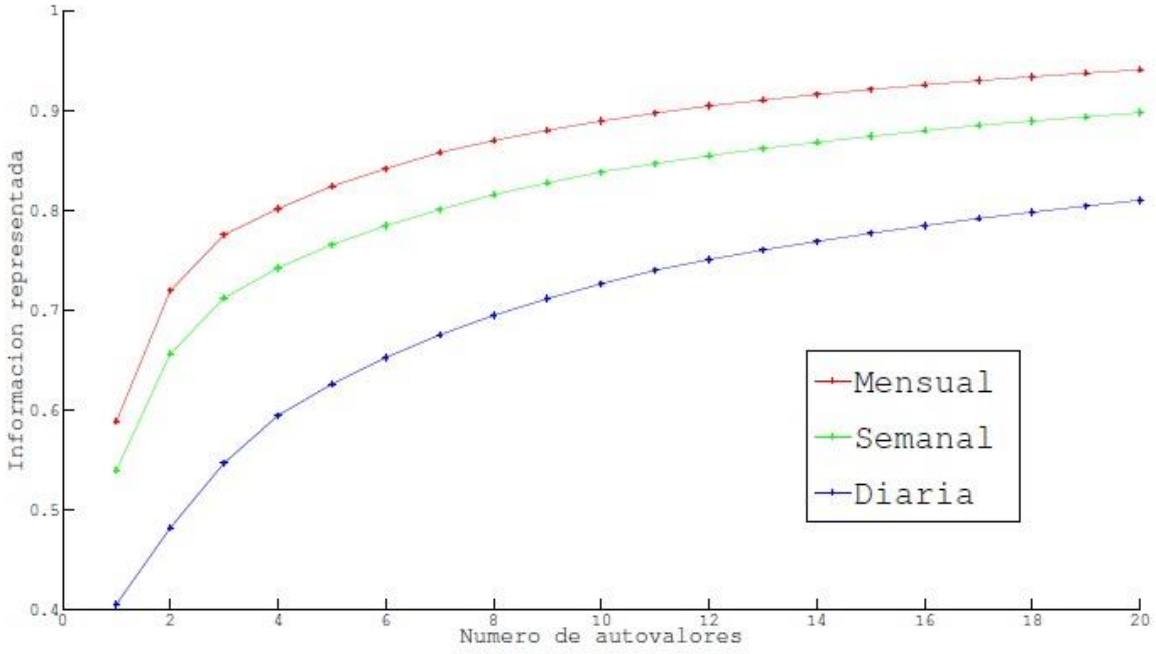

Figura 3.5: Información retenida en función del número de autovalores seleccionados para 174 series temporales financieras, para frecuencias de tipo mensual, semanal y diaria.

En la figura 3.5 se observa como para el primer autovalor, correspondiente a la primera componente principal, se es capaz de retener en torno al 50% de la información total, estando por debajo de este valor para la curva diaria, y por encima para las curvas semanales y mensuales (cuanto más grande es el período, más correlacionados están los datos). También se puede apreciar cómo hasta el sexto autovalor las pendientes de las tres curvas son mayores que a partir de dicho autovalor, tendiendo a estabilizarse después de él. Hecho que evidencia la menor información contenida por los restantes autovalores.

Los errores aumentan al disminuir la cantidad de información total, sin embargo, las necesidades computacionales son mucho menores. A pesar de ello, este aumento del error puede ser asumible en comparación con las ventajas que ofrece la disminución de las necesidades computacionales.

Después de proyectar los datos en este espacio de menor dimensionalidad, se puede recuperar [6]

una aproximación de los datos iniciales proyectando los datos de las componentes principales al espacio de alta dimensión original de partida:

$$x\_rec = y \cdot W \tag{3.10}$$

Siendo $x\_rec \in R^n$.

En la siguiente gráfica se muestran los mismos datos originales normalizados de la figura 3.4, en círculos de color azul, y los datos recuperados a partir de la proyección sobre la primera componente principal calculada con PCA, siendo ésta la de mayor varianza de los datos, con círculos de color rojo. En la gráfica también aparecen representadas unas líneas discontinuas las cuales unen cada dato original con el dato recuperado:

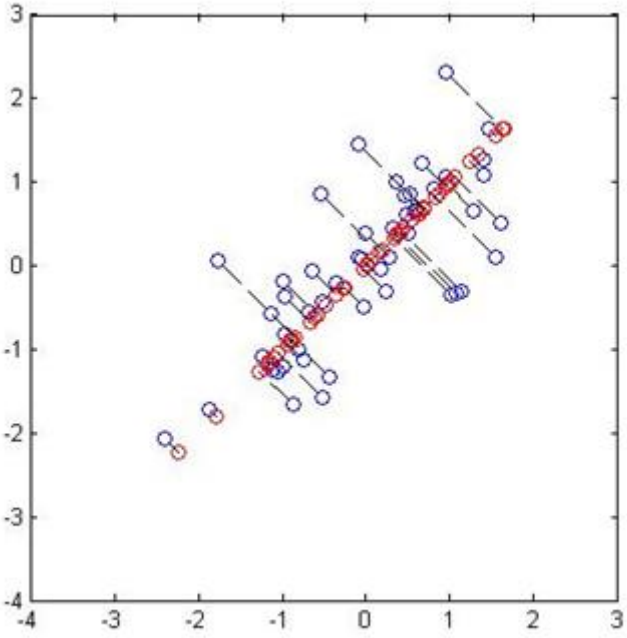

Figura 3.6: Representación de los datos originales normalizados con círculos azules y de los datos recuperados después de aplicar PCA con círculos rojos. Se representan también las líneas que unen a cada dato original con su dato recuperado.

A continuación se muestra un ejemplo gráfico obtenido de un capítulo introductorio de Machine Learning sobre PCA, pero ahora sobre un conjunto de datos de imágenes de caras:

Figura 3.7: Ejemplo de aplicación de PCA sobre un conjunto de imágenes de caras.

En la figura 3.7, en la imagen de la izquierda se aprecian las caras del conjunto de datos original. En la imagen del medio, se observa ya el resultado de calcular mediante PCA las componentes principales de los datos originales. Por último, en la imagen de la derecha se pueden ver las caras reconstruidas usando solamente el conjunto de datos proyectados sobre las primeras componentes principales.

3.1.3. PCA aplicado en series temporales financieras

En el apartado anterior se detallaba tanto el objetivo que logra PCA como los pasos seguidos para lograr su implementación. En este apartado, se pasa a hablar del uso de este algoritmo en nuestro objeto de estudio, las series temporales financieras, así como las ventajas que su aplicación nos aporta.

La base de datos de la que se dispone cuenta con un número de series financieras igual a 278, con información de cada una de ellas de 3065 días. En este caso, las variables son las series temporales financieras, y los datos, u observaciones, los días para los que se tiene la información de cada serie. Si esta cantidad de datos se compara con la cantidad de datos que se puede tener en otros escenarios como por ejemplo en las bases de datos utilizadas en voz, los datos de los que se disponen en este caso no resultan abundantes. Además, el acceso a esta base de datos es de carácter privado. Por tanto, los datos disponibles han de ser tratados de una manera lo suficientemente eficiente.

Las series financieras se encuentran correladas entre ellas, dado que en la realidad se producen acontecimientos que provocan movimientos en determinadas series, las cuales a su vez guardan algún tipo de relación con otras, provocando cambios también en estas últimas. Tras la aplicación de PCA, se tienen unas nuevas Eigenseries (autoseries) en el nuevo espacio proyectado, en el que se tratan estas Eigenseries (E) de manera individual, se realiza un suavizado ($\widehat{E}$) y se aplican las medias móviles para estimar las predicciones:

$$\widehat{E_t} = suavizado(E_t) \tag{3.11}$$

$$Predicción\ Eigenseries_t = \frac{\widehat{E_t} - \widehat{E_{t-\alpha}}}{\widehat{E_{t-\alpha}}} \qquad (3.12)$$

Sin embargo, después de realizar el algoritmo de predicción sobre ellas y proyectarlas de nuevo al espacio original, siguen conservando las relaciones existentes entre las series.

El siguiente diagrama de bloques trata de ilustrar el proceso de tratamiento aplicado sobre las series:

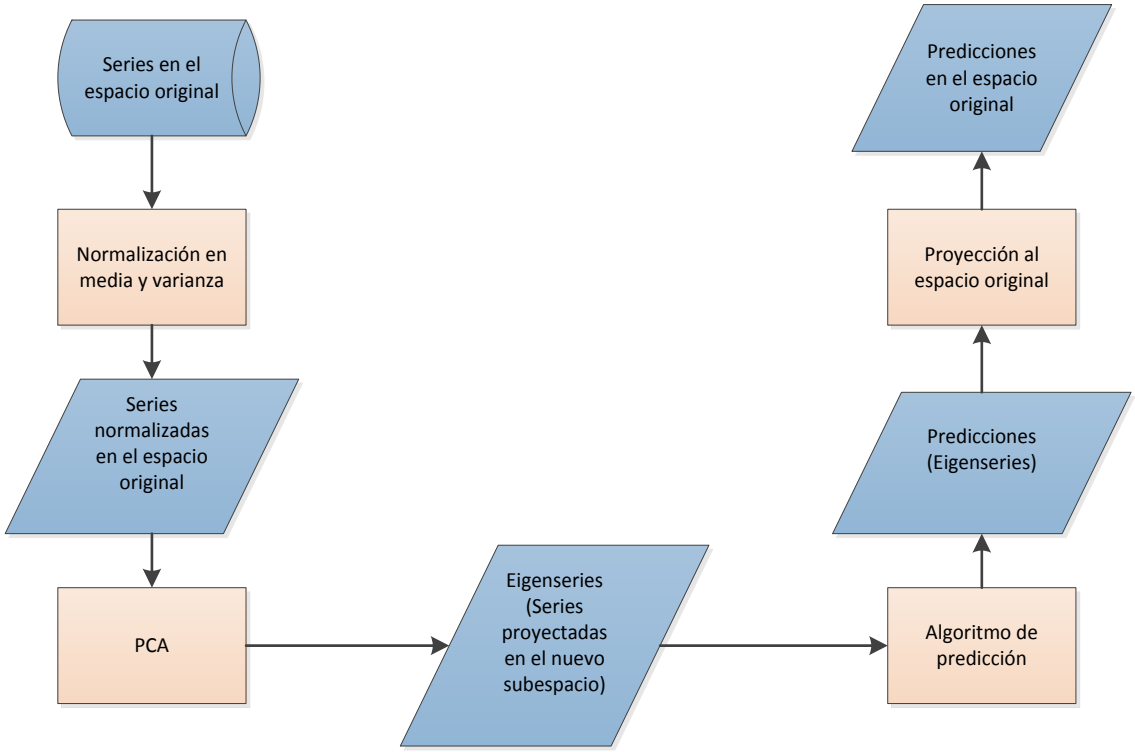

Figura 3.8: Proceso de tratamiento aplicado a las series financieras.

En la siguiente imagen se muestra un ejemplo de las series originales disponibles en la base de datos y de sus proyecciones en el nuevo espacio generado:

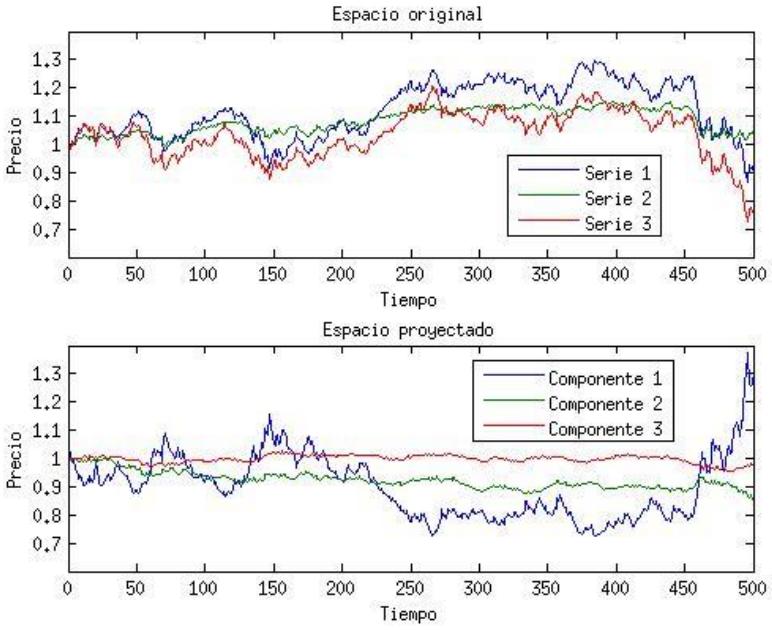

Figura 3.9: Serie reales y sus proyecciones en el espacio ortogonal.

La utilización de PCA en series temporales financieras se debe a las diversas ventajas que aporta su uso en nuestro estudio. Si hubiera que hablar de las ventajas más relevantes, estas dos serían las principales:

1. Se obtienen unas nuevas series Eigenseries (autoseries) incorreladas unas con otras, permitiendo trabajar serie a serie en el nuevo subespacio sin perder la información existente entre ellas, recuperándola después al proyectarlas de nuevo al espacio original.

2. Permite reducir la dimensionalidad de los datos.

3.2. Posicionamiento-sistemas de medición

El objetivo perseguido con la implementación de un nuevo algoritmo de predicción es el de mejorar los resultados obtenidos con anteriores métodos, es decir, conseguir estimaciones de futuros valores cada vez más próximas a los valores de las series reales. Estos resultados que se buscan mejorar no son más que medidas de distancia obtenidas entre las predicciones y los datos reales.

Si se quisiera invertir en una serie o conjunto de series con el objetivo de conseguir una determinada rentabilidad sin previo conocimiento de la situación financiera reciente, sería de gran utilidad tener una visión de la posición que ocupan estas series en las que se invertiría con respecto al resto de series existentes en el mercado.

