

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES



TRABAJO DE FIN DE GRADO

**TÉCNICAS DE MACHINE LEARNING EN ECONOMETRÍA:
LA EFICIENCIA DEL MERCADO FINANCIERO ESPAÑOL.**

Tutor:
Manuel Ángel Domínguez Toribio

Noé Muñoz Alonso

Doble Grado en Economía - Matemáticas y Estadística

Curso académico 2021-22

Convocatoria de febrero

Resumen

El objetivo de este trabajo es estudiar la eficiencia del mercado financiero español en 2021. En primer lugar se hará una breve revisión de la literatura existente y se proporcionará una fundamentación teórica de la Hipótesis del Mercado Eficiente (Fama, 1970). En segundo lugar, se introducirán las técnicas econométricas necesarias para el análisis del mercado financiero y para el contraste de la Hipótesis del Mercado Eficiente, poniendo el foco en aquellas que hacen uso de los modernos métodos de Aprendizaje Automático o *Machine Learning*. Finalmente, se llevará a cabo un contraste de la Hipótesis del Mercado Eficiente para la serie diaria del IBEX-35, comparando el desempeño de los métodos tradicionales, basados en la metodología Box-Jenkins, con los métodos modernos, basados en la predicción a través de Redes Neuronales Artificiales.

Abstract

This project aims to study the efficiency of the Spanish financial market in 2021. First and foremost, we will revise the existent literature and we will provide a theoretical framework for establishing the Efficient Market Hypothesis (Fama, 1970). Moreover, we will introduce the econometric techniques that will need to be used to analyse the financial market and to test the Efficient Market Hypothesis, focusing on those that make use of the modern *Machine Learning* methods. Finally, we will test the Efficient Market Hypothesis using the daily IBEX-35 series, comparing the performance of traditional methods, based on the Box-Jenkins methodology against the modern methods, based on prediction through Artificial Neural Networks.

Índice general

Resumen	2
Introducción	9
1 Fundamentación: Hipótesis del Mercado Eficiente	11
1.1 La Hipótesis del Mercado Eficiente	11
1.1.1 Hechos Estilizados de las Series Temporales Financieras	12
1.1.2 La Hipótesis de la Diferencia de Martingalas	13
2 Modelización: Técnicas Econométricas	15
2.1 Métodos Lineales para Series Temporales	15
2.1.1 Los procesos ARIMA	15
2.1.2 La metodología Box-Jenkins	16
2.2 Métodos No-Lineales para Series Temporales	17
2.2.1 Redes Neuronales Artificiales	18
2.3 Contraste de la EMH	20
2.3.1 El contraste LWG	20
2.3.2 Métricas de comparación	21
3 Aplicación Empírica: Eficiencia del IBEX-35	25
3.1 Análisis Exploratorio y Perfilado de los Datos	25
3.2 Métodos Lineales	26
3.2.1 Fase de identificación	27
3.2.2 Fase de validación	28
3.3 Métodos No-Lineales	30
3.3.1 Contraste LWG	30
3.3.2 Contraste direccional	31
3.4 Conclusiones	32
A Anexo	33

Índice de figuras

2.1	Estructura básica de una red neuronal <i>feed-forward</i> con una capa oculta.	18
3.1	Exploración de la serie diaria del IBEX-35. Elaboración propia con <i>R</i>	25
3.2	Estructura de autocorrelación de P_t . Elaboración propia con <i>R</i>	27
3.3	Estructura de autocorrelación de $\nabla \log P_t$. Elaboración propia con <i>R</i>	27
3.4	Contrastes gráficos asociados al residuo. Elaboración propia con <i>R</i>	29
3.5	Estructura y predicciones de la ANN. Elaboración propia con <i>R</i>	31

Índice de cuadros

3.1	Análisis de estadísticos descriptivos principales de la muestra.	26
3.2	División de la muestra para entrenamiento y validación.	26
3.3	Contraste aumentado de Dickey-Fuller.	26
3.4	Estimación de constante y efecto lunes.	28
3.5	Contraste de Ljung-Box para la serie del rendimiento.	28
3.6	Contrastes de Normalidad para el rendimiento.	29
3.7	Contraste LM de Engle para efectos ARCH.	29
3.8	Selección de la arquitectura de la ANN.	30
3.9	Resultados del contraste de Lee-White-Granger.	31
3.10	Tabla de contingencia para el contraste de Hylleberg-Dahl.	31
3.11	Resultados del contraste de Hylleberg-Dahl (Pearson).	32

Introducción

En el mundo de las Finanzas, tanto académicas como empíricas, una preocupación constante ha sido descubrir la estructura subyacente a los procesos que generan los precios de los activos financieros. En efecto, el interés académico pasa por comprender de manera profunda el funcionamiento de los mercados financieros y el interés aplicado pasa por obtener beneficio utilizando las predicciones que se pueden obtener si se conoce el proceso generador de los precios de activos.

Una de las grandes protagonistas en este contexto es la Hipótesis del Mercado Eficiente (Fama, 1970) o EMH por sus siglas en inglés, que postula que los mercados financieros reflejan toda la información pasada en los precios actuales. Una de las implicaciones de la EMH es que los rendimientos de los activos financieros vienen generados por un proceso de Martingala en diferencias y, por lo tanto, no se deberían poder predecir para obtener rendimientos superiores a los del mercado.

La herramienta fundamental para contrastar la EMH ha sido la Econometría, en particular las técnicas del análisis de series temporales, que utilizan como caballo de batalla la metodología Box-Jenkins y los procesos ARIMA. La modelización de series temporales basada en esta metodología se basa en el Teorema de Wold, que permite expresar todo proceso puramente estocástico estacionario como un proceso de ruido blanco al que se le ha aplicado un cierto filtro lineal. Por lo tanto, asume una estructura lineal para el nivel de las series temporales que modeliza. Los estudios realizados utilizando estas herramientas, en general, no consiguen rechazar la EMH (al menos, la forma más débil de la misma). No obstante, esto no implica necesariamente que los mercados sean eficientes, sino simplemente que podrían existir estructuras no lineales para el nivel de la serie temporal que el modelo no es capaz de capturar. Esto, sumado a las muy particulares características estadísticas de las series temporales financieras, hace que la predicción del nivel de tales series se convierta en una tarea especialmente ardua.

De especial relevancia en este punto son los métodos de predicción basados en la Inteligencia Artificial (IA) y el Machine Learning (ML), que utilizan la información proporcionada por los datos de forma novedosa, dando lugar a modelizaciones muy flexibles, que son capaces de capturar complejas estructuras no lineales y de mejorar frecuentemente las predicciones obtenidas por medio de métodos clásicos.

A lo largo del TFG, realizaremos un estudio de la eficiencia del mercado financiero español, a través del análisis econométrico de la serie del índice de precios del IBEX-35. Comprobaremos que los modelos ARIMA no son capaces de capturar la estructura de la serie de los rendimientos de dicho índice y propondremos algunos modelos algo más complejos, basados en Machine Learning, para realizar predicciones de la dirección en la que se mueve el mercado, con el objetivo de, posteriormente, diseñar alguna estrategia de inversión en el caso en el que se obtuviese alguna información relevante. Desde luego, si las predicciones obtenidas no resultasen de utilidad, se obtendría evidencia a favor de la EMH en el mercado financiero español actual.

Fundamentación: Hipótesis del Mercado Eficiente

La predicción financiera engloba a las técnicas de predicción de los precios o rendimientos de una gran clase de activos financieros, como pueden ser bonos, opciones, tipos de interés, tipos de cambio... Muchos agentes económicos, como los fondos de inversión, que mueven grandes sumas de dinero, están interesados en la predicción de los precios financieros de por sí. No obstante, la importancia académica de la predicción financiera está íntimamente relacionada con el rol de los mercados financieros dentro de la modelización macroeconómica. El desarrollo de nuevos instrumentos e instituciones financieras contribuye al crecimiento y a la estabilidad de la economía, precisamente por el papel que tienen los mercados financieros en la economía real, permitiendo trasladar recursos entre distintos instantes de tiempo y proporcionar completitud a los mercados. Dado que dicha modelización macroeconómica permite ofrecer un tratamiento de los mercados financieros esencialmente idéntico al de cualquier otro mercado, las técnicas y modelos estadísticos que se utilizan para predecir el crecimiento de macro-magnitudes como el PIB, el consumo, la inflación o los precios de activos reales, también pueden ser utilizadas para la predicción del rendimiento de los activos financieros.

No obstante, los mercados financieros presentan algunas diferencias que resultan ser de gran relevancia a la hora de realizar un análisis profundo de los mismos, tanto desde el punto de vista teórico como desde el punto de vista aplicado. Dedicamos la primera sección a explorar ambos aspectos.

1.1 La Hipótesis del Mercado Eficiente

En primer lugar, en el centro de una de las grandes discusiones de la Economía Financiera se encuentra la Hipótesis del Mercado Eficiente o EMH (Fama, 1970). La EMH es una de las ideas más influyentes en las Finanzas modernas, y se basa en la noción de que los precios de los mercados financieros reflejan completamente toda la información disponible y no existen estrategias de inversión que produzcan rendimientos ajustados por riesgo que sean positivos. Si bien dicha hipótesis es ampliamente discutida, es la consecuencia del funcionamiento de los mercados competitivos: la respuesta de los inversores a la llegada de nueva información es inmediata y, dado que los agentes se suponen racionales, las ventajas derivadas del uso de la nueva información a la hora de invertir desaparecen rápidamente.

En cualquier caso, Fama no es demasiado concreto con su definición de eficiencia. De manera informal, se diría que un mercado es eficiente en sentido informacional (no en el sentido habitual, que se utiliza en la economía neoclásica) cuando los precios reflejan completamente toda la información disponible, que vendría reflejada por un cierto conjunto de información Ω_t . No obstante, no se ha especificado la forma concreta de los conjuntos de información Ω_t . La distinción según las variables que se incluyen en ellos conduce a las tres formas de la EMH:

1. EMH, forma débil. En este caso, se supone que Ω_t está engendrado por el histórico de la serie,

es decir, sólo por los valores pasados del proceso. En la forma débil de eficiencia, los precios futuros no se pueden predecir analizando los propios precios pasados. Por lo tanto, no se pueden obtener rentabilidades extraordinarias en el largo plazo usando estrategias basadas en el análisis histórico de precios. De este modo, el Análisis Técnico (AT) no puede ayudar a obtener rendimientos extraordinarios. En definitiva, esta versión implica que no existen relaciones de dependencia entre las series de precios, luego no se pueden determinar patrones de precios.

2. EMH, forma semifuerte. En este caso, se supone que Ω_t está engendrado por el histórico de la serie y por toda la información pública disponible en el período t . Esta forma de eficiencia implica que los precios de las acciones reflejan toda la información pública y de una forma prácticamente instantánea. Bajo este supuesto, ningún participante del mercado puede obtener rentabilidades extraordinarias evaluando las nuevas informaciones publicadas. La implicación principal es que tanto el Análisis Técnico como el Análisis Fundamental (AF) son inútiles para obtener rendimientos extraordinarios.