En la siguiente gráfica se trata de ilustrar un ejemplo de dicho posicionamiento con valores ficticios, en el cual se representan dos series: una de ellas denominada Serie A, en color azul, y la otra Serie B, en color rojo, ambas curvas representando una supuesta serie de precios, donde se indica el porcentaje de rentabilidad calculado a partir de los valores de la serie de precios para cada una de las series:

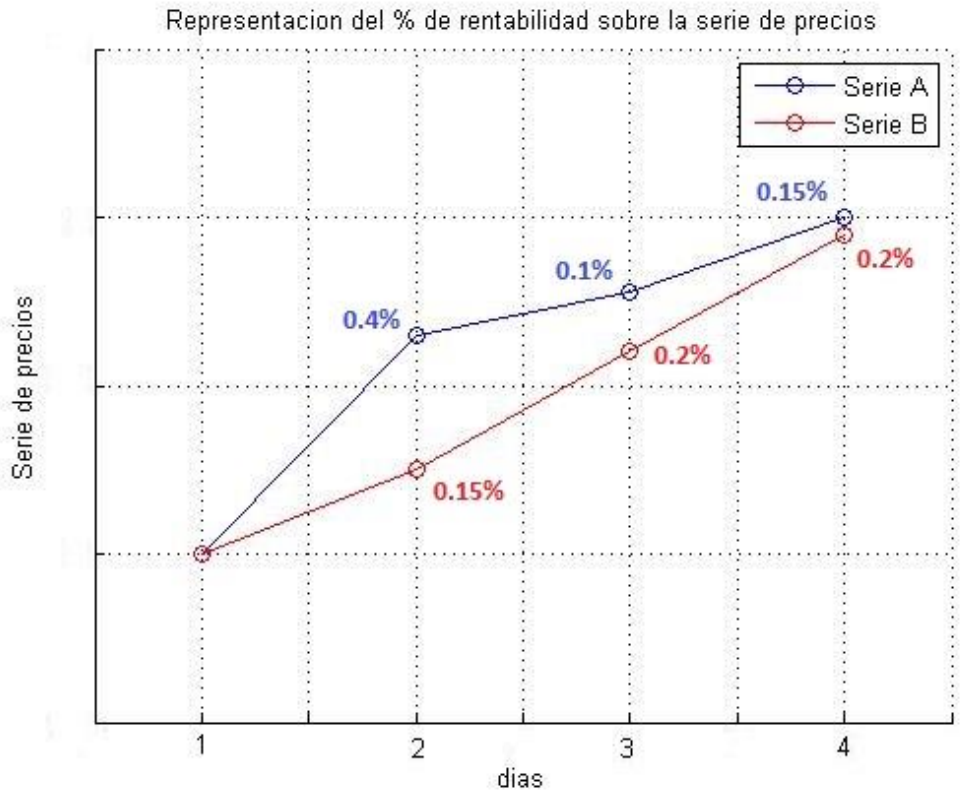

Figura 3.10: Representación de dos supuestas series de precios con valores ficticios para tratar de ilustrar un ejemplo de posicionamiento.

	Posición día 2	Posición día 3	Posición día 4	Rentabilidad total
Serie A	1º	2º	2º	0.65%
Serie B	2º	1º	1º	0.55%

En la anterior tabla se mostraban las posiciones que alcanzaban cada una de las series para los días en los que se disponían de datos, y también, la suma total de rentabilidad acumulada en esos días.
Si se buscara la serie que ha logrado obtener una mayor rentabilidad en este período de tiempo, la Serie A sería la que mayor valor ha alcanzado, con un 0.65% en comparación con el 0.55% alcanzado por la Serie B, pero si se buscara la serie la cual ha obtenido mejores posiciones en este período, la Serie B sería la que mayor número de veces ha alcanzado la primera posición, con un total de dos veces, en comparación con la Serie A, con un total de una.

Por tanto, este posicionamiento podría realizarse de dos maneras diferentes según se tengan en cuenta dos criterios distintos:

1. Las posiciones alcanzadas en base a la ordenación de rentabilidades.
2. La rentabilidad total acumulada.

Nosotros, en nuestro estudio, hemos realizado el posicionamiento según el primer criterio, dado que nos va a permitir realizar medidas más complejas donde se miden otras características adicionales.

Este tipo de posicionamiento consiste en ordenar las series en posiciones en base a los mayores

porcentajes de rentabilidad obtenidos con respecto a las otras series en el mismo período de tiempo a considerar. De este modo, la serie mejor posicionada sería la serie que se encuentra en la posición número uno, habiendo obtenido dicha posición debido a que su porcentaje de rentabilidad ha sido el mayor de entre las series a considerar para ese día.

En el presente proyecto, dicho posicionamiento se ha aplicado a los valores estimados de las series, de tal manera que se tienen ordenadas conforme a sus respectivos valores de rentabilidad predichos.

De este modo, los errores entre los datos reales y los datos sintéticos son medidos como distancia entre posiciones, es decir, diferencia entre las posiciones alcanzadas por los valores de las series estimadas con cada uno de los métodos de predicción con respecto a las posiciones que ocupa cada serie real.

En el siguiente apartado se explicarán los distintos medidores de error utilizados, los cuales nos servirán tanto para cuantificar la diferencia entre los datos reales y los datos sintéticos como para comparar entre los resultados obtenidos con cada uno de los métodos empleados.

3.2.1. Unidades de medida del error

Dado que hasta el momento no se han encontrado algoritmos capaces de generar predicciones de valores idénticas a los valores que se producen en la realidad, se tiene que hablar inevitablemente de errores cometidos en predicción. Dentro de la medición de estos errores, se han establecido diversos criterios y condiciones para su medida de manera que se han generado también distintos tipos de medición de errores.

En la siguiente tabla se muestran unos valores extraídos de una simulación en los que se muestran las rentabilidades y las posiciones para cinco series de los datos reales y de las predicciones con el método multivariable, es decir, el método el cual primero aplica PCA y después calcula las medias móviles:

	Datos reales		Predicciones (Multivariable)	
Series	Rentabilidad	Posición	Rentabilidad	Posición
Serie A	-0.0132	4º	0.0005	2º
Serie B	0.0045	1º	0.0003	3º
Serie C	0.0004	2º	0.0007	1º
Serie D	-0.0197	5º	-0.0004	5º
Serie E	-0.0062	3º	0.0001	4º

Error tipo 1

La primera medida de error utilizada y la cual recibió el nombre de error de tipo 1, es la que mide la diferencia entre las posiciones de las series predichas con respecto a las posiciones alcanzadas por las series reales. Una vez que se tienen calculadas estas diferencias de posiciones (en valor absoluto), se suman todas ellas y se divide entre el número total de series utilizadas para calcular estas diferencias:

$$Diferencia\ de\ posiciones = Posiciones\ (Real) - Posiciones\ (Predicha) \quad (3.13)$$

$$Error\ tipo\ 1 = \frac{1}{N}\sum_{1}^{N}|Diferencia\ de\ posiciones| \quad (3.14)$$

Donde N es el número total de series utilizadas, igual a 5 en el ejemplo.

A continuación se muestra un ejemplo de cálculo del error tipo 1:

Series	Diferencia de posiciones	Error tipo 1
Serie A	2	$\frac{2+2+1+0+1}{5} = 1.2$
Serie B	2	
Serie C	1	
Serie D	0	
Serie E	1	

Para el ejemplo con las anteriores cinco series, el error de tipo 1 calculado es igual a 1.2 posiciones, habiendo calculado las diferencias de posiciones en valor absoluto.

Error aleatorio

El error aleatorio surge de la necesidad de establecer una cota superior de referencia con la que poder comparar los resultados obtenidos con ambos métodos de predicción. Esta cota vendría determinada por la realización de una predicción aleatoria, dependiente del número de series sobre las que calcular dicha predicción. El error aleatorio se define como el número de errores medio cometido si la predicción de las series fuera completamente aleatoria y la hemos obtenido experimentalmente mediante un número de simulaciones sobre los datos.

Error tipo 2

En este nuevo tipo de error, se tendrán en cuenta solamente para el cálculo del error las series que se predicen con rentabilidad positiva, dado que las que se predicen con rentabilidad negativa no nos interesarían en un principio de cara a una futura inversión.

Para su cálculo, se calculan las diferencias de posiciones de todas las series que se predicen con rentabilidades positivas y se divide entre el número total de series utilizadas para calcular estas diferencias:

$$Error\ tipo\ 2 = \frac{1}{n}\sum_{1}^{n}|Diferencia\ de\ posiciones| \quad (3.15)$$

Donde *n* son las series únicamente que se predicen con rentabilidad positiva.

A continuación se muestra la misma tabla de datos inicial pero con las rentabilidades positivas remarcadas de las series predichas:

	Datos reales		Predicciones (Multivariable)	
Serie	Rentabilidad	Posición	Rentabilidad	Posición
Serie A	-0.0132	4º	0.0005	2º
Serie B	0.0045	1º	0.0003	3º
Serie C	0.0004	2º	0.0007	1º
Serie D	-0.0197	5º	-0.0004	5º
Serie E	-0.0062	3º	0.0001	4º

De tal manera que el cálculo del error de tipo 2 en el ejemplo es el siguiente:

Serie	Diferencia de posiciones	Error tipo 2
Serie A	2	$\frac{2+2+1+1}{4} = 1.5$
Serie B	2	
Serie C	1	
Serie D		
Serie E	1	

Resultando el error de tipo 2 de 1.5 posiciones, un valor mayor si se compara con el error de tipo 1 anterior.

Error tipo 3

En este tipo de error, también se tendrán en cuenta solamente las series predichas con rentabilidad positiva, con la inclusión además de una penalización mayor a los errores cometidos por las series que se predigan con mayor rentabilidad que a los errores cometidos por series que se predicen con menor rentabilidad.

Esto es debido a que en las series predichas con mayor rentabilidad se podría focalizar un mayor porcentaje de una futura inversión dada su mayor rentabilidad, quedando penalizadas en mayor medida de este modo dichas series.

El cálculo del error de tipo 3 es el siguiente:

$$Error\ tipo\ 3 = \frac{1}{n}\sum_{1}^{n}|(Diferencia\ de\ posiciones) \cdot Penalización| \quad (3.16)$$

Donde *n* son las series únicamente que se predicen con rentabilidad positiva.

La penalización aplicada sobre los errores cometidos en posiciones por las predicciones para cada serie se calcula de la siguiente manera:

- Se tienen en cuenta solamente las posiciones de las series con rentabilidades positivas:

	Predicciones (Multivariable)		
Serie	Rentabilidad	Posición	Posición
Serie A	0.0005	2º	2º
Serie B	0.0003	3º	3º
Serie C	0.0007	1º	1º
Serie D	-0.0004	5º	
Serie E	0.0001	4º	4º

- Se calcula el peso de la penalización para cada serie en función de su posición. Para ello se selecciona el número resultante de invertir las posiciones (de tal manera que la 1º serie pasaría a ser la 4º, seleccionando el número 4 en este ejemplo) y se divide entre el sumatorio del número de series con rentabilidad positiva (con n=4 en el ejemplo):

Serie	Posición	Nº resultante de invertir la posición	$\sum_{i=1}^{n=4} i = 10$	Penalización
Serie A	2º	3	10	3/10 = 0.3
Serie B	3º	2	10	2/10 = 0.2
Serie C	1º	4	10	4/10 = 0.4
Serie D				
Serie E	4º	1	10	1/10 = 0.1

Una vez calculada la penalización, se procede a calcular el error de tipo 3:

Serie	Diferencia de posiciones	Penalización	Error tipo 3
Serie A	2	3/10 = 0.3	$2 * \frac{3}{10} + 2 * \frac{2}{10} + 1 * \frac{4}{10} + 1 * \frac{1}{10} = 1.5$
Serie B	2	2/10 = 0.2	
Serie C	1	4/10 = 0.4	
Serie D			
Serie E	1	1/10 = 0.1	

Resultando un error de tipo 3 de 1.5 posiciones para este ejemplo.

Error tipo 4

Este error surge de la búsqueda de una penalización mayor para las predicciones que han cometido un gran error respecto a los datos reales. Este error de tipo 4 se calcula a partir del error de tipo 3 explicado antes, al cual se le aplica una función de penalización de la forma:

$$Error\ tipo\ 4 = Error\ tipo\ 3\ \cdot \log(Error\ tipo\ 3) \qquad (3.17)$$

Siendo log ( ) el logaritmo neperiano.

Resultando un error de tipo 4 en este caso de:

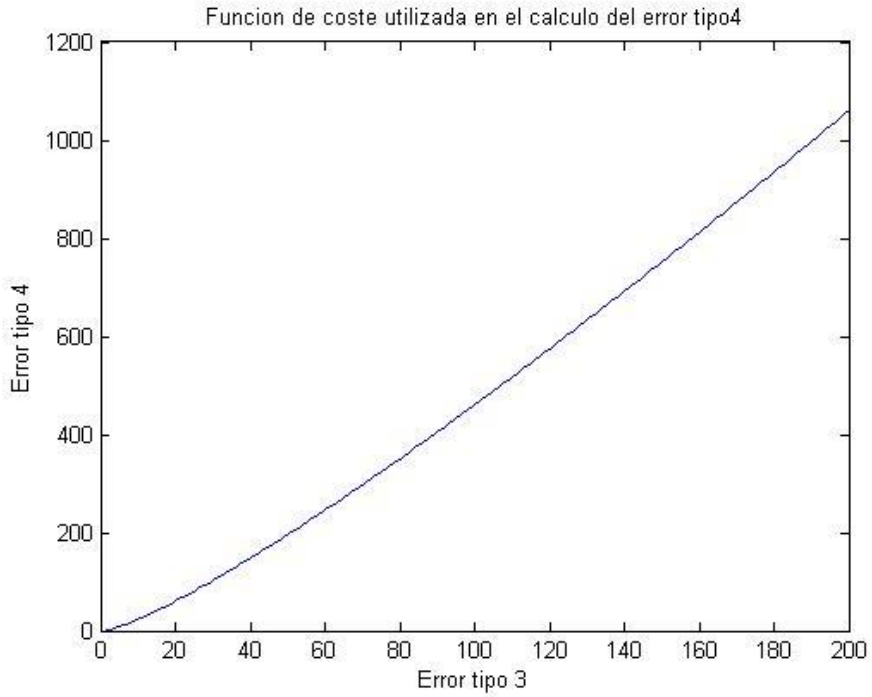

Figura 3.11: Función de coste utilizada para calcular el error de tipo 4 a partir del error de tipo 3.

De este modo, los grandes errores cometidos en predicción quedarán más penalizados que los errores más pequeños, debido a la no linealidad de la función de coste utilizada para el cálculo del error tipo 4.

4. Interfaz Gráfica de Usuario (GUI)

La base de datos de la que se dispone cuenta originalmente con un número de series igual a 278, con información de 3065 días. Como se puede observar en la figura 2.4 representada anteriormente, resulta complicado ver con claridad una determinada serie de la que se pueda extraer cierta información.

Surge entonces la idea de la implementación de una GUI, la cual permita por ejemplo visualizar las series junto a algunas de sus características sin la necesidad de lanzar una línea de comandos en Matlab cada vez que se quiera representar. Para ello, la interfaz carga previamente los datos que se elijan respecto del conjunto total disponible.

En las gráficas que se muestran en los siguientes apartados se han utilizado los datos de los dos mil primeros días (los mismos utilizados para el entrenamiento) y las 100 primeras series disponibles de la base de datos original.