3. EMH, forma fuerte. En este caso, se supone que Ω_t está engendrado por el histórico de la serie y por toda la información tanto pública como privada disponible en el período t , por lo que los precios de los activos se corresponden con sus precios teóricos, la información es accesible para todo el mundo y se incorpora inmediatamente a la hora de realizar el ajuste de precios. Por último, se añade la racionalidad de los agentes y de las expectativas a la hora de realizar estrategias de inversión: el precio P_t incorpora toda la información relevante hasta el instante t y, por lo tanto, el incremento en el precio $P_t - P_{t-1}$ es debido solamente a las noticias no previstas por los agentes entre $t - 1$ y t . De este modo, los precios son correctos y no existen oportunidades de arbitraje.

1.1.1 Modelización: Hechos Estilizados de las Series Temporales Financieras

Para realizar una modelización estadística que permita contrastar la EMH mediante técnicas de análisis de series temporales financieras, a lo largo del trabajo nos centraremos en la forma débil de ésta. Para realizar dicha modelización, es importante tener en cuenta algunas regularidades empíricas que suelen observarse al recoger datos de las mismas y también tener en cuenta que afectarán a las conclusiones que extraigamos del análisis realizado. Según Ruiz (1994), fundamentalmente son:

1. Nula o escasa estructura regular dinámica en la media. En muchas ocasiones, al intentar suavizar la heterocedasticidad de la serie y eliminar la tendencia, se obtiene una serie de ruido blanco. En los casos en los que aparece estructura autorregresiva o de media móvil, suele ser AR(1) o MA(1) con un parámetro pequeño.

2. Distribuciones asimétricas y apuntadas (leptocúrticas). Las series temporales financieras no suelen seguir una distribución Normal y, en muchos casos, una vez ajustada una ecuación para la media, el proceso residual también se aleja de esta distribución.

3. Agrupamiento de la volatilidad. Mandelbrot (1963) y Fama (1965) señalaron que en las series temporales financieras para niveles mayores de la media también se obtienen niveles mayores de la varianza. Este hecho suele venir reflejado por funciones de autocorrelación significativas para el cuadrado del residuo, que indica precisamente la persistencia de esta volatilidad para los distintos períodos.

4. Efecto apalancamiento. Cuando las series financieras analizadas son rendimientos de activos financieros, se suele observar una respuesta asimétrica de la volatilidad al nivel de los rendimientos. Habitualmente la volatilidad se incrementa en mayor medida cuando el rendimiento es negativo que cuando es positivo; ver, por ejemplo, Black (1976) y Christie (1982). No obstante, algunos de los modelos más populares (ARCH) no son capaces de modelizar adecuadamente este hecho.

5. Efectos estacionales. Cuando se analizan series financieras diarias, normalmente, se observan algunos efectos estacionales en el viernes y lunes debidos a que no existen transacciones comerciales

durante el fin de semana, como señalan Baillie y Bollerslev (1989). En los casos en los que estos efectos no sean significativos, en la práctica se suele optar por ignorarlos y trabajar con las series financieras ignorando los fines de semana.

1.1.2 Formalización: la Hipótesis de la Diferencia de Martingalas

La elaboración de una versión formal de la hipótesis de Fama que sea posible incorporar a los modelos macroeconómicos pasa necesariamente por suponer que los precios y los rendimientos de los activos financieros vienen generados por un cierto proceso estocástico. Un concepto esencial en este contexto es el de *martingala*, un tipo de proceso estocástico cuyas propiedades reflejan la eficiencia del mercado en el sentido de que la probabilidad de que el precio de un cierto activo suba o baje es la misma.

Definición 1.1. Se dice que un proceso estocástico $\{Z_t\}$ es una **martingala** si, para el conjunto de información Ω_t que contiene toda la información relevante relativa al histórico de dicho proceso, se verifica

$$E[Z_{t+1}|\Omega_t] = Z_t \quad (1.1)$$

Al diferenciar un proceso que forma una martingala, se observa que se tiene

$$E[Z_{t+1} - Z_t|\Omega_t] = E[Z_{t+1}|\Omega_t] - Z_t = 0$$

Esta propiedad, sencilla de ver pero de gran importancia, da lugar a la siguiente definición:

Definición 1.2. Se dice que un proceso estocástico $\{Z_t\}$ es una **diferencia de martingalas** si, para el conjunto de información Ω_t que contiene toda la información relativa al histórico de dicho proceso, se verifica

$$E[Z_{t+1}|\Omega_t] = 0 \quad (1.2)$$

Teniendo en cuenta los hechos estilizados expuestos anteriormente más el hecho de que, en términos informales, la EMH postula que el precio de los activos es correcto y que no existen oportunidades de arbitraje basadas en la utilización de información pasada, se podría modelizar el rendimiento de un activo como

$$r_t = \mu_t + e_t \quad (1.3)$$

donde e_t tiene media cero. Aplicando esperanzas condicionadas, es $E[r_t|\Omega_{t-1}] = \mu_t$ y por lo tanto e_t representa el rendimiento del activo que se debe a cambios en el precio no anticipados por los agentes. En la literatura, se suele suponer $\mu_t = \mu$, constante, tomando el modelo la forma

$$r_t = \mu + e_t \quad (1.4)$$

Esta modelización es consistente con los hechos estilizados de las series financieras comentados anteriormente. Al modelizar el rendimiento, se pueden hacer distintos supuestos acerca del término e_t , cuyas propiedades se transmitirán directamente a la serie del rendimiento:

1. Hipótesis de Ruido Blanco (WNH): si se cumple $e_t \sim RB(0, \sigma_e^2)$, entonces $Cov(r_t, r_s) = 0$ para $t \neq s$. La hipótesis de ruido blanco es consistente con los hechos estilizados de las series financieras y, además de aparecer constantemente en las series de rendimientos financieros, es una implicación directa de la EMH, dado que sólo impone que no se pueden predecir los rendimientos de los activos utilizando predictores lineales. No obstante, no impone condiciones sobre momentos de orden superior y, desde

luego, no aporta información acerca de una posible estructura no-lineal en el proceso de rendimientos, por lo que en general es poco informativa.

2. Hipótesis del Paseo Aleatorio (RWH): En este caso, se supone que el proceso de error verifica $e_t \sim \text{IID}(0, \sigma_e^2)$. Bajo estos supuestos distribucionales tan fuertes, dado que se puede aproximar el rendimiento como $r_t \simeq \log P_t - \log P_{t-1}$, se tiene que el logaritmo del proceso de precios verifica

$$\log P_t = \mu + \log P_{t-1} + e_t \quad (1.5)$$

y entonces el proceso $\{\log P_t\}$ es un paseo aleatorio. Así, el proceso de precios sigue lo que se denomina paseo aleatorio geométrico, muy conocido y tratado en la literatura financiera. Lo que implica esta hipótesis es que el futuro es independiente tanto del presente como del pasado, por lo que la EMH es válida en sentido estricto y el futuro no puede ser, de ningún modo, predicho a partir de la información disponible en el presente.

Aunque esta estructura puede ser plausible en algunos casos (Fan y Yao, 2017), se reconoce demasiado fuerte para ser cierta en general. Es necesario, por lo tanto, proponer una modelización intermedia, entre la WNH y la RWH, que consiga capturar de manera general los hechos estilizados.

3. Hipótesis de la Diferencia de Martingalas (MDH): En este caso, e_t verifica

$$E[e_t | \Omega_{t-1}] = 0 \quad (1.6)$$

Una modelización frecuente de dicho término de error en este caso es

$$e_t = \sigma_t \eta_t \quad (1.7)$$

donde η_t tiene media nula y varianza unitaria y σ_t es el proceso de volatilidad, predecible, conocido en $t - 1$ y que verifica

$$E[\sigma_t | \Omega_{t-1}] = \sigma_t$$

Dado que los procesos ARCH y GARCH son casos particulares de (1.7), lo que muestran las ecuaciones (1.6) y (1.7) es que la existencia de heterocedasticidad condicional en los rendimientos no contradice la MDH.

La forma más apropiada de modelizar el proceso de rendimientos de un activo financiero es, desde luego, la MDH. Mantiene la propiedad de independencia lineal del ruido blanco, pero permite que existan relaciones para momentos de orden superior, de modo que por ejemplo $\{r_t^2\}$ o $\{|r_t|\}$ pueden presentar patrones de dependencia. La MDH captura la propiedad de que no es posible predecir e_{t+1} utilizando sus valores pasados y, así, la predicción óptima (de mínimos cuadrados) de r_{t+1} es

$$E[r_{t+1} | \Omega_t] = E[\mu + e_{t+1} | \Omega_t] = \mu + E[e_{t+1} | \Omega_t] = \mu \quad (1.8)$$

En resumen, la MDH asegura que los rendimientos de los activos no pueden predecirse, esto es, no tienen ningún tipo de estructura en media, ni lineal ni no-lineal (sin realizar ninguna afirmación acerca de la volatilidad o de momentos de orden superior). Dado que una cuestión fundamental en este contexto es saber si la EMH es consistente con los datos empíricos, sería interesante comprobar si dicha hipótesis es falsa o cierta. Un modo de contrastar la EMH es tratar de predecir los rendimientos, dado que si fuese posible se encontraría evidencia en contra de la hipótesis. Dedicamos el siguiente capítulo a introducir y desarrollar las técnicas econométricas que utilizaremos para tal fin.

2

Modelización: Técnicas Econométricas

La Econometría y el Machine Learning tienen una meta común: la construcción de un modelo que explique y/o prediga el valor de una variable de interés utilizando variables explicativas como inputs. Sin embargo, la Econometría trata de construir modelos estadísticos capaces de explicar los fenómenos económicos y, en la medida de lo posible, proporcionar predicciones para los mismos; mientras que el Machine Learning construye algoritmos que son capaces de aprender de sus propios errores y cuyo objetivo principal es la predicción, y no la descripción de los fenómenos.

Una de las principales ventajas de los métodos de Machine Learning es que permiten detectar aspectos que los modelos lineales ignoran, por su gran capacidad para aprender de los datos. Uno de estos aspectos es la existencia de patrones no-lineales en los datos, que podrían potencialmente aprovecharse para posibles aplicaciones de la predicción de los valores de la serie (en el caso de las Finanzas, el diseño de estrategias de inversión para obtener beneficio gracias al arbitraje). Dedicamos este capítulo a exponer someramente las técnicas clásicas de predicción con datos de series temporales y el papel que juegan las técnicas de Machine Learning no sólo en el ámbito de la predicción sino, como se ha comentado, también en la validación de los modelos econométricos.

2.1 Métodos Lineales para Series Temporales

2.1.1 Los procesos ARIMA

La Econometría de series temporales ha utilizado, tradicionalmente, modelos lineales basados en la metodología Box-Jenkins. El objetivo de estos modelos no es tanto realizar inferencia causal, como ocurre con los modelos de regresión en general, sino más bien predecir, suponiendo que existe la suficiente persistencia en el proceso como para que el pasado pueda, al menos a corto plazo, explicar el futuro. La justificación teórica para la utilización de este tipo de modelos viene dada por el siguiente resultado (Wold, 1938), que puede encontrarse en (Peña, 2010).

Teorema 2.1 (Wold). *Todo proceso estocástico débilmente estacionario Z_t , de media finita μ , que no contenga componentes deterministas, puede escribirse como una función lineal de variables aleatorias incorreladas, e_t , como*

$$Z_t = \mu + \sum_{i=0}^{\infty} \psi_i e_{t-i} \quad (2.1)$$

siendo $\psi_0 = 1$ y $\sum_{i=0}^{\infty} \psi_i^2 < \infty$. Las variables e_t verifican $E[e_t] = 0$, $V(e_t) = \sigma_e^2$ y $E[e_t e_{t-k}] = 0$, $k > 1$.

La expresión 2.1 se conoce como representación $MA(\infty)$ del proceso Z_t . Para evitar la estimación de un número infinito de parámetros, se utiliza la aproximación $ARMA(p, q)$

$$\phi_p(B)Z_t = \theta_q(B)e_t \quad (2.2)$$

De especial interés en el análisis econométrico es la modelización de series no estacionarias, que presentan variaciones en el nivel (como pueden ser las series con tendencia, como la del PIB) o variaciones en la volatilidad (como las series de precios de activos financieros). Permitiendo raíces unitarias en el polinomio autorregresivo para modelizar series no estacionarias, se llega a la aproximación $ARIMA(p, d, q)$

$$\phi_p(B)(1 - B)^d Z_t = \theta_q(B)e_t \quad (2.3)$$

2.1.2 La metodología Box-Jenkins

Box y Jenkins (1970) propusieron una metodología de modelización de series temporales para estimar la estructura de autocorrelación de dichas series y realizar inferencia acerca de los parámetros que identifican al proceso $ARIMA$ subyacente a los datos. Los pasos principales en el proceso de ajuste de una serie temporal según la metodología Box-Jenkins son:

- 1. Hacer estacionaria la serie.** Para ello, se suelen utilizar transformaciones que inducen estacionariedad en varianza (Transformaciones Box-Cox) y en media (diferencias regulares y/o estacionales).
- 2. Fase de identificación.** En esta fase, se deben identificar los órdenes del proceso $ARMA$ subyacente bajo la serie. Esta identificación se puede realizar comparando el correlograma muestral con los correlogramas teóricos de los procesos $ARMA$.
- 3. Fase de validación.** En esta fase, se debe contrastar que el proceso residual resultante tras el ajuste del modelo es la imagen empírica de un proceso de ruido blanco, lo que significa que el modelo está correctamente especificado.
- 4. Fase de predicción.** Una vez validado el modelo, elaborar predicciones para la media a horizonte h , dadas por $\hat{Z}_{t+h} = E_t[Z_{t+h}]$. Siempre que sea posible, es conveniente realizar un *backtesting*, para calcular métricas de precisión sobre el conjunto de validación y comparar la calidad de las predicciones con las de otros modelos con el mismo objetivo.

De especial interés en este trabajo son los dos últimos apartados de la metodología Box-Jenkins. En primer lugar, el contraste de la hipótesis de ruido blanco para el ajuste de la serie temporal se corresponde precisamente con la hipótesis de ruido blanco para la innovación en el rendimiento que se comentaba en el capítulo 1. La forma más popular de contrastar de manera global la autocorrelación en los primeros k retardos de la función de autocorrelación total viene dada por el contraste Q de Ljung-Box. El estadístico del contraste viene dado por

$$Q(k) = n(n+2) \sum_{i=1}^k \frac{\hat{\rho}_i^2}{n-i} \quad (2.4)$$

Bajo la hipótesis nula de ausencia de autocorrelación, se tiene la convergencia $Q(k) \xrightarrow{d} \chi^2(k-s)$, siendo s el número de parámetros estimados en el modelo. Una limitación de este contraste y que precisamente proporciona relevancia a la introducción de métodos de Machine Learning es que la hipótesis nula supone que existe independencia en los datos, por lo que puede no llegar a ser demasiado informativo.

Además, dado que se basa en las funciones de autocorrelación, que son lineales por definición, sólo es capaz de detectar estructura lineal en el residuo y no es capaz de capturar ni patrones no-lineales en media ni patrones en momentos de orden superior.

En relación a la estructura en los momentos de orden superior del proceso residual, nótese que la MDH no descarta la presencia de efectos ARCH. Para contrastar la presencia de estos efectos, se suele utilizar el contraste LM de Engle (1982). Para llevar a cabo el contraste, se realiza la siguiente regresión auxiliar

$$\hat{e}_t^2 = \alpha_0 + \sum_{i=1}^{\tilde{p}} \alpha_i \hat{e}_{t-i}^2 + v_t \quad (2.5)$$

donde v_t es ruido blanco y \tilde{p} es el orden del efecto ARCH. La hipótesis nula es $H_0 : \alpha_1 = \dots = \alpha_{\tilde{p}} = 0$, es decir, la ausencia de efectos ARCH hasta el retardo \tilde{p} . Se tiene que

$$LM = (n - \tilde{p})R^2 \xrightarrow{d} \chi^2(\tilde{p}) \quad (2.6)$$

por lo que valores grandes del estadístico de contraste proporcionarán evidencia a favor de la existencia de efectos ARCH en el residuo.

La justificación de este contraste es que la volatilidad resulta

$$\sigma_t^2 = V_{t-1}[e_t] = E_{t-1}[e_t^2] - E_{t-1}[e_t]^2 = E_{t-1}[e_t^2]$$

y por lo tanto se puede aproximar el error al cuadrado por el residuo al cuadrado, que, bajo efectos ARCH, debería presentar una fuerte estructura de autocorrelación. Si existiesen efectos ARCH, entonces se podría seguir la metodología Box-Jenkins y ajustar un modelo $AR(\tilde{p})$ para la volatilidad de la serie o un modelo $ARMA(\tilde{p}, \tilde{q})$, en cuyo caso se diría que se está ajustando un modelo $GARCH(\tilde{p}, \tilde{q})$ (Bollerslev, 1986). Una vez estimadas las ecuaciones para la media y para la volatilidad por separado, se deberían reestimar de manera conjunta, dado que suele mejorar la calidad de las predicciones para la media al tener en cuenta la estructura de la volatilidad.

2.2 Métodos No-Lineales para Series Temporales

Para entender cómo se pueden detectar los patrones no-lineales en los datos, primero es necesario entender qué son. Sea un proceso estocástico $\{Z_t\}$, que particionamos como $Z_t = (y_t; X_t')'$, de tal modo que y_t es un escalar, la variable respuesta, X_t es un vector de dimensión $k \times 1$ que contiene los retardos de la variable respuesta y \tilde{X}_t es el vector X_t incluyendo posiblemente una constante. Ponderando los retardos y la constante por un cierto parámetro $\theta \in \mathbb{R}^k$, Bierens (1984) proporciona la siguiente definición:

Definición 2.2. *Se dice que el proceso $\{y_t\}$ es lineal en media condicionada respecto a X_t cuando*

$$P[E[y_t|X_t] = X_t'\theta^*] = 1 \quad (2.7)$$

para algún $\theta^* \in \mathbb{R}^k$.

Nótese que la definición 2.2 es consistente con el concepto informal de linealidad manejado en la sección anterior para los modelos ARIMA y para algunos procesos de interés como el proceso de ruido blanco. Según esta definición, si existiesen no-linealidades en el proceso, entonces sería $P(E[y_t|X_t] = X_t'\theta^*) < 1$. Cuando esto ocurre, se dice que los modelos lineales sufren de *no-linealidad ignorada*, por lo que no

parecería lógico pensar que la descripción que ofrecen de la realidad es buena y desde luego conduce a dudar sobre la calidad de las predicciones que proporcionan.

Otro aspecto importante a notar de la definición es que no descarta la presencia de efectos ARCH. Dado que no se hace ningún supuesto sobre los momentos de orden superior del proceso, los modelos que combinan una ecuación ARIMA para la media con una ecuación GARCH para la volatilidad siguen siendo lineales en media condicionada.

2.2.1 Redes Neuronales Artificiales

Presentamos en esta sección un método de Machine Learning, las Redes Neuronales Artificiales (ANN). Las ANN son ampliamente utilizadas en el ámbito académico para realizar predicciones de distintas variables macroeconómicas y, especialmente, en el ámbito empresarial. En esencia, son una clase de métodos no-paramétricos que persiguen aproximar una variable de interés a través de una función de regresión

$$y_t = f(X_t) + e_t$$

siendo e_t un término de error sobre el que, en principio, no se establece ningún supuesto. Existen muchos tipos de redes neuronales, y en las últimas décadas el volumen de investigación en este campo ha aumentado enormemente. Explicar con detalle los distintos tipos de ANN queda fuera del alcance de este trabajo, por lo que pasamos a exponer el tipo más popular de ANN y que utilizaremos para nuestros propósitos: la red neuronal *feed-forward* con una capa oculta.

En un problema de predicción a partir de observaciones de otras variables, se supone que el cerebro observa como funciona el fenómeno de interés tratando de extraer de esa experiencia algunas regularidades que le permitan predecir el comportamiento de la variable objetivo, es decir, recoge información tanto de la variable a predecir (output) como de las variables que se utilizarán para la predicción inputs. Aparta el output por el momento y construye estructuras con los inputs que compara con el output. A partir del resultado de dicha comparación, modifica la estructura creada para los inputs para mejorar el resultado.

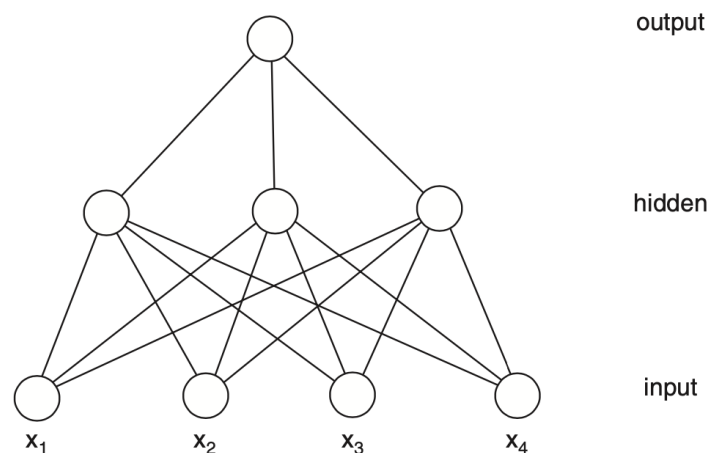


Figura 2.1: Estructura básica de una red neuronal *feed-forward* con una capa oculta.

En ANN, las estructuras se obtienen transformando los valores de los inputs con diferentes funciones paramétricas que llaman funciones de activación y que arrojan un resultado. Cada uno de estos procesos de transformación tiene lugar dentro de lo que se conoce como una neurona. A mayor número de

neuronas, más complejo es la estructura que se está creando a fin de obtener un resultado final suficientemente parecido. Cuando no lo es, se vuelve al inicio cambiando los parámetros de las funciones de activación forma adecuada para que el resultado final sea el deseado.

La figura 2.1 corresponde a una red neuronal con ocho neuronas representadas mediante nodos. Las líneas que unen los nodos representan relaciones entre las neuronas. En este caso, al ser una red *feed-forward*, el esquema corresponde a un grafo dirigido u orientado en el que las neuronas se relacionan de abajo a arriba. Con esto quiere decirse que el resultado de la función de activación de una neurona se toma como input en todas las neuronas situadas por encima que están relacionadas con ella. Es habitual que la relación entre neuronas se defina en forma de capas como en el gráfico. Así, la capa inferior se llama capa de entrada o *input layer*, la superior capa de salida o *output layer* y las intermedias capas ocultas o *hidden layers* (de las que sólo hay una en el gráfico).