En la siguiente imagen se muestra la GUI implementada finalmente:

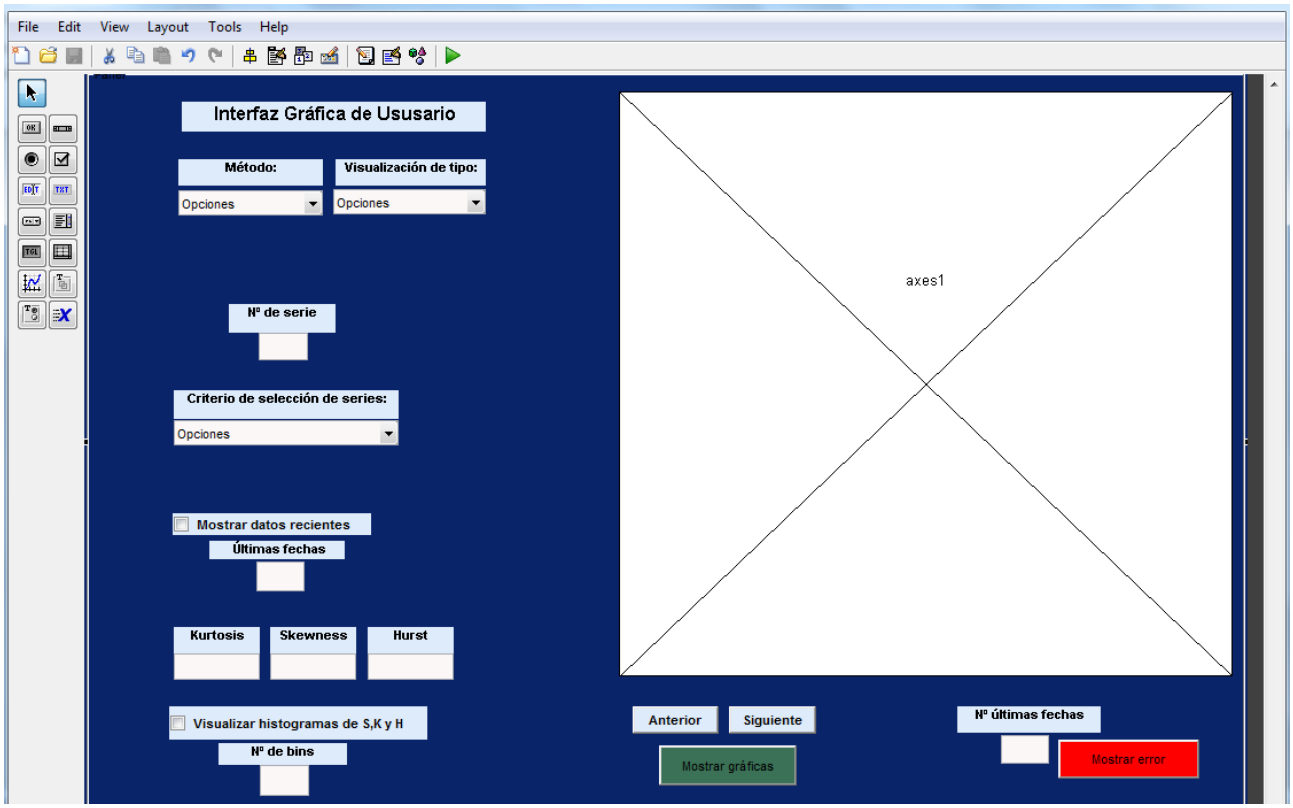

Figura 4.1: Interfaz Gráfica de Usuario implementada.

La GUI implementada se ha desarrollado completamente en entorno Matlab.

4.1. Funcionalidades de la GUI

Entre las diversas funcionalidades de la GUI se encuentran:

- La representación de series.
- La visualización de datos.
- La representación de errores de predicción.

Estas funcionalidades se detallan en los siguientes apartados.

4.1.1. Representación de series

La interfaz gráfica de usuario puede representar tanto las series reales de las que disponemos, como las series predichas calculadas con cada uno de los métodos, tanto con el método univariable como con el método multivariable. Esta opción se elige mediante un pop-up menú:

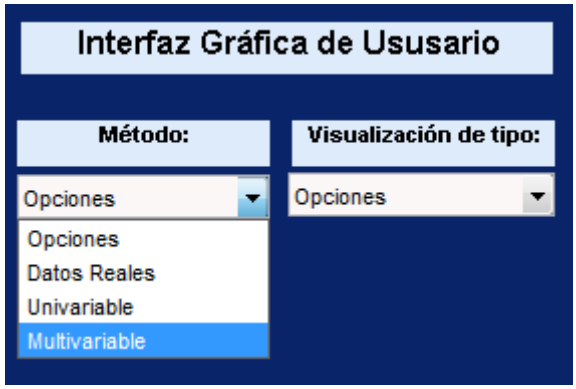

Figura 4.2: Elección del conjunto del cual se representa la serie.

Una vez se elija si representar una serie del conjunto de las reales o por el contrario una serie predicha con el método univariable o multivariable, se deberá escoger la frecuencia con la que se representa la serie, pudiendo ser ésta: diaria, semanal, bisemanal o mensual. Esta opción se elige también mediante un pop-up menú:

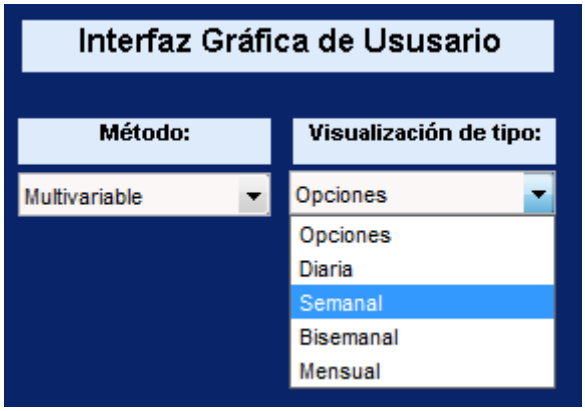

Figura 4.3: Elección de la frecuencia de representación de la serie.

En este caso se elige una frecuencia semanal, donde para este caso los retornos semanales se calculan de la misma manera que en la ecuación 2.7.

Posteriormente, se introducirá el número de la serie a representar, introduciendo este número directamente por la interfaz y seleccionando en el menú despegable la opción visualizar serie elegida:

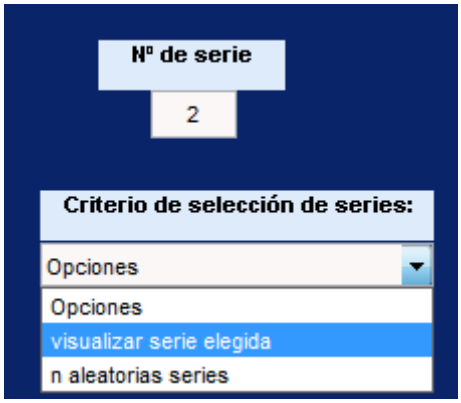

Figura 4.4: Introducción del número de serie o series a representar dependiendo del criterio de selección elegido.

La otra posible opción del menú despegable, n aleatorias series, representará simultáneamente un número de series aleatorias igual al número introducido por la interfaz.

Con las opciones seleccionadas que se han mostrado en las tres figuras anteriores, la interfaz genera la siguiente gráfica:

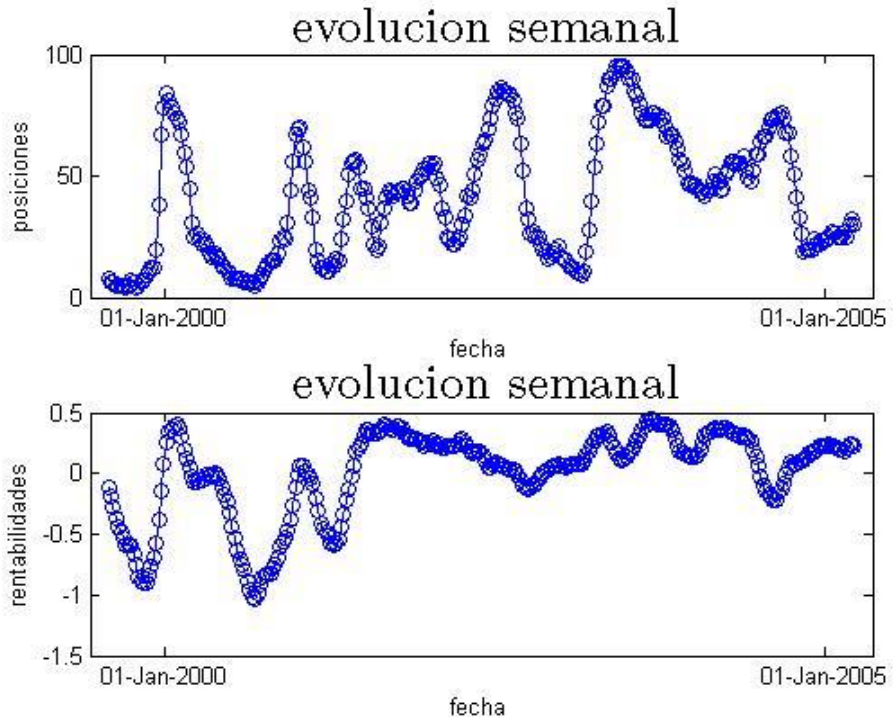

Figura 4.5: Posiciones y rentabilidades semanales alcanzadas por una serie.

En la gráfica anterior se observan tanto las rentabilidades como las posiciones semanales que ha alcanzado la serie introducida por la interfaz. En la representación de las posiciones, la mejor posición posible es la que se representa con un valor igual a 100 y la peor con un valor igual a 1.

Entre las opciones de esta interfaz, se puede encontrar también la posibilidad de representar en las gráficas los datos únicamente de las fechas más recientes. Para ello se debe seleccionar el Checkbox, e introducir el número de las últimas fechas a representar, con lo que se mostrarán las gráficas que contienen las posiciones y las rentabilidades de la serie o series aleatorias elegidas para estas últimas fechas (decir que el valor que se introduce en últimas fechas se interpreta como los últimos días si está seleccionada la representación diaria, como últimas semanas si está seleccionado la representación semanal, y del mismo modo para bisemanal y mensual):

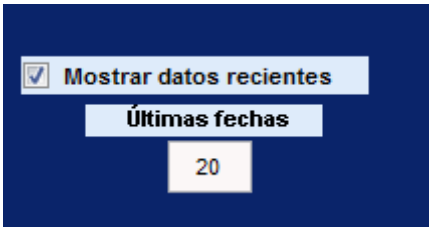

Figura 4.6: Posibilidad de representar los datos de las fechas más recientes y de elegir la cantidad.

Generándose la siguiente gráfica con esta última opción seleccionada:

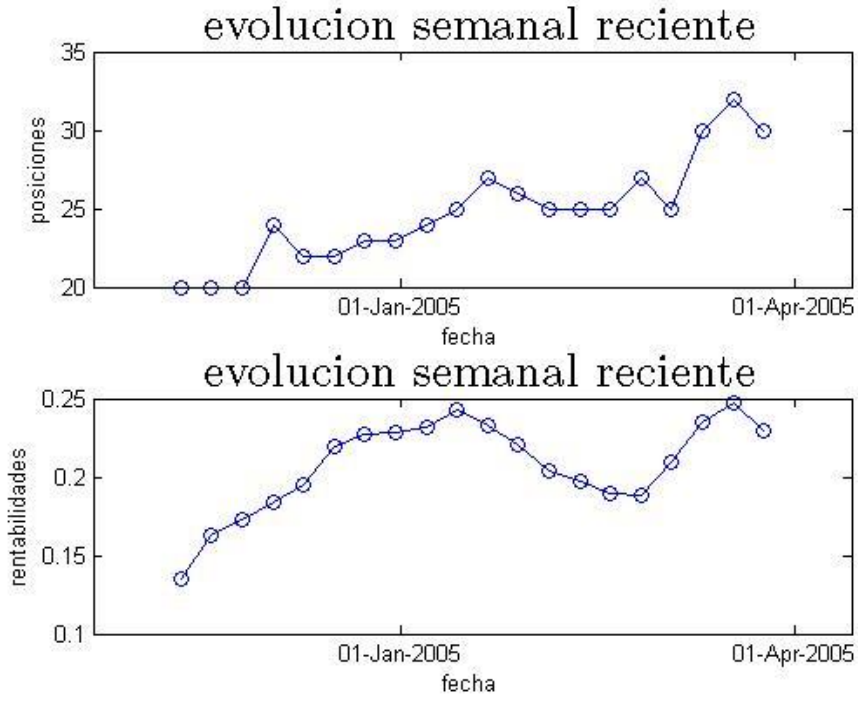

Figura 4.7: Posiciones y rentabilidades semanales más recientes alcanzadas por una serie.

Coincidiendo los datos de la figura 4.7 con los últimos 20 datos de las posiciones y rentabilidades de la figura 4.5 (esto se podría apreciar más claramente si se hiciera zoom en dicha ilustración).

4.1.2. Visualización de datos

En la interfaz también se muestran algunos datos acerca de las series representadas, como son los valores de características como la Kurtosis, el Skewness y el Hurst, para cada una de las series que se van introduciendo por la propia interfaz y que se hayan seleccionado del conjunto de las reales. Estos valores se calculan para las series reales cuando se ha seleccionado previamente la frecuencia de visualización diaria:

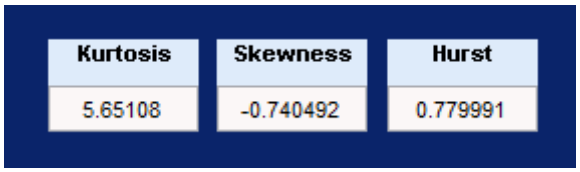

Figura 4.8: Visualización de la Kurtosis, el Skewness y el Hurst de la serie real seleccionada.

Adicionalmente, respecto a los valores de la Kurtosis, el Skewness y el Hurst, se da la opción de generar tres gráficos con los histogramas de estos valores calculados sobre las series reales disponibles en los datos cargados previamente por la interfaz, donde el número de intervalos utilizados en el histograma es introducido por el usuario a través de la interfaz:

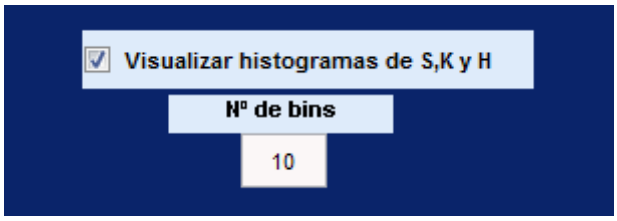

Figura 4.9: Seleccionando esta opción se mostrarán los histogramas de la Kurtosis, Skewness y Hurst de las series reales disponibles con el número de intervalos que se introduzca.

Generando en este caso los siguientes tres gráficos para un número de intervalos igual a 10:

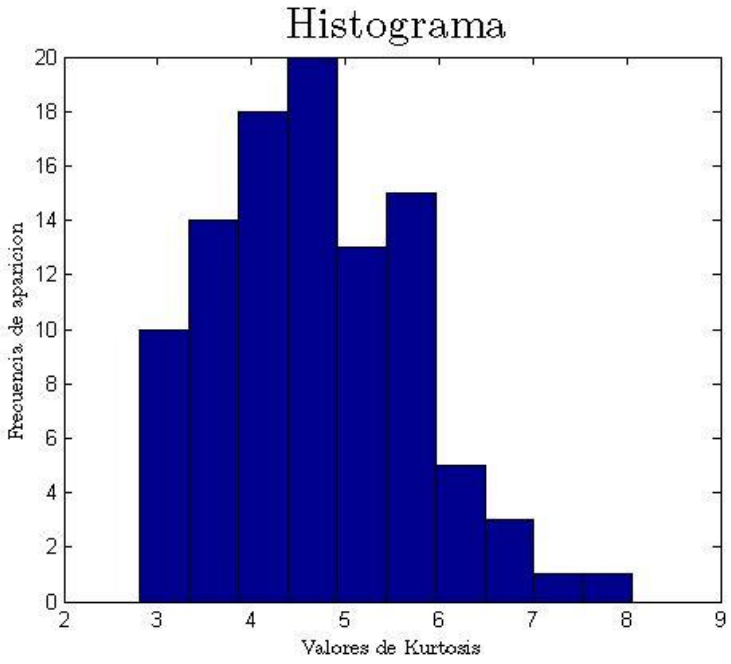

Figura 4.10: Histograma de los valores de Kurtosis de las series reales.