De manera formal, los elementos básicos que conforman una neurona son:

- a) Un espacio de inputs o variables explicativas, $\mathcal{X} \subset \mathbb{R}^k$ con $k \geq 1$.
- b) Un espacio de outputs o predicciones, $\mathcal{Y} \subset \mathbb{R}$.
- c) Un vector de parámetros de ponderación de la información, $\gamma \in \mathbb{R}^k$.
- d) Una función de activación $\psi : \mathbb{R} \rightarrow \mathbb{R}$. Para asegurar el funcionamiento correcto de la neurona, se suele exigir que la función sea monótona, diferenciable y acotada (en muchos casos, que verifique $\lim_{x \rightarrow \infty} \psi(x) = 1$ y $\lim_{x \rightarrow -\infty} \psi(x) = 0$).

Una neurona, por tanto, es una función $f_\gamma : \mathcal{X} \rightarrow \mathcal{Y}$ definida por

$$y = f_\gamma(x) = \psi(x'\gamma) \quad (2.8)$$

En la capa oculta normalmente hay más de una neurona (en la figura 2.1 hay tres), por lo que la neurona de la capa de salida debe repetir el proceso. Para problemas de regresión también es lo usual utilizar la identidad como función de activación en la capa de salida, por lo que el output de la red neuronal que tiene q neuronas en la capa oculta será

$$y = \sum_{j=1}^q \beta_j f_{\gamma_j}(x) = \sum_{j=1}^q \beta_j \psi(x'\gamma_j) \quad (2.9)$$

y los coeficientes β_j son las ponderaciones que otorga la neurona de salida a cada uno de los datos que recibe de las neuronas de la capa oculta.

En primer lugar debe notarse que, si bien el uso de funciones de activación recuerda a las funciones de enlace de los modelos lineales generalizados (algunos son de amplio uso en Econometría, como los modelos *Logit* y *Probit*), a la hora de definir y establecer la estructura de la ANN no se está imponiendo ningún modelo probabilístico subyacente. Esto, es no establecemos suposiciones acerca de la distribución de $y|X$ y por lo tanto no podemos hablar propiamente de un modelo de regresión, sino sólo de un método de regresión.

Otro aspecto importante que se debe tener en cuenta es que hasta ahora no se ha hablado de la capacidad de las ANN para ajustar una función de regresión arbitraria. El siguiente teorema (Hornik-Stinchcombe-White, 1989) asegura que, con el suficiente número de neuronas, una ANN puede aproximar arbitrariamente bien una amplia clase de funciones de regresión.

Teorema 2.3 (Aproximación Universal). *Sea ψ una función de activación, sea k la dimensión del espacio de inputs y sea μ una medida de probabilidad sobre (\mathbb{R}^k, β^k) . Con la métrica*

$$\rho(f, g) = \inf\{\varepsilon > 0 : \mu\{x : |f(x) - g(x)| < \varepsilon\} > \varepsilon\} \quad (2.10)$$

el espacio de neuronas es denso en el espacio de funciones medibles-Borel sobre \mathbb{R}^k .

Lo que indica el teorema 2.3 es que, para cualquier función de activación, cualquier dimensión del espacio de inputs y cualquier ley de probabilidad subyacente a los datos, una red neuronal *feed-forward* con una capa oculta puede aproximar cualquier función de regresión que sea medible, siempre que se pueda disponer del número suficiente de neuronas. En efecto, como el espacio de neuronas es denso, se podrá aproximar la función de regresión como

$$f(x) \simeq \sum_{j=1}^q \beta_j \psi(x' \gamma_j) \quad (2.11)$$

que es precisamente el output de la red neuronal.

2.3 Contraste de la EMH

En general, los contrastes asociados al residuo de la modelización Box-Jenkins en principio permitirían comprobar si existe evidencia en contra de la EMH. El problema es que, dado que se basan en la modelización lineal de las series temporales, pueden no capturar las no-linealidades existentes en dichas series y en realidad sólo permiten contrastar la hipótesis de ruido blanco para el término de error.

2.3.1 El contraste LWG

Si los métodos lineales no permiten extraer ningún tipo de patrón de la serie del rendimiento, sería de especial relevancia comprobar si los métodos no lineales consiguen identificar algún tipo de no-linealidad que los primeros no estén consiguiendo capturar. En este punto es de especial relevancia la introducción de métodos basados en Machine Learning.

Dado que el teorema 2.3 justifica la utilización de redes neuronales para aproximar una clase amplia de funciones de regresión a partir de los datos, Lee et al. (1993) proponen un contraste de no-linealidad basado en dichas redes neuronales que podemos aplicar a la serie del rendimiento. Siguiendo la notación de la definición 2.2, el contraste se basa en realizar una regresión de dicha serie como

$$y_t \simeq \tilde{X}_t' \theta + \sum_{j=1}^q \beta_j \psi(\tilde{X}_t' \gamma_j) \quad (2.12)$$

Si no existiesen no linealidades en los datos, entonces sería cierta

$$H_0 : P \left[E[y_t | X_t] = \tilde{X}_t' \theta^* \right] = 1 \quad (2.13)$$

para algún parámetro $\theta^* \in \mathbb{R}^k$. En este caso, deberían ser para 2.12 los coeficientes $\beta_1^* = \dots = \beta_q^* = 0$. Para determinadas elecciones de q y de γ_j sería, por tanto, la hipótesis a contrastar. Es claro que, si es posible extraer estructura de $e_t^* = y_t - \tilde{X}_t' \theta^*$, entonces el test rechazaría H_0 .

El contraste de significación conjunta de los coeficientes β_j^* se puede llevar a cabo, en principio, con un estadístico tipo F (Wald). De manera adicional Lee et al. (1993) proponen la siguiente hipótesis nula, equivalente, para poder expresarlo en términos de un contraste LM:

$$H_0' : E[\Psi_t e_t^*] = 0 \quad (2.14)$$

donde $\Psi_t = (\psi(\tilde{X}_t' \Gamma_1), \dots, \psi(\tilde{X}_t' \Gamma_q))'$ y $\Gamma = (\Gamma_1, \dots, \Gamma_q)$ se elige a priori, independientemente de $\{X_t\}$ para $q \in \mathbb{N}$. Según Bierens (1987), se puede elegir Γ de manera aleatoria sin perder generalidad.

La correlación entre e_t^* y Ψ_t significaría que existe evidencia a favor de la existencia de no-linealidades en los datos, y que por lo tanto se rechazaría H'_0 . El contraste se basa, por lo tanto, en estimar $E[\Psi_t e_t^*]$. A la hora de construir un estadístico para el contraste, se reemplazaría e_t^* por el residuo $\hat{e}_t = y_t - \tilde{X}_t' \hat{\theta}$, siendo $\hat{\theta}$ los parámetros estimados de una regresión estándar de mínimos cuadrados. Se tendría así el estadístico de contraste

$$M_n = \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n \Psi_t \hat{e}_t \right)' \widehat{W}_n^{-1} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n \Psi_t \hat{e}_t \right) \quad (2.15)$$

donde \widehat{W}_n verifica

$$\text{plim}_n \widehat{W}_n = V \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n \Psi_t e_t^* \right)$$

La aplicación del método Delta asegura que, bajo H'_0 , se tiene la convergencia

$$M_n \xrightarrow{d} \chi^2(q)$$

Lee et al. (1993) señalan dos problemas fundamentales a la hora de llevar a cabo el contraste:

- 1) Existe una alta multicolinealidad entre Ψ_t y X_t y también entre los propios Ψ_t . Para remediar este problema, se reduce la dimensionalidad utilizando $q^* < q$ componentes principales, menos que el número de neuronas en la capa oculta.
- 2) Es computacionalmente costoso calcular \widehat{W}_n . Para evitar dicho cálculo, se puede realizar una regresión de \hat{e}_t sobre \hat{X}_t y Ψ_t^* y calcular

$$\text{LM} = nR^2 \xrightarrow{d} \chi^2(q^*) \quad (2.16)$$

siendo R^2 el coeficiente de determinación estándar de la regresión auxiliar. Bajo heterocedasticidad condicional, se utiliza una versión robusta del estadístico LM (Wooldridge, 1990). Para ello:

- a) Se realiza una regresión auxiliar de las Ψ_t^* sobre X_t , obteniendo conjuntos de residuos $\{\tilde{e}_t^i\}_{i=1}^k$.
- b) Se obtienen los nuevos residuos $\mathbf{e}_t^i = \tilde{e}_t^i \cdot \hat{e}_t$
- c) Se realiza una regresión sin constante de un vector de unos, $\mathbf{1}$, sobre los residuos \mathbf{e}_t^i , y se calcula la suma de cuadrados de los residuos, SCR_1 . Se verifica

$$\text{LM} = n - SCR_1 \xrightarrow{d} \chi^2(q^*) \quad (2.17)$$

2.3.2 Métricas de comparación

A la hora de ajustar una ANN, es necesario realizar una estimación de las ponderaciones γ_{ij} para cada uno de los inputs en cada una de las neuronas. El método general de ajuste de una ANN se denomina *Retropropagación* y, si bien queda fuera del alcance de este trabajo y no es realmente relevante para nuestros objetivos, sí lo es determinar la función de pérdida, L , que se utiliza para realizar dicho ajuste. Una vez se determina la arquitectura de la ANN (es decir, el número de neuronas de la capa oculta y

la función de activación) trataríamos de determinar las ponderaciones que minimicen lo que se conoce como *riesgo empírico*.

$$\hat{\mathcal{R}}_n(x) = \frac{1}{n} \sum_{t=1}^n L(y_t, x_t) \quad (2.18)$$

La función $\hat{\mathcal{R}}_n(x)$ trata, de algún modo, de determinar la distancia promedio entre los verdaderos valores y las predicciones de la ANN. En general, como función de pérdida se suele utilizar el error cuadrático medio, de modo que volvemos a la función de riesgo empírico utilizada en el contexto de la regresión lineal, la suma de los cuadrados de los residuos.

$$\hat{\mathcal{R}}_n(x) = \frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2 \quad (2.19)$$

con $\hat{y}_t = \hat{y}_t(x)$, dado que las predicciones como es evidente se realizan a partir de los inputs.

De este modo, las ponderaciones se eligen de manera que el output de la ANN se ajuste lo mejor posible a los datos disponibles (conjunto de entrenamiento), en el sentido de los mínimos cuadrados. Para comprobar la calidad de las predicciones, se debe reservar una parte de los datos (conjunto de validación). Al calcular el valor de la función de riesgo empírico sobre este conjunto, debería esperarse un valor similar al obtenido sobre el conjunto de entrenamiento. En caso contrario, la ANN habría captado perfectamente la estructura de los datos del conjunto de entrenamiento pero no sería capaz de reconocer la estructura de nuevos datos, fenómeno conocido como *sobreajuste*.