En la figura 4.10 se muestra el histograma de los valores de Kurtosis de las 100 series en los 2000 días de datos que utiliza en este caso la interfaz. La mayoría de valores son mayores que 3, por lo que se puede concluir que la probabilidad de que se produzcan eventos extremos, es decir, la probabilidad de que haya retornos alejados del valor medio, es elevada.

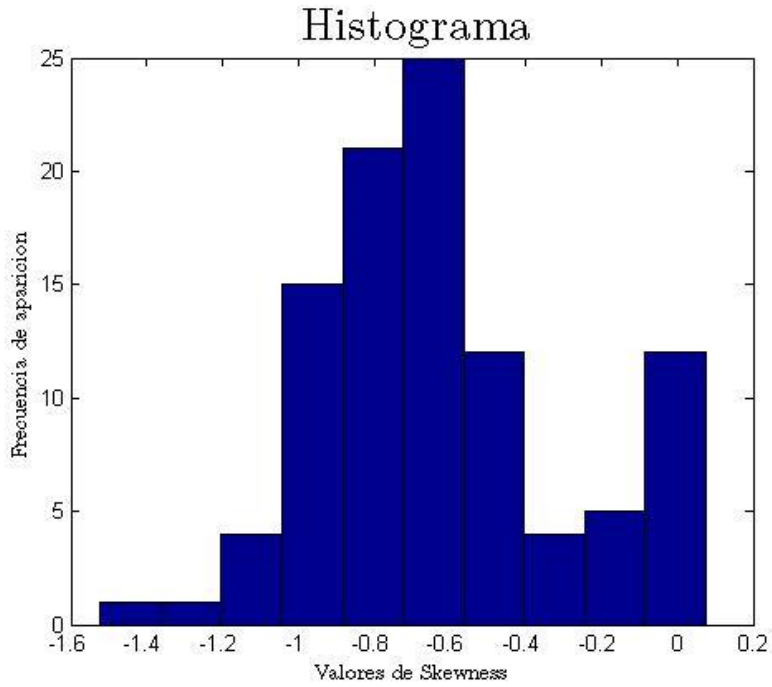

Figura 4.11: Histograma de los valores de Skewness de las series reales.

En el histograma de figura 4.11, la gran mayoría de valores está por debajo de cero, con lo que se deduce el hecho de que los retornos negativos "pesan" más que los retornos positivos en estas 100 series.

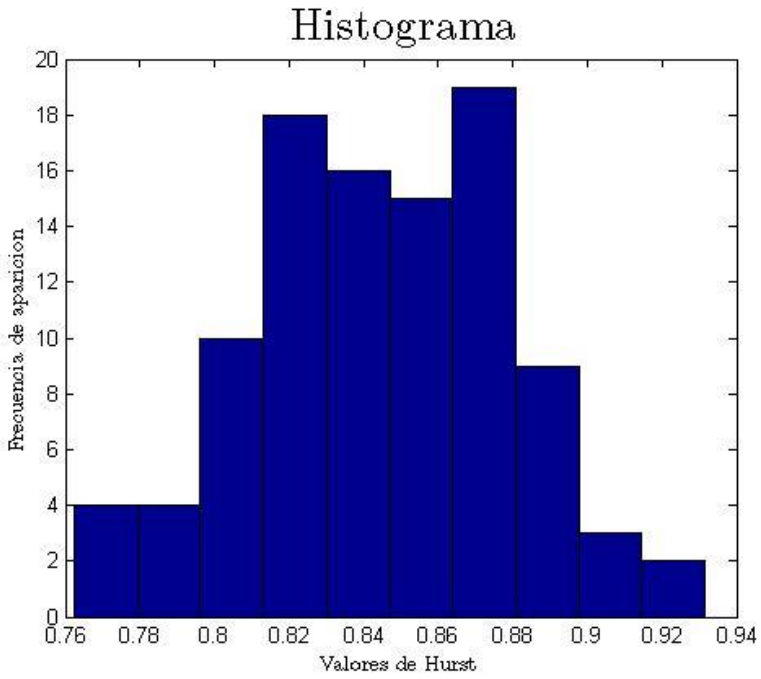

Figura 4.12: Histograma de los valores de Hurst de las series reales.

En la figura 4.12 se puede apreciar como todos los valores se encuentran por encima de 0.5 y por debajo de 1 para estas 100 series, con lo que se tendrá una correlación positiva entre incrementos.

4.1.3. Representación de errores de predicción

Otra de las funcionalidades de la GUI es la de representar los errores cometidos en predicción con ambos métodos de diferentes maneras. Todas estas gráficas que se representan se calculan para los días más recientes, siendo este valor introducido también mediante la interfaz:

Figura 4.13: Introducción del número de días más recientes de los que se mostrarán los errores.

A continuación se expondrán las distintas representaciones generadas por la GUI. Con los botones de anterior y siguiente de la interfaz que se muestran en la ilustración anterior, se podrá ir alternando la visualización de una u otra gráfica de entre las generadas.

En la siguiente gráfica se muestra el error medio cometido para cada día (de los días más recientes seleccionados, en este caso ochenta) con ambos métodos de predicción junto al error aleatorio, siendo este igual a 37.75, tal y como se explica en el apartado 3.2.1:

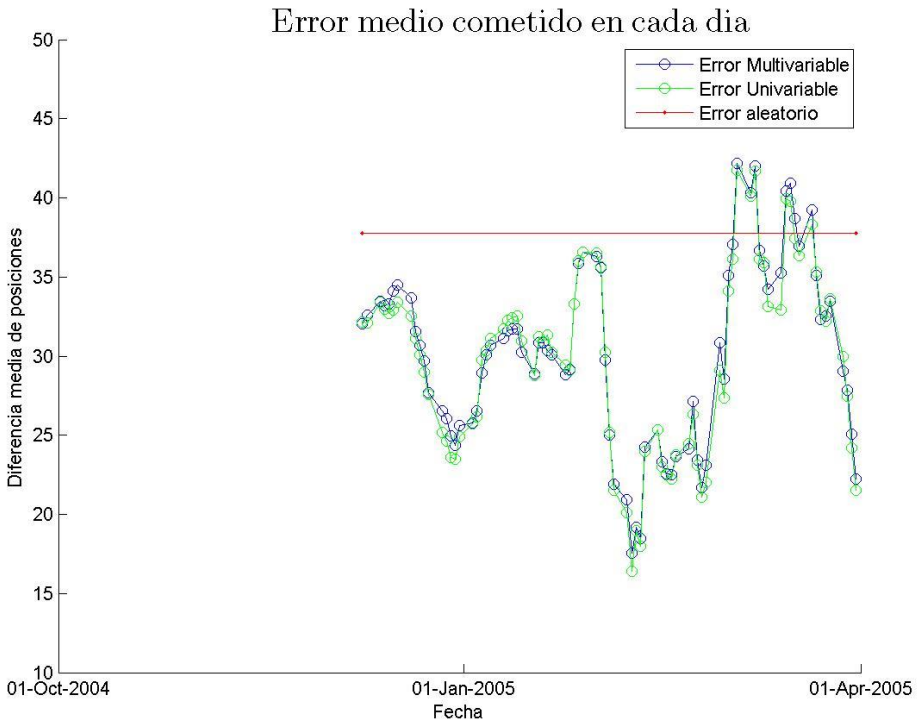

Figura 4.14: Representación del error medio cometido con cada uno de los métodos de predicción junto al error aleatorio para los días más recientes de los que se disponen datos.

En la figura 4.14 vemos que hay algunos días en los que los errores correspondientes a las predicciones son mayores que el error aleatorio, días en los que ambos errores están por debajo, días en los que una predicción es mejor que otra o incluso días con errores cometidos en predicciones muy similares.

En la siguiente imagen se representa el mismo tipo de error pero usando esta vez gráficos de barras, de manera que se pueda visualizar más claramente:

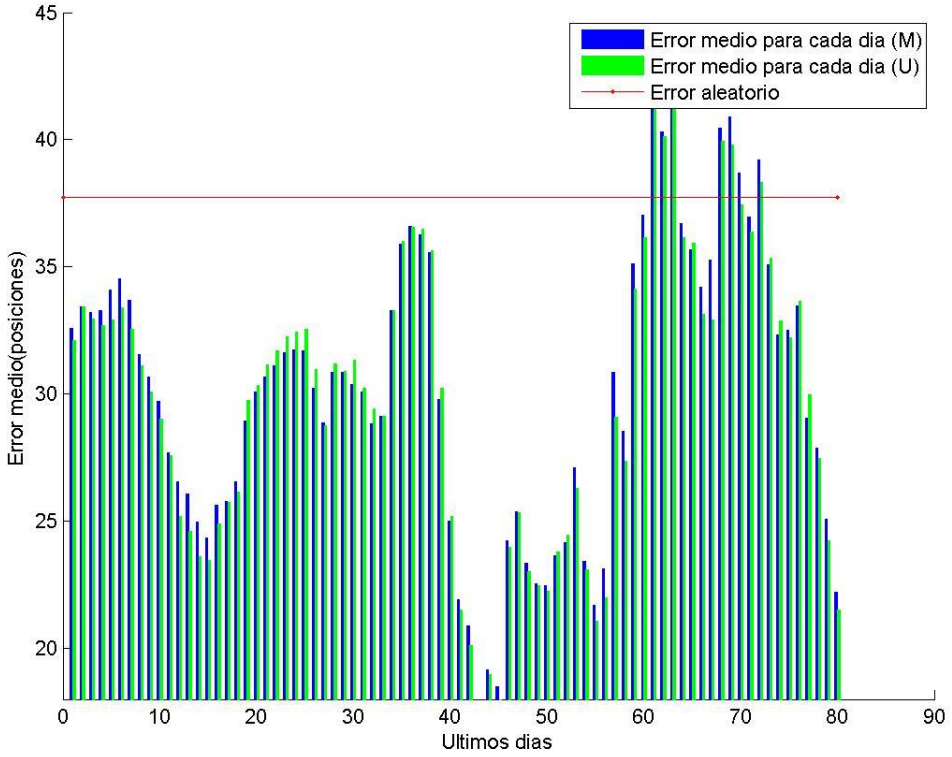

Figura 4.15: Representación del error medio cometido con cada uno de los métodos de predicción junto al error aleatorio utilizando gráficos de barras.

La siguiente imagen muestra un histograma de los valores de los errores de ambos métodos de predicción calculados a partir de los datos de las fechas más recientes:

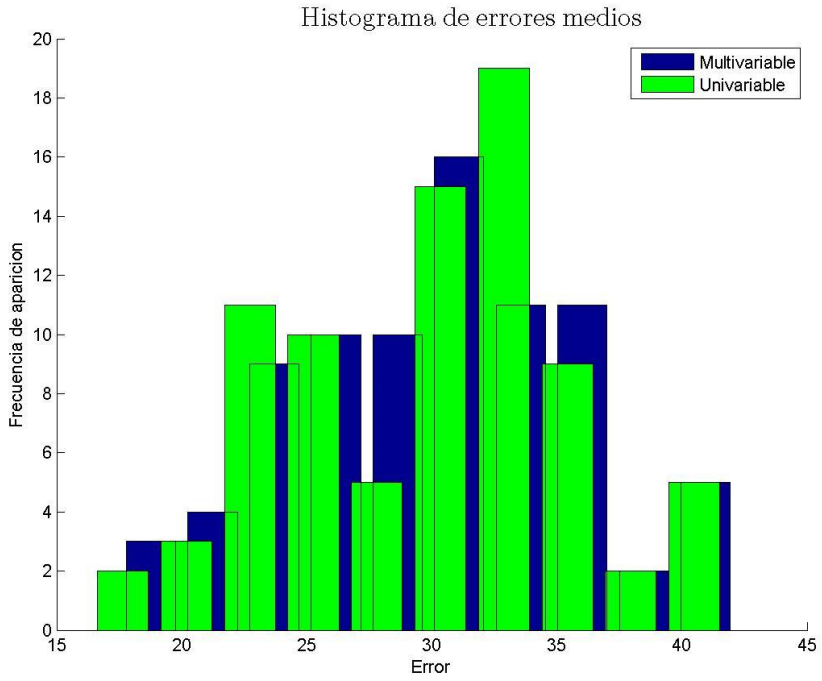

Figura 4.16: Histograma de los errores medios de posiciones cometidos con ambos métodos de predicción calculado con los datos de las fechas más recientes.

En el histograma de errores medios de la figura 4.16 se pueden apreciar los distintos errores medios cometidos en los últimos días seleccionados con cada uno de los métodos de predicción, donde la mayoría de valores se encuentran por debajo del error aleatorio (siendo este de 37.75 posiciones). Con respecto a los errores multivariables, se observa como hay tanto valores de error con menor frecuencia de aparición que los errores univariables, como valores con una frecuencia de aparición mayor.

En la siguiente gráfica que genera la interfaz se representa el ranking que ocupa una determinada serie según se tengan en cuenta sus posiciones alcanzadas o la suma total de sus rentabilidades, es decir, dependiendo que cual de los dos tipos de posicionamiento se realice. El 100 representa la mejor serie, y el 1 la peor, tanto para rentabilidades como para posiciones:

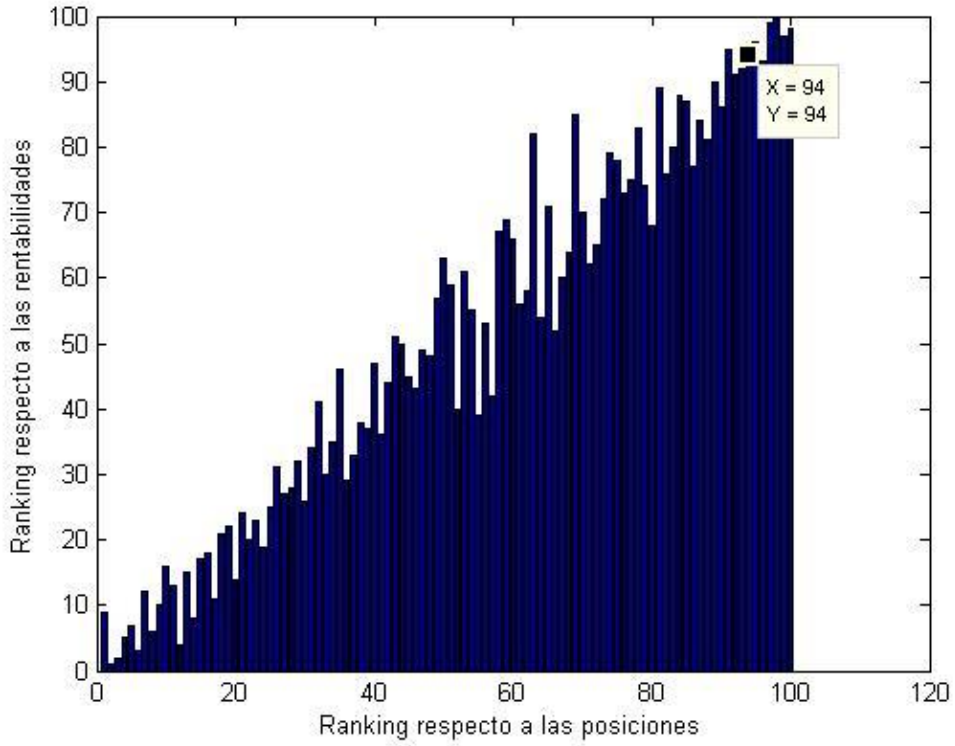

Figura 4.17: Ranking que ocupa cada serie para cada uno de los dos tipos de posicionamiento.