Para elegir entre distintos modelos, por lo tanto, se evalúa el valor de la métrica calculado sobre el conjunto de validación (proceso de *backtesting*). Es importante recalcar que la comparación de la métrica de error ¹ permite comparar ANN con distinta arquitectura, por lo que se pueden utilizar para determinar la arquitectura óptima de la ANN, tal y como se hará en la siguiente sección.

El contraste direccional de Hylleberg-Dahl

Tratando con series financieras, además de la evaluación de la capacidad predictiva en términos cuantitativos, también es relevante la precisión de la predicción en términos cualitativos. Es decir, para el diseño de una buena estrategia de inversión sería conveniente conocer si la serie de los precios del activo en cuestión se encuentra en un período de subida o de bajada. Para evaluar, complementariamente a la calidad de la predicción en términos exactos, la capacidad de detección de subidas y bajadas, se utilizará una métrica conocida como Medida de Precisión Direccional o MDA.

$$\text{MDA} = \frac{100}{n} \sum_{t=1}^n \mathbf{1}_{\{\text{sgn}(y_t - y_{t-1}) = \text{sgn}(\hat{y}_t - y_{t-1})\}} \quad (2.20)$$

Esta métrica proporciona una estimación en porcentaje de la probabilidad de que el método utilizado para realizar las predicciones detecte de manera correcta la dirección que toma la serie temporal. Para contrastar estadísticamente si la precisión cualitativa de las predicciones es significativa, se plantea un contraste de medida direccional, basado en la corrección del contraste clásico de Henriksson-Merton (1981) propuesta por Dahl y Hylleberg (1999). El contraste se basa en una tabla de contingencia bidimensional en la que se recogen el número de aciertos y errores del modelo a la hora de predecir las subidas y bajadas de la serie del logaritmo del precio en el conjunto de validación.

¹También se podrían utilizar métricas distintas, como el criterio de Akaike (AIC), el criterio bayesiano de Schwarz (BIC) o el error absoluto medio (MAE), entre otras.

Real	Predicción		
	Sube	Baja	
Sube	m_{11}	m_{12}	m_{10}
Baja	m_{21}	m_{22}	m_{20}
	m_{01}	m_{02}	m

En esta tabla, m es el número de elementos en el conjunto de validación. El valor m_{11} es el número de aciertos al predecir movimientos ascendentes, m_{21} el número de errores al predecir movimientos ascendentes, y análogamente para m_{12} y m_{22} . La hipótesis nula es que las predicciones de la red neuronal tienen la misma distribución que las que se obtendrían si el logaritmo del precio fuese un paseo aleatorio. Como es evidente, cuanto mayores sean los valores de m_{11} y m_{22} mayor será la precisión en la predicción cualitativa de la red neuronal. El estadístico de contraste es

$$X^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(m_{ij} - m_{i0}m_{0j}/m)^2}{m_{i0}m_{0j}/m} \quad (2.21)$$

Bajo condiciones muy generales, se tiene la convergencia (Hylleberg y Dahl, 1999)

$$X^2 \xrightarrow{d} \chi^2(1)$$

Un valor alto de X^2 indicará que la red neuronal produce predicciones mejores que las de un paseo aleatorio, por lo que, de obtenerlo con los datos de la muestra, se rechazaría la EMH en su forma débil.

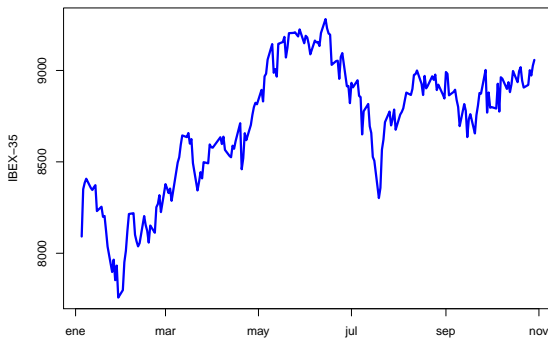
3

Aplicación Empírica: Eficiencia del IBEX-35

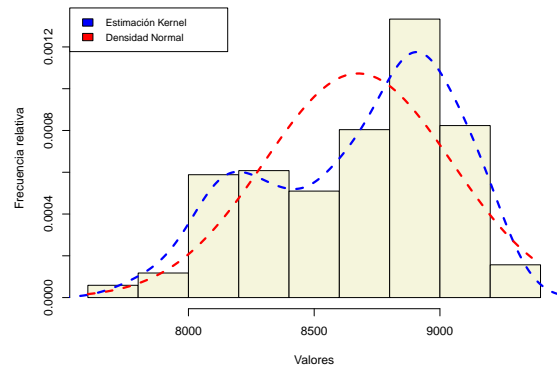
En esta última sección utilizamos las herramientas econométricas que hemos desarrollado en la segunda sección para contrastar la eficiencia en sentido débil del mercado financiero español actual, representado a través de la serie diaria del IBEX-35, disponible para el acceso público en Yahoo! Finance, considerando el histórico desde diciembre de 2020. Consideramos este período para el ajuste y la predicción porque se cuenta con suficiente número de datos como para ajustar de manera satisfactoria un modelo de series temporales, sea un modelo lineal o un modelo no-lineal (Walczak, 2001) capturando correctamente la estructura actual del mercado..

3.1 Análisis Exploratorio y Perfilado de los Datos

Como paso previo al ajuste de modelos predictivos, es importante realizar una exploración de los datos, para detectar posibles tendencias, desviaciones de las suposiciones habituales de Normalidad y, en definitiva, obtener una visión general del conjunto de datos con el que estamos tratando. Con tal fin, presentamos en la figura 3.1 una exploración gráfica de los datos, como primera aproximación.



(a) Serie diaria del IBEX-35



(b) Histograma de la serie diaria del IBEX-35.

Figura 3.1: Exploración de la serie diaria del IBEX-35. Elaboración propia con R .

De la exploración gráfica de la serie, conviene destacar que no parece que la distribución de la serie de precios sea Normal. En la figura 3.1b se ha realizado una estimación de la densidad basada en la muestra vía Kernel (Gaussiano), y se ha comparado con la distribución teórica de una Normal, y el ajuste no resulta satisfactorio. Como se puede ver en la figura 3.1a, una de las causas de la asimetría es que la

serie parece exhibir una tendencia ascendente a lo largo del año. Para obtener un reflejo cuantitativo de los aspectos que se han detectado en la exploración gráfica y alguna información adicional, presentamos en la siguiente tabla los estadísticos principales de la muestra recogida.

Estadísticos principales				
Media	Mediana	Mínimo	Máximo	Desviación Típica
8671.74	8775.10	7757.50	9281.10	349.19
Asimetría	Curtosis	Percentil 5 %	Percentil 95 %	Rango Intercuartílico
-0.48	-0.86	8037.60	9183.90	611.30

Cuadro 3.1: Análisis de estadísticos descriptivos principales de la muestra.

Para el ajuste de los modelos, y en la línea de la literatura existente (Walczak, 2001), se ha realizado una división de la muestra en dos conjuntos de datos:

- 1. Conjunto de entrenamiento.** Se utiliza este conjunto para ajustar o entrenar los modelos. Corresponde al período comprendido entre el 08/12/2020 y el 15/10/2021.
- 2. Conjunto de validación.** Se utiliza este conjunto para validar el poder predictivo de los modelos. Corresponde al período comprendido entre el 18/10/2021 y el 24/11/2021.

Se ha dejado algo más del último mes para predecir, aproximadamente el 15 % de la muestra. Representamos en la siguiente tabla esta división del conjunto de datos, especificando el número de elementos asignado a cada conjunto.

División de la muestra		
	Entrenamiento	Validación
Inicio	08/12/2020	18/10/2021
Final	15/10/2021	24/11/2021
Total	220	28

Cuadro 3.2: División de la muestra para entrenamiento y validación.

3.2 Métodos Lineales

En primer lugar, según la figura 3.1a la serie de precios no parece estacionaria, dado que el nivel va cambiando con el tiempo y además la intensidad de las fluctuaciones también parece ir cambiando a medida que pasa el tiempo. La no-estacionariedad viene reflejada en la función de autocorrelación de la serie, que presenta un patrón de decrecimiento prácticamente lineal.

Para realizar un contraste formal, podemos llevar a cabo un contraste ADF (Augmented Dickey-Fuller) de raíces unitarias.

Contraste aumentado de Dickey-Fuller		
DF = -1,9078	Retardos = 6	p-valor = 0,6147
Hipótesis alternativa: la serie es estacionaria		

Cuadro 3.3: Contraste aumentado de Dickey-Fuller.

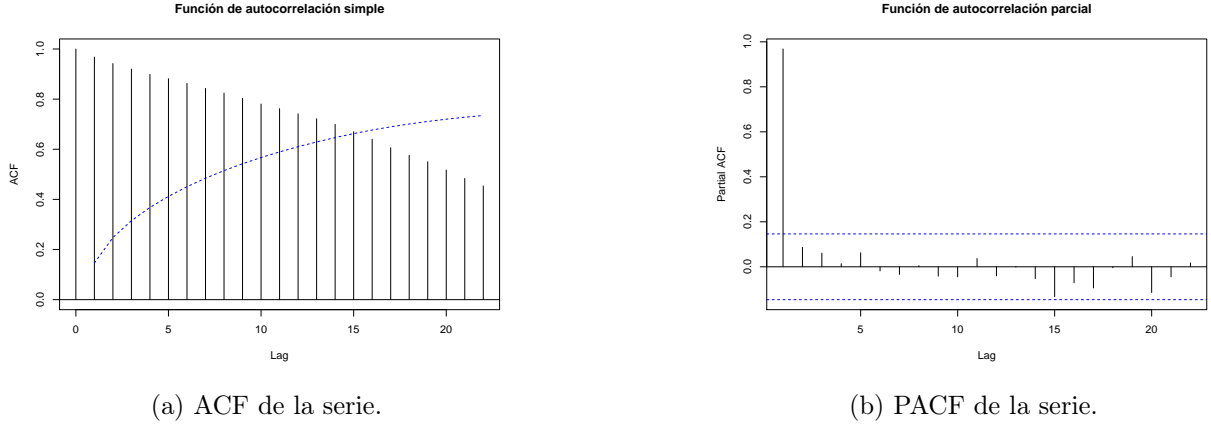


Figura 3.2: Estructura de autocorrelación de P_t . Elaboración propia con R .

El filtrado de la serie de precios según la metodología Box-Jenkins conduce a aplicar una diferencia regular al logaritmo de la serie de precios. Esto no sólo es conveniente para inducir estacionariedad (el logaritmo induce estacionariedad en varianza y la diferencia regular en media), sino que, dado que se tiene

$$r_t = \nabla \log P_t = \log P_t - \log P_{t-1} = \log \left(\frac{P_t}{P_{t-1}} \right) = \log \left(1 + \frac{P_t - P_{t-1}}{P_{t-1}} \right) \simeq \frac{P_t - P_{t-1}}{P_{t-1}} \quad (3.1)$$

por ser $\lim_{x \rightarrow 0} \frac{\log(1+x)}{x} = 1$, el resultado de aplicar logaritmo y la diferencia regular es precisamente la serie del rendimiento, que es la realmente interesante de predecir. Para esta serie, ya no se detecta ningún patrón de autocorrelación, como se puede comprobar a continuación.

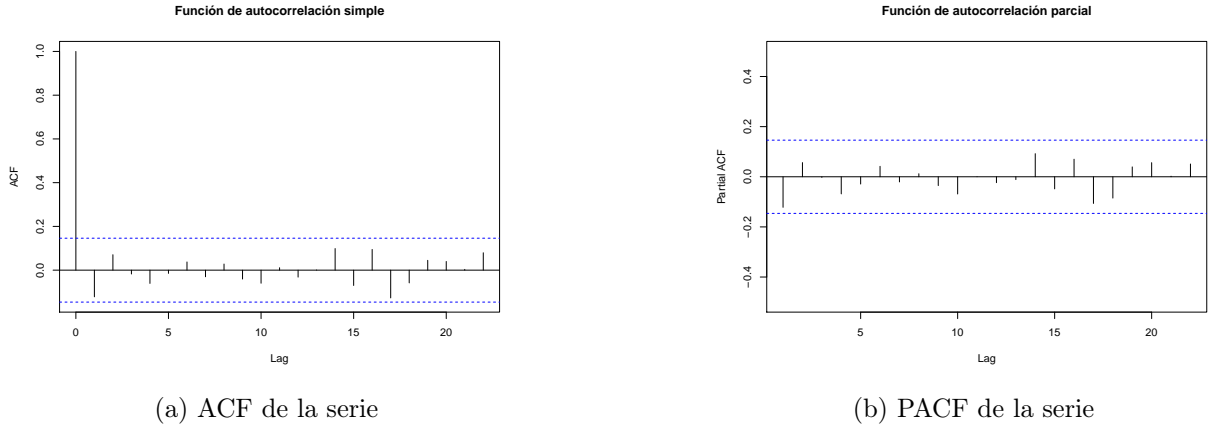


Figura 3.3: Estructura de autocorrelación de $\nabla \log P_t$. Elaboración propia con R .

3.2.1 Fase de identificación

Con el objetivo de descartar la existencia de efectos semanales, se define una variable binaria que tiene en cuenta el efecto entre los viernes y los lunes sobre el nivel de la serie, además de una constante. El modelo así estimado sería

$$r_t = \mu + \omega \nabla \text{Lunes}_t + e_t \quad (3.2)$$

y la estimación de dicho modelo se muestra en la siguiente tabla:

Regresión con errores ARIMA(0,1,0)			
Transformación Box-Cox: $\lambda = 0$ (Logaritmo)			
	Coefficiente	Desv. Típica	p -valor
Constante	0.0005	0.0007	0.425
Lunes	0.0007	0.0011	0.482

Cuadro 3.4: Estimación de constante y efecto lunes.

Como se observa, ni la constante ni la variable binaria resultan significativas al llevar a cabo la estimación. Por lo tanto, al aplicar un modelo $I(1)$ puro al logaritmo de la serie de precios, se obtiene un proceso residual trivial correspondiente a la serie del rendimiento. Como es obvio, omitiremos la fase de predicción, dado que, para el modelo que se ha identificado, la predicción a cualquier horizonte h es $\hat{r}_h = 0$.

3.2.2 Fase de validación

Una vez estimado el modelo, es el momento de realizar los contrastes asociados al residuo, dado que en este caso no se ha tenido que estimar ningún coeficiente. En primer lugar, la figura 3.3 muestra que el rendimiento no presenta autocorrelación serial. En efecto, el contraste de autocorrelación de Ljung-Box así lo confirma.

Contraste Q de Ljung-Box	
Retardo	p -valor
6	0,168
12	0,134
18	0,274
24	0,595
30	0,503

Cuadro 3.5: Contraste de Ljung-Box para la serie del rendimiento.

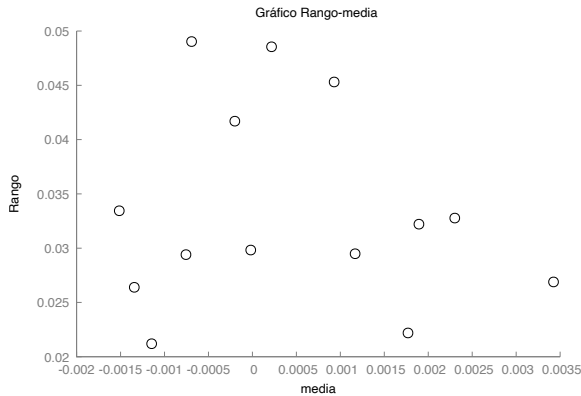
El contraste de Ljung-Box, que tiene como hipótesis nula la ausencia de autocorrelación serial, presenta unos p -valores muy altos. Por lo tanto, no permite descartar que la serie del rendimiento se corresponde con un proceso de ruido blanco.

Aunque la transformación Box-Cox de la serie, correspondiente al logaritmo, garantiza prácticamente la homocedasticidad marginal, podemos llevar a cabo un contraste rango-media para confirmar que no tenemos evidencia de que exista heterocedasticidad en el proceso residual. Tomando 14 submuestras de tamaño 13 del conjunto de entrenamiento, se obtiene el gráfico rango media de la figura 3.4a. Al ajustar la regresión por mínimos cuadrados del rango con respecto a la media

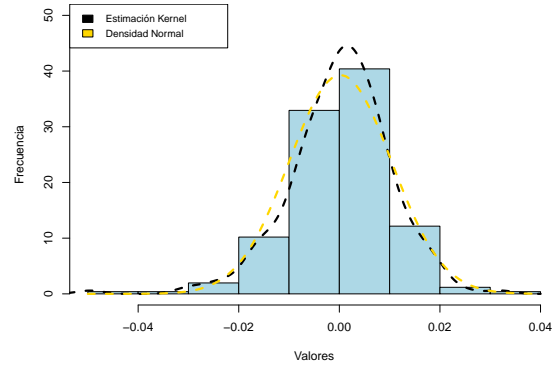
$$r = \beta_0 + \beta_1 m \quad (3.3)$$

se obtiene una pendiente de $-0,7503$ que no es significativa, dado que su p -valor para el contraste de hipótesis alternativa $\beta_1 = 0$ es 0,673. Por lo tanto, se puede asumir que el rendimiento es incondicionalmente homocedástico.

Otro aspecto interesante sería comprobar la Normalidad de la serie, dado que si no se consiguiese rechazar, entonces se podría estar presentando la EMH en su forma de paseo aleatorio, más fuerte que la forma de diferencia de martingalas.



(a) Gráfico rango-media de la serie r_t .



(b) Histograma de r_t .

Figura 3.4: Contrastes gráficos asociados al residuo. Elaboración propia con R .

Para contrastar la hipótesis de Normalidad de manera formal, llevamos a cabo algunos contrastes muy conocidos y muy presentes en la literatura econométrica: el contraste LM de Jarque-Bera, el contraste de Kolmogorov-Smirnov tanto en su versión habitual como con la corrección asintótica de Lilliefors y el contraste de Shapiro-Wilk.

Contraste	Valor	Distribución	p -valor
Jarque-Bera	5.141	$\chi^2(2)$	0.076
Kolmogorov-Smirnov	0.052	KS	0.585
Lilliefors	0.053	KS	0.151
Shapiro-Wilk	0.991	W	0.186

Cuadro 3.6: Contrastes de Normalidad para el rendimiento.

Los contrastes toman como hipótesis nula la normalidad del residuo y ninguno la rechaza, por lo que no se puede descartar que r_t tenga distribución Normal, con media $\mu = 0$ y desviación típica $\sigma = 0.0031$. Dado que en presencia de Normalidad la incorrelación implica independencia, descartamos la presencia de efectos ARCH. Por asegurar que no existen tales efectos, podemos llevar a cabo el test LM de Engle.

Contraste LM de Engle	
Retardo	p -valor
1	0,839
2	0,366
3	0,297
4	0,404
5	0,506

Cuadro 3.7: Contraste LM de Engle para efectos ARCH.

Los p -valores son muy grandes, y dado que la hipótesis nula es que no existen efectos ARCH hasta el retardo correspondiente, el contraste no detecta heterocedasticidad condicional en la serie del rendimiento. De este modo, los contrastes relativos al residuo no permiten rechazar la EMH para el IBEX-35 y no sólo eso sino que sugieren que la forma de EMH que se presenta en la serie del rendimiento del índice es la RWH.

3.3 Métodos No-Lineales

Elegir la correcta arquitectura para la ANN es crucial. Dado que no existe una metodología de identificación de la arquitectura óptima a partir de los datos, como en el caso de los métodos lineales y la metodología Box-Jenkins, se ha ajustado una gran cantidad de ANN, con distinto número de inputs y de neuronas. La función de activación utilizada ha sido la función logística

$$\psi(x) = \frac{1}{1 + e^{-x}} \quad (3.4)$$

En la siguiente tabla se recogen los modelos que proporcionan menores valores del RMSE *out-of-sample*, es decir, sobre el conjunto de validación.

Retardos	Neuronas	RMSE
2	3	0.0109
2	4	0.0108
3	3	0.0093
3	4	0.0098
4	3	0.0094
4	4	0.0102
5	3	0.0076
5	4	0.0104

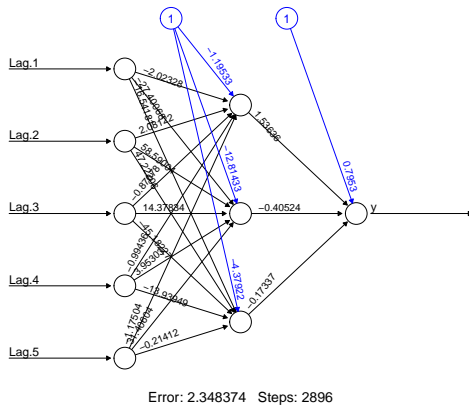
Cuadro 3.8: Selección de la arquitectura de la ANN.

Conviene destacar que a partir de 5 neuronas en la capa oculta, se obtienen valores mayores para el RMSE sobre el conjunto de validación, por lo que se descartan dichos modelos dado que indican sobreajuste sobre el conjunto de entrenamiento. Con el objetivo de predecir la serie y de llevar a cabo los contrastes de la EMH, se utilizará la ANN que consigue un menor error de predicción sobre el conjunto de validación, que toma como input 5 retardos y tiene 3 neuronas en la capa oculta. En la figura 3.5 se pueden observar tanto la arquitectura de la red (3.5a) como la comparación de las predicciones con los valores reales de la serie sobre el conjunto de validación (3.5b).

La ANN(5 × 3) consigue un RMSE sobre el conjunto de validación de 0,0076 frente al 0,0097 que proporcionaba la predicción del modelo I(1).

3.3.1 Contraste LWG

Para contrastar estadísticamente si la ANN está captando verdaderamente una estructura no-lineal en la serie del rendimiento o simplemente está captando el ruido aleatorio, llevamos a cabo el contraste LWG desarrollado en el capítulo 2. El contraste toma como hipótesis nula la ausencia de estructura no-lineal en los datos, por lo que valores altos del mismo indicarán que existe evidencia a favor de estructuras no-lineales que los métodos clásicos están pasando por alto y, por lo tanto, evidencia en contra de la Hipótesis del Mercado Eficiente.