En la figura 4.17 se puede observar como el ranking que ocupa cada serie según se haya realizado un tipo de posicionamiento u otro es bastante similar en la mayoría de series, aunque no en todas. Por ejemplo, la serie que se encuentra en el ranking 94 respecto a las posiciones, ha obtenido el mismo puesto en el ranking teniendo en cuenta sus rentabilidades alcanzadas. Sin embargo, también se encuentran algunas series con las que se ha obtenido un ranking diferente con ambos tipos de posicionamiento.

En la siguiente imagen se trata de mostrar la misma idea anterior pero esta vez representada de una manera diferente:

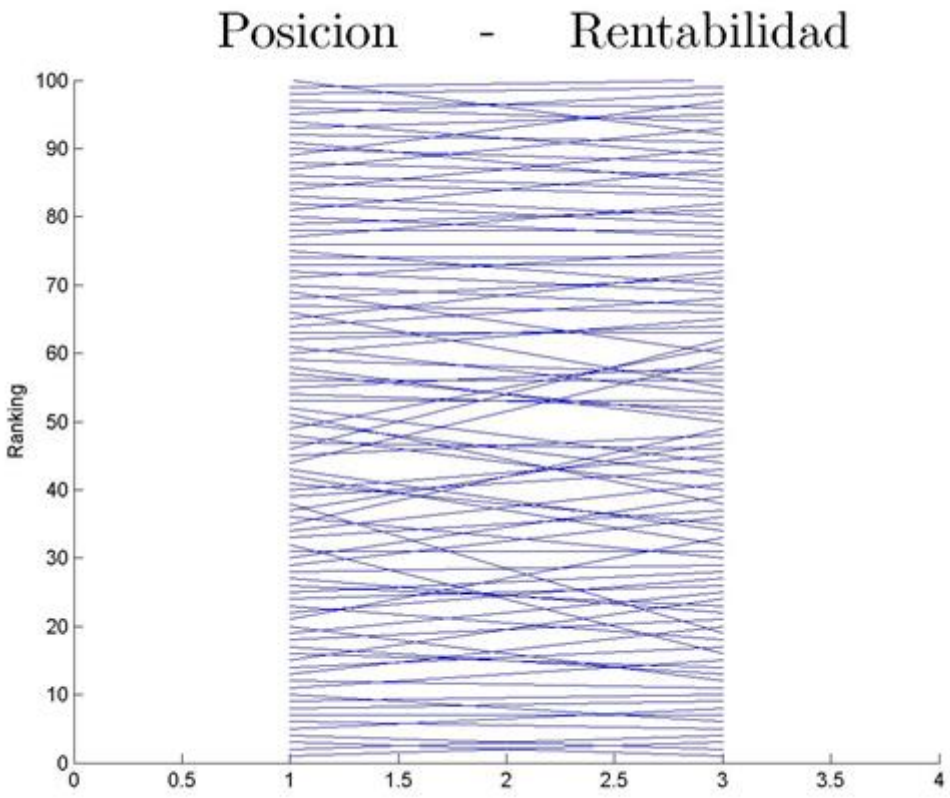

Figura 4.18: Unión mediante líneas de los dos rankings obtenidos para cada serie con cada uno de los posicionamientos.

En la figura 4.18 se unen los dos rankings alcanzados por cada serie mediante una línea azul. Las líneas que presentan una leve pendiente representan a las series con rankings similares para ambos tipos de posicionamiento, mientras que las líneas con mayor pendiente representan a las series con mayor diferencia de rankings obtenidos.

5. Experimentos

En este apartado se mostrarán los experimentos realizados a lo largo de este trabajo. Inicialmente se hablará del entorno experimental del trabajo realizado, para posteriormente pasar a detallar estos experimentos.

Entre los experimentos realizados por el autor se encuentra la comparativa entre dos mecanismos de predicción. Por un lado, aplicación de medias móviles en las series financieras, y por otro, aplicación primero de PCA sobre estas series con posterior cálculo de medias móviles. El siguiente diagrama de bloques trata de ilustrar dicho proceso:

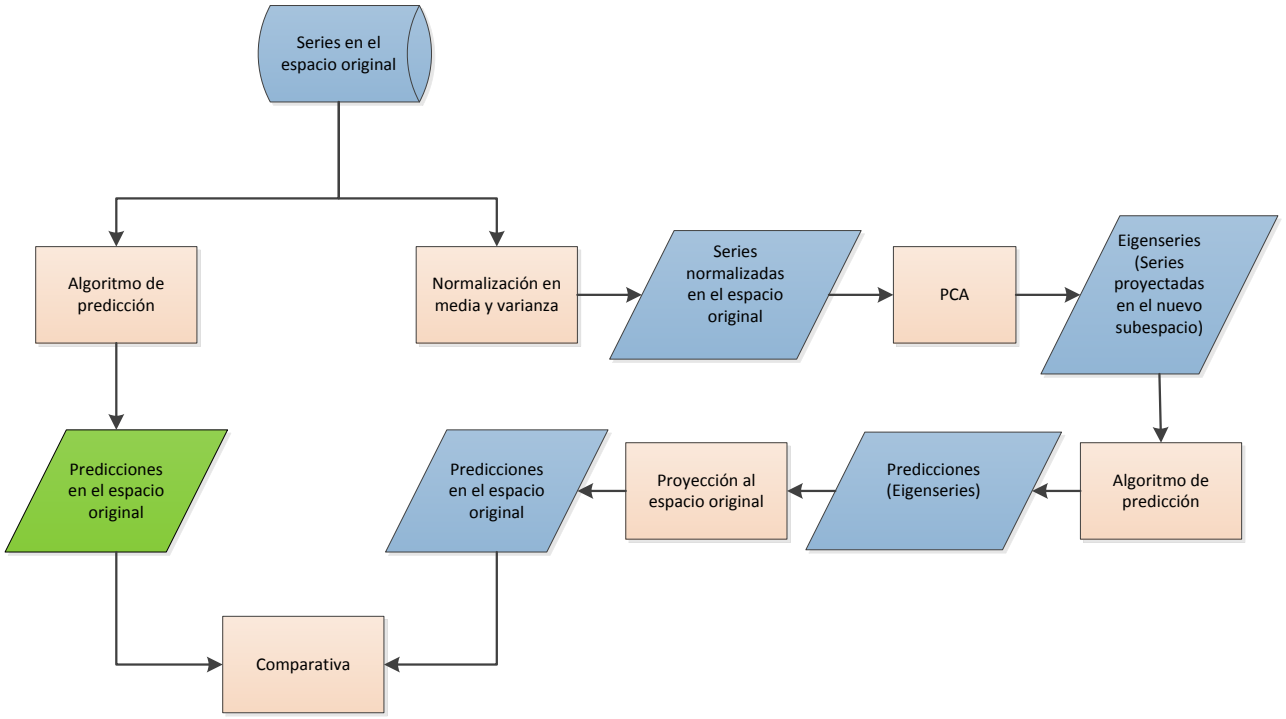

Figura 5.1: Representación del proceso seguido por las series financieras para calcular las predicciones con cada uno de los métodos y poder realizar la comparativa entre ellos.

También se analiza el resultado de usar determinados parámetros en predicción así como la búsqueda de una configuración óptima de ellos la cual aporte unos mejores resultados.

En este apartado de experimentos además se ha estudiado el resultado de aplicar las configuraciones de parámetros que mejor resultado han obtenido en predicción, en el generador de series sintéticas desarrollado por el grupo ATVS previamente.

Este objetivo consiste en analizar si esta configuración mejora o no la verosimilitud de las características de las series sintéticas generadas con respecto a las características que presentan las series reales comparando con la configuración de parámetros usada anteriormente en el generador.

En definitiva, todas las tareas realizadas se pueden resumir y ver sus relaciones entre ellas en la siguiente figura:

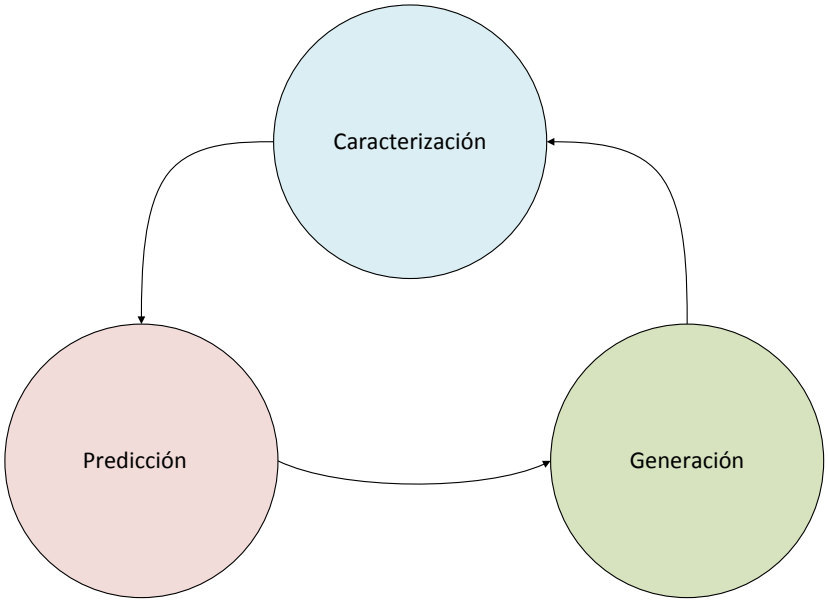

Figura 5.2: Resumen de las tareas realizadas y sus conexiones entre ellas.

Las conclusiones de los mejores resultados conseguidos en predicción se aplicarán en generación, donde la bondad de las series sintéticas generadas se medirá y comparará con las características explicadas en caracterización.

5.1. Entorno experimental

En este apartado se menciona el entorno de trabajo en el cual se han podido realizar los experimentos de este proyecto y las causas que motivaron su uso.

5.1.1. Software usado

El código implementado a lo largo de este trabajo y las gráficas generadas que se muestran han sido desarrollados en entorno Matlab debido a las ventajas que se exponen a continuación:

- Facilidad de su uso.
- Buena manejabilidad a la hora de realizar cálculos matriciales.
- Permite la inclusión de toolbox adicionales.
- Facilidad de generación y tratamiento de gráficas.

Para realizar este trabajo se han utilizado algunas funciones incluidas en los siguientes toolbox en Matlab:

- **Econometrics**: proporciona funciones para el modelado de datos económicos. Entre las funciones que se han utilizado pertenecientes a este toolbox se encuentran:

    - ret2price: calcula la serie de precios a partir de los retornos.
    - price2ret: calcula los retornos a partir de la serie de precios.

- **Curve fitting**: realiza un ajuste de curvas y superficies mediante suavizado (entre otros). La función que se utiliza de este toolbox es:

    - smooth: realiza el suavizado a la serie de precios.

- **Financial:** proporciona funciones para el modelado matemático y el análisis estadístico de datos financieros. Entre sus diversas funciones, se han utilizado:

    - mean: calcula la media aritmética.
    - std: calcula la desviación estándar.
    - cumsum: realiza la suma acumulativa.
    - hist: calcula y representa el histograma de los datos que recibe como argumento.
    - log: calcula el logaritmo natural.

- **Statistics**: proporciona algoritmos y herramientas para organizar, analizar y modelar datos. Las funciones utilizadas pertenecientes a este toolbox son:

    - skewness: calcula el Skewness de los datos que recibe como argumento.
    - kurtosis: calcula la Kurtosis de los datos que recibe como argumento.
    - ksdensity: calcula una estimación de densidad de probabilidad de las muestras que recibe como argumento.

5.1.2. Visualización de resultados

Los resultados que se muestran aparecen representados de dos formas distintas:

1. De forma gráfica. Se representan tanto gráficas con valores para distintos momentos temporales, como histogramas con frecuencias de aparición para ciertos valores.

2. De forma analítica. Se presentan resultados en forma numérica.

5.2. Experimentos de predicción

En este apartado se pasan a explicar los experimentos realizados en predicción de series financieras.

Para calcular las predicciones con cada uno de los dos métodos de predicción y medir los diferentes errores, se han utilizado los siguientes parámetros, los cuales han tomado diferentes valores a la hora de realizar los experimentos, con el objetivo de ver cómo afectan los distintos valores tomados por estos parámetros en la predicción y que combinación de ellos nos aportan unos mejores resultados:

- α
- T
- *Información*
- *Ponderación*
- *Normalización*
- *Días ventana PCA*

A continuación se definen los anteriores parámetros y sus valores utilizados:

- α: período de días hacia atrás que se tienen en cuenta para el cálculo de las medias móviles (ecuación 3.2).

Parámetro	Valores utilizados					
α (días)	1	5	10	20	50	80

- T: días hacia delante con los que calcular la serie diferencia de precios proporcional y comparar las predicciones (ecuación 3.3).

Parámetro	Valores utilizados			
T (días)	1	5	10	20

- *Información*: porcentaje de información retenida respecto del total (ecuación 3.9).

Parámetro	Valores utilizados		
Información (%)	98	99	100

- *Ponderación*: dependiendo del valor que se introduzca para este parámetro, se dará más o menos importancia a los datos de los últimos días. Esto se consigue mediante la repetición de los valores de las fechas más recientes, lo que afectará al nuevo cálculo de la media y la varianza de este conjunto de datos, aplicando después estos valores de media y varianza para la normalización del conjunto de datos original del primer paso del algoritmo PCA.

Parámetro	Valores utilizados		
Ponderación	Baja	Media	Alta

- *Normalización*: para este parámetro se tienen dos posibles valores. Uno de ellos normaliza la matriz de datos original, y el otro no la normaliza.

Parámetro	Valores utilizados	
Normalización	Si	No

- *Días ventana PCA*: con el valor de este parámetro se van seleccionando los datos de los *Días ventana PCA* días anteriores de las Eigenseries para calcular las predicciones, mediante la aplicación de medias móviles, de los $\frac{\text{Dís ventana PCA}}{2}$ días siguientes. De manera que supone un cálculo de las predicciones mediante la utilización de una ventana deslizante, que se va desplazando desde el primer día que se tienen datos hasta el último día de los 2000 días que se utilizan como entrenamiento.

Parámetro	Valores utilizados		
Días ventana PCA (días)	125	250	500

Se realizarán los experimentos con todas las combinaciones posibles de estos parámetros, es decir:

$$\#\,Combinaciones =$$

$$= \#\,\alpha \cdot \#\,T \cdot \#\,Información \cdot \#\,Ponderación \cdot \#Normalización \cdot \#\,Días\ ventana\ PCA$$

$$= 6 \cdot 4 \cdot 3 \cdot 3 \cdot 2 \cdot 3 = 1296\ combinaciones$$

Para realizar el experimento con estas 1296 combinaciones, se ha contado con 2000 días de entrenamiento, donde finalmente se han calculado para cada combinación los 4 diferentes tipos de error. Este experimento se ha realizado para un número de series igual a 200, número éste suficiente para poder extraer conclusiones de los métodos de predicción.

5.2.1. Resultados analíticos

En los siguientes apartados, se hablará del error multivariable como el error medido entre las predicciones aportadas con la aplicación de PCA primero seguido del cálculo de medias móviles con respecto a los datos reales, y de error univariable, para las predicciones aportadas con la aplicación de las medias móviles.