(a) Estructura de la ANN.

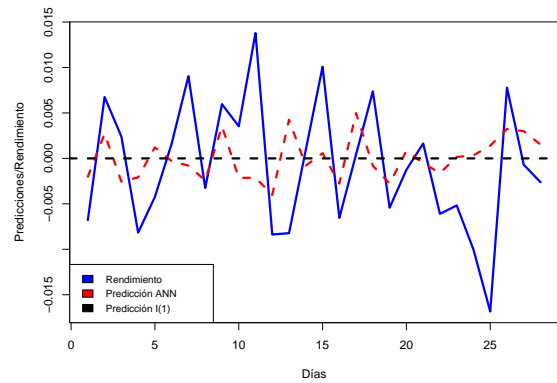

(b) Predicciones y valores reales de r_t .

Figura 3.5: Estructura y predicciones de la ANN. Elaboración propia con R .

Presentamos en el cuadro 3.9 los resultados del contraste LWG para la serie del rendimiento. Dado que el p -valor del contraste superan ampliamente el valor de 0,05, concluimos que la ANN que hemos construido proporciona evidencia a favor de la eficiencia del mercado financiero español durante 2021. Nótese que esto es consistente con la Normalidad de la serie del rendimiento, que implica que la serie del rendimiento, al estar incorrelada consigo misma, está formada por variables aleatorias independientes e idénticamente distribuidas.

Contraste de Lee-White-Granger			
Versión	Valor	Distribución	p -valor
LM (robusto)	3,675	$\chi^2(2)$	0,159

Cuadro 3.9: Resultados del contraste de Lee-White-Granger.

3.3.2 Contraste direccional

Con las predicciones obtenidas con la ANN, podemos evaluar la capacidad predictiva del método de Machine Learning en términos cualitativos, para determinar si al menos es capaz de capturar la estructura de subidas y bajadas de los precios.

En primer lugar, se calcula la métrica MDA para la red neuronal, obteniendo $MDA = 64,29\%$, y a continuación se comprueba si las predicciones cualitativas difieren estadísticamente de un paseo aleatorio a través del contraste direccional de Hylleberg-Dahl.

Real	Predicción		
	Sube	Baja	
Sube	9	4	13
Baja	6	9	15
	13	15	28

Cuadro 3.10: Tabla de contingencia para el contraste de Hylleberg-Dahl.

Con la tabla de contingencia, es inmediato llevar a cabo el contraste direccional de Hylleberg-Dahl, que es en esencia un contraste χ^2 de Pearson. Dado que en una de las categorías caen menos de 5

observaciones, presentamos en el cuadro 3.11 tanto la versión estándar como la versión que incorpora la corrección por continuidad de Yates que se encuentra implementada en todos los paquetes estadísticos.

Contraste Direccional de Hylleberg-Dahl			
Versión	Valor	Distribución	p -valor
Estándar	2,967	$\chi^2(1)$	0,128
Yates	1,365	$\chi^2(1)$	0,243

Cuadro 3.11: Resultados del contraste de Hylleberg-Dahl (Pearson).

En ninguno de los casos se rechaza la hipótesis nula, ya que el valor crítico al 5 % de la distribución $\chi^2(1)$ es $\chi_{0,05}^2(1) = 3,84146$. Por lo tanto, se concluye que las predicciones sobre el conjunto de validación no son estadísticamente distintas de las que se obtienen asumiendo que el rendimiento es puramente aleatorio (es decir, que el logaritmo de los precios es un paseo aleatorio). Esto supone aún más evidencia a favor de la MDH y, por tanto, de la EMH en el mercado español en 2021.

3.4 Conclusiones

La Hipótesis del Mercado Eficiente de Fama implica, por contradictorio que pueda parecer de primeras, que la eficiencia informacional del mercado implica la mayor aleatoriedad posible. Como ya se ha comentado, esto no es más que el reflejo de la competencia perfecta sobre los mercados financieros: si la incorporación de la información es rápida y eficiente, se propaga inmediatamente a través de todos los agentes, por lo que los únicos cambios que no hayan incorporado serán aquellos cambios no anticipados, esto es, aleatorios.

El estudio que hemos realizado, tratando de predecir los rendimientos del mercado tanto con métodos clásicos como con métodos de IA, no ha proporcionado evidencia estadística de que exista predecibilidad en dicha serie. De hecho, se han obtenido algunos resultados que, si bien no son demasiado relevantes en el contraste de la MDH, son poco comunes en el tratamiento de series financieras, tales como la normalidad del rendimiento y la ausencia de efectos ARCH. Aunque las conclusiones no cambiarían de no haberse obtenido creemos que, junto a los resultados obtenidos con los métodos de IA, constituyen una fuerte evidencia a favor de la EMH en el mercado español a lo largo del último año.

Como líneas de investigación posteriores, creemos que podría ser interesante llevar a cabo un estudio utilizando otras técnicas de Machine Learning, como *Support Vector Machines*, *Random Forests* o regresión vía *K-Nearest-Neighbors*, que quizás podrían aportar perspectivas diferentes. Además, también podría ser interesante desarrollar un modelo, si resulta relevante, que trate de predecir las subidas y las bajadas del mercado como objetivo principal (del estilo de los modelos *Probit* o *Logit*, o como los algoritmos de clasificación basados en Machine Learning) y no como objetivo subsidiario, como se ha realizado en este trabajo, simplemente para aprovechar la información que se tenía disponible y evaluar la calidad predictiva de la ANN.

Para finalizar, es muy importante remarcar que los resultados obtenidos están limitados a la metodología seguida y el alcance de las herramientas econométricas utilizadas.

Anexo

En el anexo se incluyen los códigos de *R* desarrollados para llevar a cabo el contraste de la EMH para el mercado financiero español en 2021.

Código de *R*.

```
### Un modelo ARIMA para la serie del IBEX-35 ###
library("plyr")
library("dplyr")
library("ggplot2")
library("plotly")

#install.packages("lubridate")
library("lubridate")

#install.packages("forecast")
library("forecast")

#install.packages("readxl")
library("readxl")

#install.packages("MLmetrics")
library("MLmetrics")

#Se carga la base de datos y se eliminan los posibles datos missing.
ibex35_diaria=read_excel("../preciosibex_nov.xls")
ibex35_diaria=na.omit(ibex35_diaria)

# Se define una variable de lunes que se utilizará en la estimación
ibex35_diaria$monday = ifelse(ibex35_diaria$dia_semana == 'lunes', 1, 0)

# Se formatean correctamente las variables a utilizar
ibex35_diaria$adj_close=as.numeric(ibex35_diaria$adj_close)
ibex35_diaria$Date=as.Date(ibex35_diaria$Date)

# Se realiza un análisis exploratorio gráfico para determinar si la serie de precios es
# Normal. Para ello, se hace el histograma de la serie, se estima la densidad empírica
# utilizando un Kernel Gaussiano y se compara con la densidad Normal.
hist(ibex35_diaria$adj_close, freq=F, col="beige", xlab="Valores", ylab="Frecuencia
relativa", main="")
lines(density(ibex35_diaria$adj_close), add=T, col="blue", lty=2, lwd=3)
curve(dnorm(x, mean(ibex35_diaria$adj_close), sd(ibex35_diaria$adj_close)), add=T,
col="red", lwd=3, lty=2)
legend(x = "topleft", legend = c("Estimación Kernel", "Densidad Normal"), fill = c("blue",
"red"),cex=0.8)
```

```
# Se divide la muestra en entrenamiento y validación
ibex35.train <- subset(ibex35_diaria, Date<=as.Date('2021/10/18',format='%Y/%m/%d') &
  Date>as.Date('2020/12/01',format='%Y/%m/%d'))
ibex35.test <- subset(ibex35_diaria, Date>as.Date('2021/10/14',format='%Y/%m/%d'))
ibex35.test=ibex35.test[-length(ibex35.test),]

# El gráfico de la serie
plot(as.Date(ibex35_diaria$Date, "%d/%m/%Y"),ibex35_diaria$adj_close, type="l", lwd=3,
  col="blue", xlab="", ylab="IBEX-35")

#Con logaritmo
plot(as.Date(ibex35_diaria$Date, "%d/%m/%Y"),log(ibex35_diaria$adj_close), type="l", lwd=3,
  col="blue", xlab="", ylab="IBEX-35")

#No parece estacionaria, se debería hacer una diferencia
library(forecast)
library(lmtest)

#Hago un test ADF para ver si hago la diferencia
library(tseries)
adf.test(ibex35_diaria$adj_close)

#Las funciones de autocorrelación y el ajuste del modelo:
acf(log(ibex35.train$adj_close), main="Función de autocorrelación simple",ci.type="ma",
  ylim=c(0,1))
pacf(log(ibex35.train$adj_close), main="Función de autocorrelación parcial")

ajuste1 <- arima(log(ibex35.train$adj_close),
  order = c(0,1,0),
  method = "ML")
summary(ajuste1)

#Se chequean los residuos para ver si presentan algún patrón
plot(ajuste1$residuals)

acf(ajuste1$residuals, main="Función de autocorrelación simple")
pacf(ajuste1$residuals, main="Función de autocorrelación parcial", ylim=c(-0.2,0.2))

#El ajuste para el rendimiento
ajuste_rdto <- arima(ibex35.train$rdto_diario,
  order = c(0,0,0),
  method = "ML")
ajuste_rdto

prediccion_rdto=predict(ajuste_rdto, n.ahead=28)
prediccion_rdto=as.data.frame(prediccion_rdto)
upper=prediccion_rdto$pred+2*prediccion_rdto$se*x
lower=prediccion_rdto$pred-2*prediccion_rdto$se

#El gráfico:
plot(as.Date(datos.real.pred$Date, "%d/%m/%Y"),datos.real.pred$adj_close, type="l", lwd=3,
  col="blue", xlab="", ylab="IBEX-35", ylim=c(7900,9700))
lines(as.Date(datos.real.pred$Date, "%d/%m/%Y"),datos.real.pred$Prediccion, lwd=3,
```

```

    col="red")
lines(as.Date(datos.real.pred$Date,"%d/%m/%Y"), datos.real.pred$LimSup, lwd=3, lty=2,
      col="orange")
lines(as.Date(datos.real.pred$Date,"%d/%m/%Y"), datos.real.pred$LimInf, lwd=3, lty=2,
      col="orange")
legend(x = "topleft", legend = c("Serie real", "Predicción I(1)"), fill = c("blue",
  "red"),cex=0.8)

#Algunos contrastes de Normalidad del residuo...
residuosDif <- ajuste1$residuals
library(e1071)

skewness(residuosDif)
kurtosis(residuosDif)
hist(residuosDif, nclass=10, col="blue", freq=FALSE, main="", xlab="Residuos",
     ylab="Densidad")
curve(dnorm(x, mean(residuosDif), sd(residuosDif)), lwd=3, col="gold", add=TRUE)

plot(as.Date(ibex35.train$Date,"%d/%m/%Y"),residuosDif, type="l", lwd=2, col="black",
     xlab="", ylab="Residuo")

adf.test(residuosDif)

varianza=9.229e-05

media_cero=mean(residuosDif)

#install.packages("DescTools")
library("DescTools")

#Hacemos un contraste de normalidad de Jarque-Bera
jarque.bera.test(residuosDif)

#... por asegurar, también hacemos un contraste de Kolmogorov-Smirnov.
ks.test(residuosDif, pnorm, mean(residuosDif), sd(residuosDif))

#... usamos también la corrección de Lilliefors.
#install.packages("nortest")
library("nortest")
lillie.test(residuosDif)

#... y aún más, el test de Shapiro-Wilk.
shapiro.test(residuosDif)

#No se rechaza la normalidad de los residuos. Vamos a mirar si es RB.

library(caschro)

Box.test.2(residuosDif,
  nlag = c(6,12,18,24,30,36,42,48),
  type="Ljung-Box")

Box.test.2(residuosDif^2,
  nlag = c(6,12,18,24,30,36,42,48),
  type="Ljung-Box")