Se tratará de buscar la mejor configuración en el método multivariable de tal manera que se consigan mejorar los resultados aportados por el método univariable.

Con el objetivo de comparar la bondad de los resultados obtenidos con el método multivariable frente al método univariable, se ha dividido la matriz de datos que contiene los errores calculados con el método multivariable entre la matriz con los datos del método univariable resultantes de la realización del experimento, de tal manera que:

$$matriz\ resultado = \frac{errores\ Multivariable}{errores\ Univariable} \qquad (5.1)$$

La finalidad de esta división es la de comparar ambos métodos. Si se analizan los datos, los valores por debajo de 1 encontrados en la matriz de resultados indicarán un mejor resultado obtenido con el método multivariable, es decir, un menor error cometido en predicción con el método multivariable frente al univariable:

$$matriz\ resultado < 1 \quad => \quad error\ Multivariable < error\ Univariable$$

En este punto nos vamos a centrar en los valores menores que la unidad que aparecen en la ecuación 5.1, los cuales se han obtenido como resultado de aplicar una configuración específica de los seis parámetros que se han variado, y para uno de los cuatro diferentes tipos de error, con la finalidad de encontrar una determinada configuración que mejore los resultados con el método multivariable frente a los datos obtenidos anteriormente con el método univariable.

De entre todos los valores menores que uno, el mínimo valor de entre todos ellos encontrado en la ecuación 5.1 representa el mejor resultado obtenido con el método multivariable respecto al método univariable. El siguiente paso será por tanto el de determinar los valores de estos parámetros con los que se ha producido este mínimo. Para ello, se ha implementado una función la cual devuelve dichos valores.

El siguiente paso será analizar para el error de tipo 1 los resultados obtenidos con estos valores óptimos, es decir, con los valores de los parámetros con los que mejor resultado se ha obtenido con el método multivariable con respecto al univariable.

Error tipo 1

Para el error tipo 1, el cual mide la diferencia entre las posiciones estimadas de las series con cada uno de los métodos con respecto a las posiciones alcanzadas en la realidad, se muestran a continuación los valores de los parámetros con los cuales se ha producido el mínimo valor resultante en la ecuación 5.1:

	Valores óptimos para los parámetros
	Error tipo 1
α (días)	5
T (días)	20
Información (%)	99
Ponderación	Baja
Normalización	No
Días ventana PCA (días)	500
Valor resultante	**0.9927**

Estos valores de los parámetros coinciden con los del menor error de tipo 1 obtenido con el método multivariable, teniendo éste un valor de 90111 posiciones en los 2000 días usados (resultando así un error medio para cada día de unas 45.05 posiciones). El error con esta misma configuración con el método univariable es de 90777 posiciones (teniéndose en este caso un error medio diario de unas 45.38 posiciones), resultando de la división entre ambos el valor mostrado en la anterior tabla.

Los pasos siguientes consisten en analizar la estabilidad del valor resultante variando un parámetro y manteniendo fijos el resto de ellos, así como la extracción de conclusiones a raíz de la utilización de un valor u otro para cada parámetro.

En primer lugar, se varían los diferentes valores usados de los parámetros α y *T*, manteniendo fijos el resto de parámetros:

	Valores óptimos para los parámetros
	Error tipo 1
Información (%)	99
Ponderación	Baja
Normalización	No
Días ventana PCA (días)	500

En la siguiente tabla se muestran los valores resultantes de calcular el error medio multivariable entre el error medio univariable con dichas condiciones:

$\frac{error\ tipo\ 1\ M}{error\ tipo\ 1\ U}$	*T* = 1	*T* = 5	*T* = 10	*T* = 20
α = 1	0.9964	0.9950	0.9939	0.9930
α = 5	0.9976	0.9951	0.9945	**0.9927**
α = 10	0.9983	0.9949	0.9951	0.9933
α = 20	0.9991	0.9960	0.9952	0.9978
α = 50	1.0017	1.0035	1.0032	1.0022
α = 80	1.0034	1.0077	1.0082	1.0104

Analizando los resultados anteriores con respecto a α, se puede apreciar cómo se han conseguido unos mejores resultados con el método de predicción multivariable respecto al univariable para valores de α menores o iguales que 20, lo que refleja que tener en cuenta cierta cantidad de días o valores superiores hacia atrás supone un empeoramiento progresivo de los resultados ofrecidos por el sistema multivariable respecto a los ofrecidos por el sistema univariable, es decir, el método univariable funciona mejor para valores superiores de α igual a 20.

Esto puede ser debido a que tener en cuenta datos con cierta antigüedad puede no tener demasiada relación o influencia con lo que está sucediendo actualmente, debido a la distancia que los separa en el tiempo y/o a las distintas características de las circunstancias o acontecimientos actuales.

Respecto al número de días usados en *T*, se observa como para valores de α menores o iguales a 20, a medida que este número de días aumenta, los resultados reflejan un mejor comportamiento del sistema multivariable frente al univariable pese a este aumento de *T*. Por tanto, parece que las bondades de la predicción multivariable son mayores cuanto mayor sea el horizonte de predicción. Dado que en la anterior tabla solamente se pueden apreciar los valores medios resultantes de la división entre errores, se muestran más abajo los histogramas de estos errores para cada uno de los métodos de predicción, donde se podrá visualizar la forma de la distribución de los errores, a partir de la cual poder realizar un segundo análisis.

A continuación se muestran los histogramas de los datos semanales de los errores de tipo 1 calculados a partir de los resultados aportados por el método de predicción univariable (a la izquierda) y multivariable (a la derecha). Los histogramas muestran para cada uno de los cuatro

valores de *T*, los errores calculados para todos los días hacia atrás (α) tenidos en cuenta en el cálculo de las medias móviles:

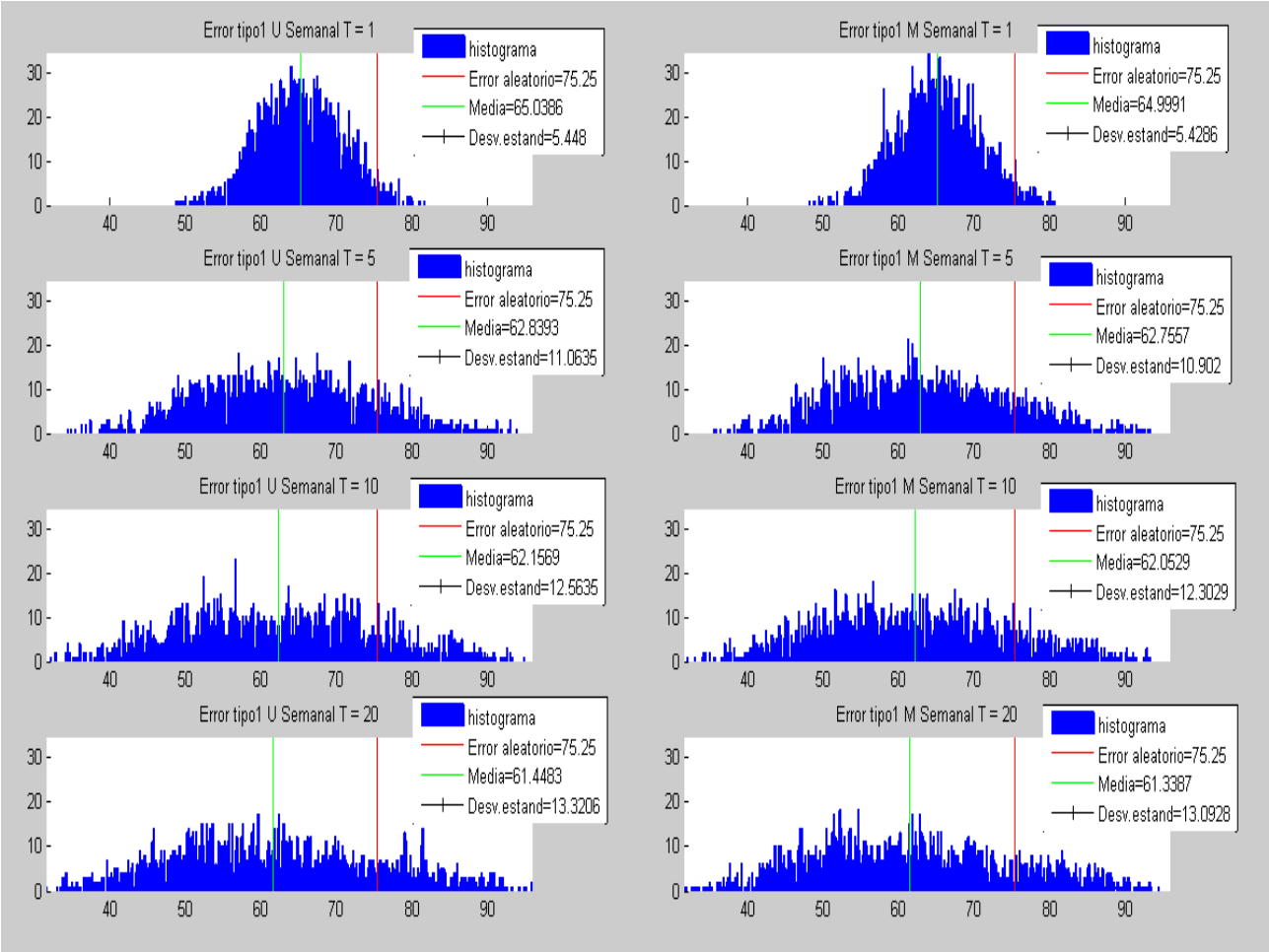

Figura 5.3: Histogramas de los errores de tipo 1 semanales obtenidos con los métodos univariable (a la izquierda) y multivariable (a la derecha) para distintos horizontes de predicción.

A la vista de las ocho gráficas diferentes, se puede observar como la media de los errores (línea verde) es menor que el error aleatorio (línea roja) en todos los casos, con lo cual con ambos métodos de predicción se obtienen unos mejores resultados que realizando una predicción de manera aleatoria, hecho este motivante para continuar optimizando el algoritmo de predicción multivariable.

Comparando los resultados del método univariable con los del multivariable, se observa un valor medio de error en posiciones menor en multivariable, para cada uno de los cuatro distintos valores del horizonte de predicción *T*, reflejando este hecho una mejor predicción con este método. Además, la desviación estándar también es menor en multivariable para los cuatro valores de *T*, lo que implica una menor dispersión de los errores con el método de predicción multivariable, presentando errores más estables concentrados en torno al valor medio (siendo menor el riesgo de obtener grandes errores).

Respecto a los distintos valores que toma *T*, a medida que éstos aumentan (tanto en un método como en otro), la media del error disminuye, pero la desviación estándar aumenta, por lo que se aprecia una mayor dispersión de los errores (esto es, distribuciones más "ensanchadas" con un rango más amplio de errores probables), encontrando por ejemplo errores de predicción más

grandes (aparición de errores por encima de 80 para valores de *T* mayores que 1) y también más pequeños (aparición de errores por debajo de 50 para valores de *T* mayores que 1), lo que supone un riesgo mayor en predicción debido a la incertidumbre de los posibles valores de error.

En segundo lugar, se varía el parámetro *Información*, manteniendo el resto de parámetros fijos:

	Valores óptimos para los parámetros
	Error tipo 1
α (días)	5
T (días)	20
Ponderación	Baja
Normalización	No
Días ventana PCA (días)	500

Obteniendo un valor resultante para cada uno de los tres valores distintos del parámetro *Información* de:

$\frac{error\ tipo\ 1\ M}{error\ tipo\ 1\ U}$	*Información* (98%)	*Información* (99%)	*Información* (100%)
Valor resultante	0.9932	**0.9927**	0.9961

Con un porcentaje del 99% de la información retenida, se consigue obtener el mejor resultado de los distintos estudiados. Aunque pueda parecer un porcentaje alto, cercano al 100%, con dicho valor se consigue reducir bastante la dimensionalidad del conjunto de series (es decir, se usa un número de series menor que el total disponible) y obtener un error con el método de predicción multivariable mejor que el que se obtiene con el univariable. El 100% de la información no aporta unos mejores resultados debido a la inclusión de los autovectores con un menor autovalor asociado, siendo éstos los que tienen mayor porcentaje de ruido [4], justificando en cierto modo también la reducción de la dimensionalidad.

En tercer lugar, se varía el parámetro *Ponderación*, manteniendo el resto de parámetros fijos:

	Valores óptimos para los parámetros
	Error tipo 1
α (días)	5
T (días)	20
Información (%)	99
Normalización	No
Días ventana PCA (días)	500

Obteniendo un valor resultante para cada uno de los tres valores distintos del parámetro *Ponderación* de:

$\frac{error\ tipo\ 1\ M}{error\ tipo\ 1\ U}$	*Ponderación* (Baja)	*Ponderación* (Media)	*Ponderación* (Alta)
Valor resultante	**0.9927**	0.9932	0.9929

Como se puede observar, el mejor resultado se ha obtenido con un nivel de ponderación bajo. Esto puede deberse a que para niveles más altos de ponderación, es decir, mayor repetición de valores dados (lo que repercute en los valores de la media y la varianza que se utiliza para normalizar el conjunto de datos original en el algoritmo de PCA), se puede producir un sobreapendizaje de los datos, obteniendo unos peores resultados en consecuencia. En definitiva, los resultados muestran que es mejor generalizar con más datos.

En cuarto lugar, se varía el parámetro *Normalización*, manteniendo el resto de parámetros fijos:

	Valores óptimos para los parámetros
	Error tipo 1
α (días)	5
T (días)	20
Información (%)	99
Ponderación	Baja
Días ventana PCA (días)	500

Obteniendo un valor resultante para cada uno de los dos valores distintos del parámetro *Normalización* de:

$\frac{error\ tipo\ 1\ M}{error\ tipo\ 1\ U}$	*Normalización* (Si)	*Normalización* (No)
Valor resultante	1.0961	**0.9927**

En este caso, la no normalización del conjunto de datos ofrece unos mejores resultados para el método multivariable con respecto al univariable. Esto puede ser debido a la naturaleza de los datos con los que se trata, presentando valores medios en torno a 0, y una varianza pseudoestacionaria, con lo que no habría una gran diferencia significativa entre los datos normalizados y no normalizados, sin embargo, ésta sería suficiente para perder el contenido diferenciador entre las series.

En quinto lugar, se varía el parámetro *Días ventana PCA*, manteniendo el resto de parámetros fijos:

	Valores óptimos para los parámetros
	Error tipo 1
α (días)	5
T (días)	20
Información (%)	99
Ponderación	Baja
Normalización	No

Obteniendo un valor resultante para cada uno de los tres valores distintos del parámetro *Días ventana PCA* de:

$\frac{error\ tipo\ 1\ M}{error\ tipo\ 1\ U}$	*Días ventana PCA* (125 días)	*Días ventana PCA* (250 días)	*Días ventana PCA* (500 días)
Valor resultante	0.9991	0.9962	**0.9927**

A la vista de los datos anteriores, se aprecia cómo con un mayor número de días utilizados para el tamaño de la ventana de PCA se ofrecen unos mejores resultados con el método multivariable. Por tanto, para calcular las correlaciones, parece ser mejor usar una cantidad significativa de días. De hecho, con 500 días, coincide como el mejor valor para todos los tipos de error.

La lista completa con todos los resultados obtenidos se pueden encontrar en el anexo.

5.2.2. Conclusiones de experimentos de predicción

Las conclusiones de predicción extraídas de los experimentos realizados indican que la bondad de los resultados ofrecidos con el método multivariable dependen en gran medida de los valores que tomen todos los parámetros utilizados: α, *T*, *Información*, *Ponderación*, *Normalización* y *Días ventana PCA*:

- Los valores óptimos de α se encuentran generalmente en 50, o lo que es lo mismo, dos meses de datos. Podemos decir, por tanto, que los dos últimos meses son los datos que más información aportan. Usar menos días supone unos peores resultados, y usar más, generaliza demasiado los datos.

- El algoritmo de predicción multivariable mejora al algoritmo univariable según aumenta el horizonte de predicción, *T*. Por tanto, su utilidad principal parece encontrarse en predicciones no a corto plazo, en las que tiene mejor estabilidad que el método univariable.

- El porcentaje de información retenida con mejores resultados es de un 99%. Por tanto, nuestros mejores resultados se encuentran cuando reducimos dimensionalidad y quitamos series del cálculo. Es decir, reducir la dimensionalidad no sólo va a disminuir las necesidades computacionales, sino que además permite eliminar "ruido" y mejorar las predicciones.

- La ponderación de los datos es muy dependiente del sistema de medición usado. Con lo que aunque parece no ser muy determinante, no deja de ser un parámetro configurable más que poder optimizar en nuestros cálculos.

- La normalización de los datos es perjudicial para la predicción. El motivo principal es que las series se encuentran todas con medias y varianzas muy parecidas. Normalizar ha supuesto perder parte del contenido diferenciador entre las series.

- El resultado más óptimo para el parámetro *Días ventana PCA* se ha conseguido con 500 días, es decir, dos años. Este resultado es razonable ya que para entrenar el sistema con tantas series se requiere un número suficiente de datos como para poder calcular correctamente la matriz de correlaciones. Sin embargo, usar datos anteriores a dos años supone entrenar el sistema con datos muy antiguos, en los que las correlaciones entre los activos eran diferentes a las actuales.

5.2.3. Parámetros óptimos en predicción

A continuación se muestran los valores de los parámetros que mejores resultados han obtenido para cada uno de los cuatro tipos de error:

	Valores óptimos para los parámetros			
	Error tipo 1	Error tipo 2	Error tipo 3	Error tipo 4
α (días)	5	50	50	50
T (días)	20	20	20	20
Información (%)	99	99	99	99
Ponderación	Baja	Alta	Media	Media
Normalización	No	No	No	No
Días ventana PCA (días)	500	500	500	500
Valor mínimo	0.9927	0.987	0.981	0.9779

5.3. Experimentos de Caracterización – Generador de series sintéticas

En este apartado se analizará la bondad de utilizar los valores óptimos de los parámetros utilizados en predicción en la generación de series sintéticas, con el fin de determinar si esta configuración ofrece una mayor similitud o no de las series sintéticas generadas con respecto a las series reales en comparación con los valores utilizados anteriormente en la generación.

5.3.1. Generación con los parámetros óptimos de predicción

Los parámetros y sus valores óptimos resultantes en predicción que tienen aplicación en la generación de series sintéticas, se muestran en la siguiente tabla, junto a los valores usados anteriormente en la configuración del generador, a los cuales nos referiremos más adelante como los valores generados con la configuración no óptima en predicción:

	Valores óptimos para los parámetros en predicción (Error tipo 1)	Valores usados anteriormente en generación
Información (%)	99	100
Ponderación	Baja	Baja
Normalización	No	Si

5.3.2. Medidores de bondad

Para determinar la bondad de los resultados obtenidos con cada una de las configuraciones utilizadas, se han calculado las funciones de distribución acumuladas (CDF) de algunas de las características de las series, tanto de las series reales como de las series sintéticas, así como la diferencia entre las CDF. Entre las características sobre las que se calculan las diferencias se encuentran:

- el Skewness
- la Kurtosis
- el exponente de Hurst
- el Sharpe Ratio[2]

Calculándose estas cuatro características a partir de las series reales y también de las series sintéticas que proporciona el generador.

El número de simulaciones utilizado en el generador con el que se calculan las diferencias entre los datos sintéticos y los reales es de 50 simulaciones. Se ha usado este valor debido a que con un número igual o mayor de simulaciones los datos sintéticos generados presentan histogramas similares de las cuatro características para distintas generaciones realizadas, es decir, se tienen ya suficientes muestras, donde los valores de los datos generados no se diferenciarán en exceso entre las distintas simulaciones.

A continuación se muestran los histogramas de una de las cuatro características, del exponente de Hurst, para dos generaciones distintas usando un número de simulaciones igual a 50:

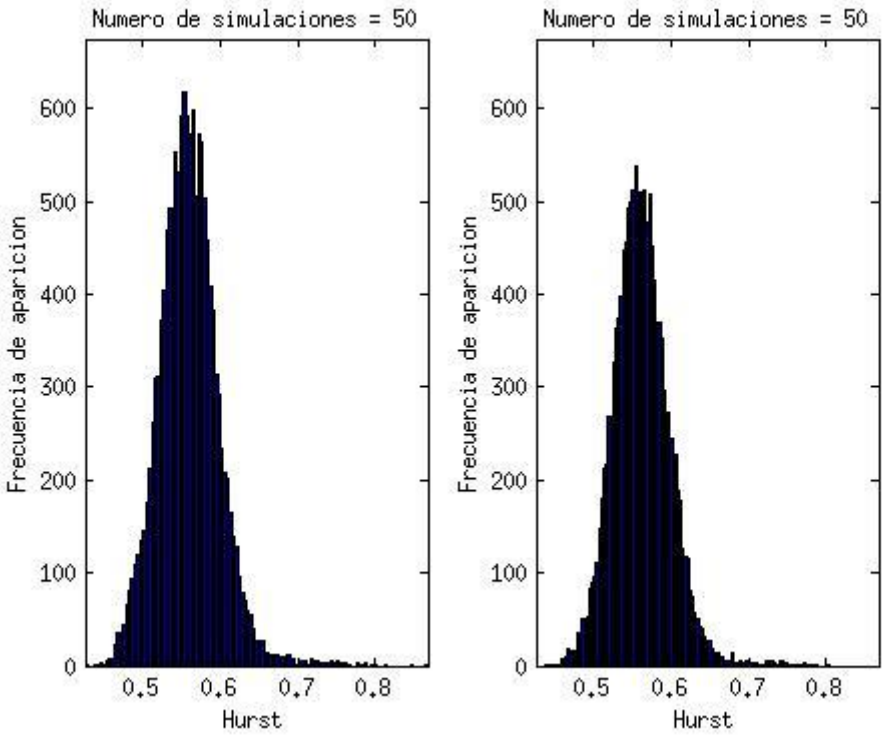

Figura 5.4: Histogramas del exponente de Hurst para dos generaciones distintas.

En la figura 5.4 se observa a simple vista una forma similar de ambos histogramas, donde se puede ver también como la mayoría de valores del exponente de Hurst se concentran entre 0.5 y 0.6 para ambas generaciones, con una diminución de la concentración de valores a partir de 0.65 aproximadamente.

[2] Uno de los mecanismos más usados para medir la bondad de una inversión es el Sharpe Ratio, que puede interpretarse como la cantidad de retorno obtenida por cada unidad de riesgo tomada (optimización media/varianza). El objetivo por tanto es el de maximizar el Sharpe Ratio.

5.3.3. Resultados

En este apartado se muestran las gráficas de las Cumulative Distribution Function (CDF) de las cuatro características para cada una de las configuraciones utilizadas así como los valores de las diferencias calculados entre las CDF.

La siguiente gráfica muestra las CDF de cada una de las cuatro características de los datos reales junto a las CDF calculadas a partir de los datos sintéticos generados con la configuración no óptima en predicción. Para representar los valores de las CDF y calcular las diferencias entre ellas, se ha definido un eje de abscisas común para los datos sintéticos y reales, estando comprendido entre el mínimo y el máximo valor de cada característica:

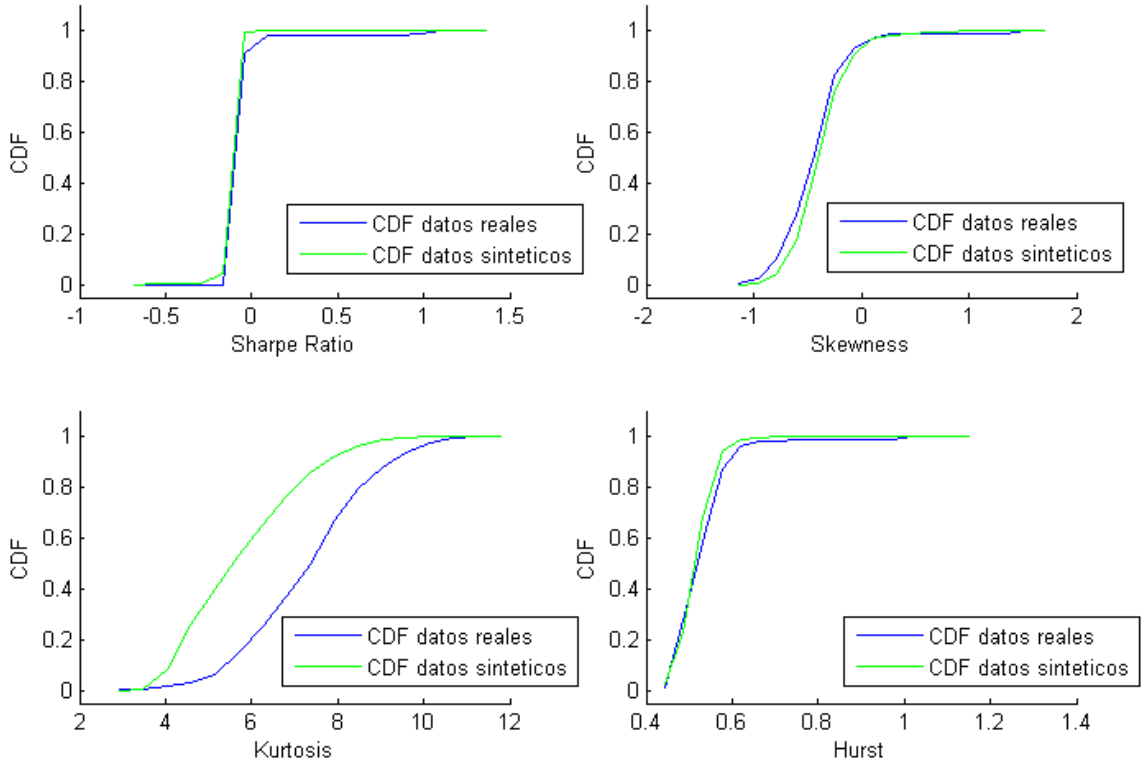

Figura 5.5: CDF de las cuatro características calculadas a partir de los datos reales y sintéticos, siendo los sintéticos generados con la configuración no óptima en predicción.

En el Sharpe Ratio, se observa una gran pendiente entre -0.2 y 0.2 aproximadamente, donde se concentran la mayoría de los valores, tanto sintéticos como reales, siendo la diferencia entre ambas curvas muy pequeña.

En la imagen del Skewness, el rango donde se concentran la mayoría de valores para ambas curvas se encuentra en torno a -0.9 y 0.2 aproximadamente, siendo también la diferencia entre ellas pequeña y observándose una pendiente algo más suave que la del Sharpe Ratio anterior. Estos valores de Skewness son los característicos de las series financieras, con funciones de distribución inclinadas hacia la izquierda de la media (en este caso con un valor igual a cero).

En la siguiente característica, la Kurtosis, se aprecia la mayor diferencia de las cuatro características entre los datos sintéticos y los reales. La mayoría de los valores sintéticos de la Kurtosis se concentran entre 4 y 9 aproximadamente (la curva sintética de la CDF empieza a crecer antes que la curva real y también alcanza antes el valor de 1), mientras que este rango en los datos reales se

encuentra entre 5 y 11, observándose una pendiente más suave en éstos. Los valores sintéticos generados por tanto tienen unos valores de Kurtosis algo menores que los reales, teniendo en definitiva una menor probabilidad de ocurrencia de outliers en las series sintéticas que se generan.

Por último, en el exponente de Hurst se puede ver la gran similitud entre ambas curvas, concentrándose la mayoría de los valores entre 0.45 y 0.65 aproximadamente, lo que refleja el comportamiento cercano a la aleatoriedad (H=0.5) de las series financieras reales, y como se puede apreciar también, de las series sintéticas generadas.

A continuación se muestran las CDF de los datos reales y los sintéticos generados ahora con los valores óptimos resultantes en predicción:

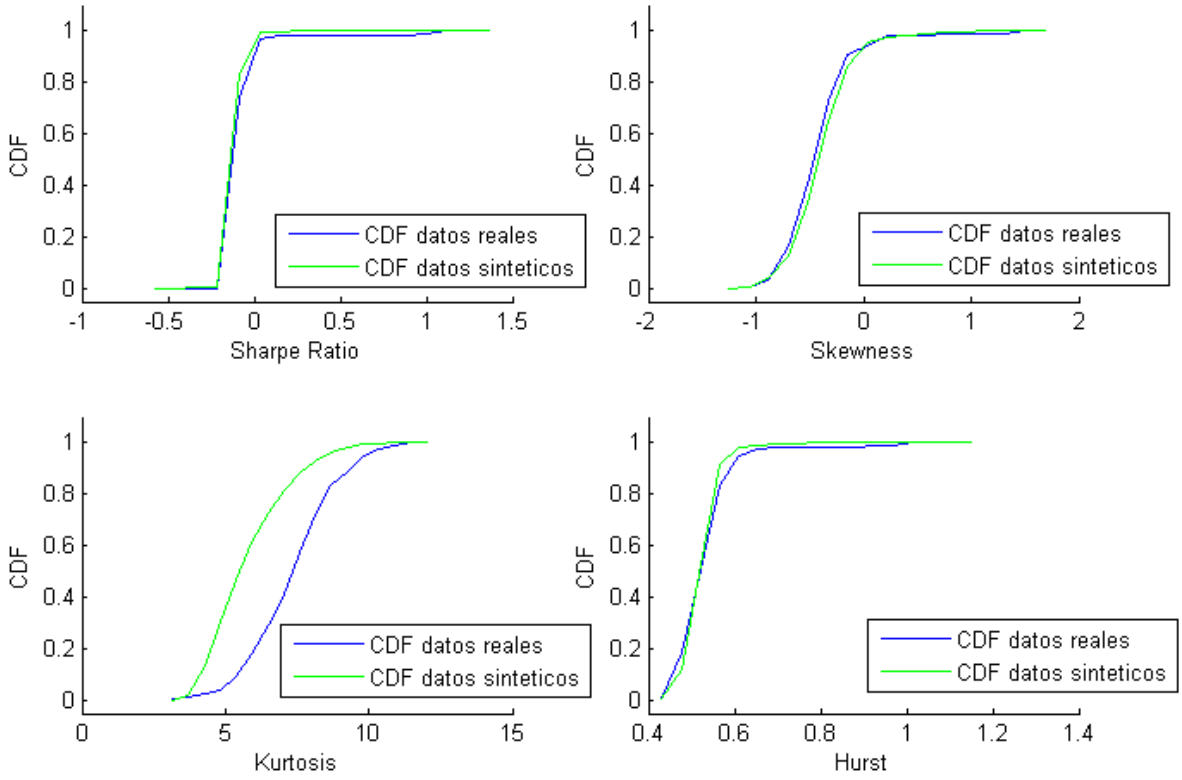

Figura 5.6: CDF de las cuatro características calculadas a partir de los datos reales y sintéticos, siendo los sintéticos generados con la configuración óptima resultante en predicción.

Con esta configuración, el Sharpe Ratio presenta un gran parecido con respecto a la configuración anterior, donde se observa también una pendiente abrupta entre -0.2 y 0.2 aproximadamente y una escasa diferencia entre la curva de los datos sintéticos y la de los reales. Sin embargo, observando los valores que se muestran en la tabla de las diferencias entre las CDF mostrada más abajo, se ha obtenido una diferencia menor con esta nueva configuración.

En el Skewness, se puede observar una pendiente menos abrupta donde la mayoría de los valores para ambas curvas se encuentran entre -0.9 y 0.2 aproximadamente, siendo también la diferencia entre curvas pequeña. De nuevo los datos sintéticos presentan unos valores de Skewness iguales a los de las series financieras reales, pero siendo la diferencia entre CDF menor con esta nueva configuración.

De nuevo, la Kurtosis es la característica que presenta la mayor diferencia entre los datos sintéticos y reales de las cuatro. También, la curva de los datos sintéticos empieza a crecer y a alcanzar el valor máximo de 1 un poco antes que la curva de los datos reales, presentando una pendiente algo

mayor que la de la curva real. Los datos de Kurtosis correspondientes a las series sintéticas por tanto presentan unos valores más pequeños que los de las series reales. En este caso, el dato de la diferencia calculada entre CDF que se puede observar en la tabla arroja un mejor resultado con la configuración no óptima resultante en predicción.

De igual modo que con la configuración anterior, en el exponente de Hurst se observa la similitud entre ambas curvas, concentrándose también la mayoría de los valores entre 0.45 y 0.65 aproximadamente, lo que refleja el comportamiento cercano a la aleatoriedad (H=0.5) de las series sintéticas generadas. A pesar del parecido que guarda con la anterior configuración, la diferencia entre ambas CDF es ahora menor, mejorándose por tanto con la configuración óptima en predicción.

En definitiva, en las gráficas de las figuras 5.5 y 5.6 anteriores se pueden ver como las curvas de los datos sintéticos presentan en general una gran verosimilitud con las reales, hecho este que resalta el buen funcionamiento del generador. Sin embargo esto no sucede con la Kurtosis, donde se observa la mayor diferencia entre los datos reales y los sintéticos. El generador aquí no es capaz de generar series con una Kurtosis tan similar a la de las reales.

En la siguiente tabla se muestran los valores de las diferencias calculadas entre las CDF a las que se ha hecho referencia anteriormente. Para calcularlas, en cada punto del eje común de abscisas definido se mide la diferencia entre ambas curvas, resultando finalmente los valores mostrados en la tabla de la suma de las diferencias de todos los puntos:

	Diferencia entre las CDF de los datos reales y sintéticos	
	Datos sintéticos generados con los valores óptimos	Datos sintéticos generados con los valores no óptimos
Sharpe Ratio	0.2628	0.2760
Skewness	0.3167	0.3969
Kurtosis	2.8480	2.7806
Exponente de Hurst	0.3059	0.3538

Los resultados anteriores reflejan una menor diferencia entre los datos reales y los datos sintéticos generados con los valores óptimos en predicción en todas las características excepto en la Kurtosis. Es decir, los resultados óptimos en predicción ayudan a mejorar la generación de series. El potencial de estos resultados empíricos es realmente grande, ya que la optimización de los resultados de predicción es un resultado más simple y directo, mientras que conocer los resultados óptimos para la generación de series es una tarea mucho más tediosa y subjetiva.

Esta última característica presenta una menor diferencia en el caso de los datos sintéticos generados con los valores no óptimos. Sin embargo, comparando los valores de las diferencias de esta característica para ambas configuraciones con el resto de características, se aprecia un valor más elevado, como ya se podía observar anteriormente en la CDF, debiéndose a la naturaleza de las series sintéticas generadas con nuestro generador, las cuales presentan unos valores de Kurtosis menores que los de las series reales.

6. Conclusiones y trabajo futuro

6.1. Conclusiones

Las conclusiones principales de este PFC se podrían resumir en:

- Las series temporales financieras presentan unas características que las diferencian del resto de series temporales.

- El método de predicción multivariable, con la necesaria optimización de sus parámetros configurables, ha sido capaz de conseguir unos mejores resultados en predicción que el método univariable.

- Las series sintéticas generadas con la configuración de los valores óptimos en predicción, han conseguido obtener una gran verosimilitud con respecto a las características que presentan las series reales, y además, han llegado a obtener en varias simulaciones, mejores resultados que con la configuración de los valores no óptimos usados anteriormente en el generador.

- La obtención de los parámetros óptimos es un proceso más simple y directo, mientras que conocer los resultados óptimos para la generación de series es una tarea mucho más tediosa y subjetiva. El punto anterior va a permitir progresar esta línea de investigación: la capacidad para medir la bondad de la generación de series sintéticas.

6.2. Trabajo futuro

Entre las futuras mejoras que se podrían añadir a este trabajo se pueden encontrar:

- La introducción de mejoras en el algoritmo multivariable de tal manera que se consiguieran mejorar aún más los resultados obtenidos en predicción con respecto al método univariable.

- Mejoras en el generador de series sintéticas, de tal manera que las series generadas, entre otras, tuvieran unos valores de Kurtosis más próximos a los que se observan en las series financieras reales.

- Introducción de nuevas mejoras y funcionalidades en la GUI implementada, como podrían ser la representación simultánea de varias series elegidas por el usuario, o la visualización de otras características.

Glosario de acrónimos

- PFC: Proyecto Fin de Carrera
- PCA: Principal Component Analysis
- GUI: Graphic User Interface
- CDF: Cumulative Distribution Function

Bibliografía

[1]. Ruey S.Tsay. *"Analysis of Financial Time Series"*. Wiley series in probability and statistics. Chapter 1, pp. 1-6, 2005.

[2]. Eric Zivot. "*Introduction to Computational Finance and Financial Econometrics"*. Chapters: 1 – 4, Coursera, University of Washington, 2011. https://class.coursera.org/compfinance-006/lecture

[3]. Armando Sánchez Vargas, Orlando Reyes Martínez. "*Regularidades probabilísticas de las series financieras y la familia de modelos Garch"*. Ciencia Ergo Sum, vol.13, número 002, pp. 149-156, Toluca, México, 2006.

[4]. Álvaro Diéguez Sánchez-Largo. "*Análisis Multifactor de Series Temporales Financieras Mediante Descomposición en Subespacios*". Trabajo Fin de Máster, EPS-UAM, 2013. http://www.eps.uam.es/nueva_web/intranet/ga/tfdm/trabajos/Alvaro_Dieguez_Sanchez-Largo.pdf

[5]. Antonio Pulido, Ana López, Jorge Rodríguez. *"Modelos econométricos uniecuacionales: la estacionariedad en la práctica"*. Curso de simulación en Economía y Gestión de Empresas, UAM, 2004.

[6]. Andrew Ng. *"Machine Learning"*. Chapter 14: Dimensionality Reduction, Coursera, Stanford, 2012. https://class.coursera.org/ml-003/lecture

[7]. Pau Micó. "*Nuevos desarrollos y aplicaciones basados en métodos estocásticos para el agrupamiento no supervisado de latidos en señales electrocardiográficas".* Capítulo 3: Reducción de características, análisis de componentes principales, pp. 64-67, Tesis Doctoral, UPV, 2005. http://riunet.upv.es/bitstream/handle/10251/1856/tesisUPV2326.pdf

[8]. Comunidad forex. *"Acercándose al análisis fundamental: los principios básicos".* 2014, www.efxto.com

[9]. David S. Matteson and David Ruppert. "*Time-Series Models of Dynamic Volatility and Correlation"*. IEEE Signal Process, 28(5), pp.72–74, 2011.

[10]. Comunidad forex. "*Las medias móviles*". 2009, www.efxto.com

Anexos

A. Resultados error tipo 2

Los valores óptimos para los parámetros empleados en predicción son los siguientes:

	Valores óptimos para los parámetros
	Error tipo 2
α (días)	50
T (días)	20
Información (%)	99
Ponderación	Alta
Normalización	No
Días ventana PCA (días)	500
Valor resultante	**0.987**

Variando los valores de los parámetros α y T:

$\frac{error\ tipo\ 2\ M}{error\ tipo\ 2\ U}$	T = 1	T = 5	T = 10	T = 20
α = 1	0.9994	0.9902	0.9951	0.9950
α = 5	0.9961	0.9907	0.9921	0.9928
α = 10	1.0002	0.9956	0.9975	0.9948
α = 20	1.0023	1.0033	1.0005	0.9983
α = 50	1.0014	0.9996	0.9947	**0.9870**
α = 80	0.9997	1.0049	1.0072	1.0083

Variando el valor del parámetro *Información*:

$\frac{error\ tipo\ 2\ M}{error\ tipo\ 2\ U}$	*Información* (98%)	*Información* (99%)	*Información* (100%)
Valor resultante	0.9901	**0.9870**	0.9960

Variando el valor del parámetro *Ponderación*:

$\frac{error\ tipo\ 2\ M}{error\ tipo\ 2\ U}$	*Ponderación* (Baja)	*Ponderación* (Media)	*Ponderación* (Alta)
Valor resultante	0.9909	0.9871	**0.9870**

Variando el valor del parámetro *Normalización*:

$\frac{error\ tipo\ 2\ M}{error\ tipo\ 2\ U}$	*Normalización* (Si)	*Normalización* (No)
Valor resultante	1.0965	**0.9870**

Variando el valor del parámetro *Días ventana PCA*:

$\frac{error\ tipo\ 2\ M}{error\ tipo\ 2\ U}$	*Días ventana PCA* (125 días)	*Días ventana PCA* (250 días)	*Días ventana PCA* (500 días)
Valor resultante	1.0156	1.0034	**0.9870**

B. Resultados error tipo 3

Los valores óptimos para los parámetros empleados en predicción son los siguientes:

	Valores óptimos para los parámetros
	Error tipo 3
α (días)	50
T (días)	20
Información (%)	99
Ponderación	Media
Normalización	No
Días ventana PCA (días)	500
Valor resultante	**0.9810**

Variando los valores de los parámetros α y *T*:

$\frac{error\ tipo\ 3\ M}{error\ tipo\ 3\ U}$	*T* = 1	*T* = 5	*T* = 10	*T* = 20
α = 1	1.0020	0.9961	0.9992	1.0008
α = 5	0.9985	0.9946	0.9957	0.9999
α = 10	1.0022	0.9950	0.9994	0.9987
α = 20	1.0041	1.0078	1.0063	0.9969
α = 50	1.0028	0.9979	0.9935	**0.9810**
α = 80	0.9993	1.0027	1.0068	1.0053

Variando el valor del parámetro *Información*:

$\frac{error\ tipo\ 3\ M}{error\ tipo\ 3\ U}$	*Información* (98%)	*Información* (99%)	*Información* (100%)
Valor resultante	0.9911	**0.9810**	0.9888

Variando el valor del parámetro *Ponderación*:

$\frac{error\ tipo\ 3\ M}{error\ tipo\ 3\ U}$	*Ponderación* (Baja)	*Ponderación* (Media)	*Ponderación* (Alta)
Valor resultante	0.9851	**0.9810**	0.9814

Variando el valor del parámetro *Normalización*:

$\frac{error\ tipo\ 3\ M}{error\ tipo\ 3\ U}$	*Normalización* (Si)	*Normalización* (No)
Valor resultante	1.1176	**0.9810**

Variando el valor del parámetro *Días ventana PCA*:

$\frac{error\ tipo\ 3\ M}{error\ tipo\ 3\ U}$	*Días ventana PCA* (125 días)	*Días ventana PCA* (250 días)	*Días ventana PCA* (500 días)
Valor resultante	1.0200	1.0075	**0.9810**

C. Resultados error tipo 4

Los valores óptimos para los parámetros empleados en predicción son los siguientes:

	Valores óptimos para los parámetros
	Error tipo 4
α (días)	50
T (días)	20
Información (%)	99
Ponderación	Media
Normalización	No
Días ventana PCA (días)	500
Valor resultante	**0.9779**

Variando los valores de los parámetros α y *T*:

$\frac{error\ tipo\ 4\ M}{error\ tipo\ 4\ U}$	*T* = 1	*T* = 5	*T* = 10	*T* = 20
α = 1	1.0019	0.9940	0.9981	1.0006
α = 5	0.9976	0.9922	0.9940	0.9994
α = 10	1.0027	0.9943	0.9996	0.9976
α = 20	1.0054	1.0096	1.0073	0.9945
α = 50	1.0046	0.9988	0.9934	**0.9779**

α = 80	1.0006	1.0048	1.0095	1.0065

Variando el valor del parámetro *Información*:

$\frac{error\ tipo\ 4\ M}{error\ tipo\ 4\ U}$	*Información* (98%)	*Información* (99%)	*Información* (100%)
Valor resultante	0.9909	**0.9779**	0.9880

Variando el valor del parámetro *Ponderación*:

$\frac{error\ tipo\ 4\ M}{error\ tipo\ 4\ U}$	*Ponderación* (Baja)	*Ponderación* (Media)	*Ponderación* (Alta)
Valor resultante	0.9828	**0.9779**	0.9785

Variando el valor del parámetro *Normalización*:

$\frac{error\ tipo\ 4\ M}{error\ tipo\ 4\ U}$	*Normalización* (Si)	*Normalización* (No)
Valor resultante	1.1285	**0.9779**

Variando el valor del parámetro *Días ventana PCA*:

$\frac{error\ tipo\ 4\ M}{error\ tipo\ 4\ U}$	*Días ventana PCA* (125 días)	*Días ventana PCA* (250 días)	*Días ventana PCA* (500 días)
Valor resultante	1.0268	1.0111	**0.9779**

D. PRESUPUESTO

1) Ejecución Material

- Compra de ordenador personal (Software incluido)..... ..... 1900 €
- Alquiler de impresora láser durante 6 meses .....200 €
- Material de oficina .....200 €
- Total de ejecución material..... 2.300 €

2) Gastos generales

- 16 % sobre Ejecución Material ..... 368 €

3) Beneficio Industrial

- 6 % sobre Ejecución Material ..... 138 €

4) Honorarios Proyecto

- 1500 horas a 15 € / hora ..... 22500 €

5) Material fungible

- Gastos de impresión ..... 150 €
- Encuadernación ..... 200 €

6) Subtotal del presupuesto

- Subtotal Presupuesto ..... 25150 €

7) I.V.A. aplicable

- 21% Subtotal Presupuesto..... 5281.5 €

8) Total presupuesto

- Total Presupuesto ..... 30431,5 €

Madrid, Junio de 2014

El Ingeniero Jefe de Proyecto

Fdo.: Alfredo Serrano Quejido
Ingeniero de Telecomunicación

E. PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de Análisis y Caracterización de Series Temporales Financieras. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato,

incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contrata" y anteriormente llamado "Presupuesto de Ejecución Material" que hoy designa otro concepto.

Condiciones particulares

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el

momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.