```

```
acf(residuosDif^2, ylim=c(-0.2,0.3))
#Nos sale RB, así que no hay que hacer nada más. El hecho de que el contraste JB
#no descarte la normalidad de los residuos impide que existan efectos ARCH. Además no hay
#estructura en el residuo al cuadrado. En efecto, el test de Engle así debería
#confirmarlo.

library("aTSA")
# Arch-test
arch.test(ajuste1)
#No se detectan efectos ARCH.

##### Una red neuronal para la serie del IBEX-35 #####

#Se instalan algunos paquetes más:
library("neuralnet")
library("quantmod")

#Se formatea adecuadamente para poder usar la ANN
rdto=as.ts(ibex35_diaria$rdto_diario)
rdto_d=ibex35_diaria$rdto_diario
rdto_d=rdto_d[-1]

#Sacamos el MSE y el RMSE en validación del ARIMA para comparar
MSE_arima=mean(rdto_d^2)
RMSE_arima=sqrt(MSE_arima)

#Seleccionamos los retardos a usar
y=as.zoo(rdto)
x1=Lag(y, k=1)
x2=Lag(y, k=2)
x3=Lag(y, k=3)
x4=Lag(y, k=4)
x5=Lag(y, k=5)

#Escalamos los datos para que la ANN los pueda leer
rdto=as.zoo(cbind(y,x1,x2,x3, x4, x5))
rdto=rdto[-(1:6),]
maxs=apply(rdto, 2, max)
mins=apply(rdto, 2, min)
rdto_2=as.data.frame(rdto)

index <- 1:220
scaled <- as.data.frame(scale(rdto, center = mins, scale = maxs - mins))
scaled = select(scaled, y, Lag.1, Lag.2, Lag.3, Lag.4, Lag.5)
train_ <- scaled[index,]
test_ <- scaled[-index,]

n <- names(train_)
f <- as.formula(paste("y ~", paste(n[!n %in% "y"], collapse = " + ")))
#Estimamos los pesos de la ANN
nn <- neuralnet(f,data=train_,hidden=3,linear.output=T)

#Graficamos la arquitectura de la ANN
plot(nn)
```

```

pr.nn <- neuralnet::compute(nn,test_[,2:6])
pr.nn_ <- pr.nn$net.result*(max(rdto$y)-min(rdto$y))+min(rdto$y)
test.r <- (test_$y)*(max(rdto$y)-min(rdto$y))+min(rdto$y)

#Calculamos MSE y RMSE de la ANN para comparar con el ARIMA
MSE.nn <- sum((test.r - pr.nn_)^2)/nrow(test_)
RMSE_nn=sqrt(MSE.nn)

tabla=as.data.frame(cbind(pr.nn_, test.r))
print(paste(sqrt(MSE.nn), sqrt(MSE_arima)))

resid=(tabla$test.r-tabla$V1)
mae=mean(abs(resid))

#Graficamos las predicciones todas juntas contra la serie real
plot(tabla$test.r, type="l", col="blue", lwd=3, xlab="Días",
      ylab="Predicciones/Rendimiento")
lines(tabla$V1, col="red", lty=2, lwd=3)
abline(h=0, lwd=3, lty=2, col="black")
legend(x = "bottomleft", legend = c("Rendimiento", "Predicción ANN", "Predicción I(1)"),
      fill = c("blue", "red", "black"),cex=0.8)

plot(1:length(tabla$test.r),tabla$test.r, type="l", col="blue", lwd=3, xlab="Días",
      ylab="Predicciones/Rendimiento", ylim=c(-0.1,0.1))
abline(h=0, lwd=3, lty=2, col="black")
legend(x = "bottomleft", legend = c("Rendimiento", "Predicción I(1)"), fill = c("blue",
      "black"),cex=0.8)

#Con esto ya hemos sacado las predicciones. Ahora hacemos el test de Lee-White-Granger,
#podemos probar para distintos retardos.

#1 retardo
white.test(as.ts(rdto$y), type="Chisq", lags=1, q=3)
white.test(as.ts(rdto$y), type="F", lags=1, q=3)

#2 retardos
white.test(as.ts(rdto$y), type="Chisq", lags=2, q=3)
white.test(as.ts(rdto$y), type="F", lags=2, q=3)

#3 retardos
white.test(as.ts(rdto$y), type="Chisq", lags=3, q=3)
white.test(as.ts(rdto$y), type="F", lags=3, q=3)

#4 retardos
white.test(as.ts(rdto$y), type="Chisq", lags=4, q=3)
white.test(as.ts(rdto$y), type="F", lags=4, q=3)

#5 retardos (la ANN que hemos usado)
white.test(as.ts(rdto$y), type="Chisq", lags=5, q=3)
white.test(as.ts(rdto$y), type="F", lags=5, q=3)

#No hemos rechazado la hipótesis nula nunca, por lo que no obtenemos evidencia en contra de
#la hipótesis del mercado eficiente.

```

```
#Recreamos algunos casos del estudio de Monte-Carlo de Lee, White y Granger para estimar el
  tamaño empírico y la potencia empírica del contraste.
```

```
#Estudio de Monte-Carlo del contraste LWG de no-linealidad
```

```
set.seed(5)
```

```
MC=150
```

```
n=190
```

```
p_ar=numeric(MC)
```

```
p_bilin=numeric(MC)
```

```
for (k in 1:MC){
```

```
  y=rnorm(n, mean=0, sd=2)
```

```
  y_1=Lag(y)
```

```
  error=rnorm(n)
```

```
  error_1=Lag(error)
```

```
  #Genero el valor del AR(1)
```

```
  valor_ar=y-0.7*y_1+error
```

```
  valor_ar=valor_ar[-1,]
```

```
  #Genero el valor del BILIN(1)
```

```
  valor_bilin=y-0.7*y_1*error_1+error
```

```
  valor_bilin=valor_bilin[-1,]
```

```
  #Hago el contraste del AR
```

```
  test_ar=white.test(as.ts(valor_ar), type="Chisq", lag=1, q=3)
```

```
  #Hago el contraste del BILIN
```

```
  test_bilin=white.test(as.ts(valor_bilin), type="Chisq", lag=1, q=3)
```

```
  #Recojo los p-valores
```

```
  p_ar[k]=test_ar$p.value
```

```
  p_bilin[k]=test_bilin$p.value
```

```
}
```

```
tamaño_empirico=1-mean(p_ar>0.05)
```

```
potencia_empirica=mean(p_bilin<0.05)
```

Bibliografía

- [1] BAILLIE, R. T. Y BOLLERSLEV, T. (1991). *The Message in Daily Exchange Rates: A Conditional Variance Tale*. Journal of Business and Economic Statistics, 7, pp. 297-305.
- [2] BIERENS, H. J. (1984). *Model Specification Testing of Time Series Models*. Journal of Econometrics, 26, pp. 323-353.
- [3] BIERENS, H.J. (1987). *ARMAX Model Specification Testing with an Application to Unemployment in the Netherlands*. Journal of Econometrics 35, pp. 161-190.
- [4] BLACK, F. (1976). *Studies of Stock Price Volatility Changes*, Proceedings from the American Statistical Association, Business and Economic Statistics Section, pp. 177-181.
- [5] BOLLERSLEV, T. (1986). *Generalized Autoregressive Conditional Heteroskedasticity*. Journal of Econometrics. 31(3), pp. 307-327.
- [6] BOLLERSLEV, T., RUSSELL, J., Y WATSON, M. (2010). *Volatility and Time Series Econometrics: Essays in Honor of Robert Engle*. Oxford: Oxford University Press. pp. 137-163.
- [7] BOX, G. E. P., JENKINS, G. M., REINSEL, G. C. Y LJUNG, G. M. (2015). *Time Series Analysis: Forecasting and Control*. 5th Revised ed., Wiley.
- [8] CHARPENTIER, A., FLACHAIRE, E. Y LY, A. (2018). *Econometrics and Machine Learning*. Economics and Statistics, 505-506, pp. 147-169.
- [9] CHRISTIE, A. A. (1982). *The Stochastic Behavior of Common Stock Variances: Value, Leverage and Interest Effects*. Journal of Financial Economics, 10, pp. 407-432.
- [10] FAMA, E. F. (1965). *The Behavior of Stock Market Prices*. , Journal of Business, 38, pp. 34-105.
- [11] FAMA, E. F. (1970). *Efficient Capital Markets: A Review of Theory and Empirical Work*. The Journal of Finance, 25(2), pp. 383-417.
- [12] FAN, J. Y YAO, Q. (2017). *The Elements of Financial Econometrics*. Cambridge University Press.
- [13] GARCIA, M. (2021). *Econometría y Predicción*. McGraw-Hill Interamericana de España, S. L.
- [14] GONZÁLEZ-RIVERA, G. Y LEE, T. H. (2007). *Nonlinear Time Series in Financial Forecasting*. Working Papers 200803, University of California at Riverside, Department of Economics.
- [15] HASTIE, T., TIBSHIRANI, R. Y FRIEDMAN, J. (2009). *The Elements of Statistical Learning*. Springer, New York.
- [16] HENRIKSSON, R.D. Y MERTON, R.C. (1981). *On Market Timing and Investment Performance II: Statistical Procedures for Evaluating Forecasting Skills*. The Journal of Business, 54(4), pp. 513-533.

- [17] HYLLEBERG, S. Y DAHL, C.M. (1999). *Specifying Nonlinear Econometric Models by Flexible Regression Models and Relative Forecast Performance*. University of Aarhus, Department of Economics, Working Paper Nors. 1999-4.
- [18] LEE, T. H., WHITE, H. Y GRANGER, C. W. J. (1993). *Testing for Neglected Nonlinearity in Time Series Models*. Journal of Econometrics, 56(3), pp. 269-290.
- [19] MANDELBROT, B. (1963). *The Variation of Certain Speculative Prices*. Journal of Business, 36, 394-419.
- [20] MCNELIS, P. (2004). *Neural Networks in Finance: Gaining Predictive Edge in the Market*. Academic Press Advanced Finance.
- [21] PEÑA, D. (2010). *Análisis de Series Temporales*. Alianza Editorial SA.
- [22] RUIZ, E. (1994). *Modelos para Series Temporales Heterocedásticas*. Cuadernos Económicos del ICE, 1994, vol. 56, pp. 73-108.
- [23] TSAY, R. S. (2005). *Analysis of Financial Time Series*. Hoboken, NJ: Wiley-Interscience.
- [24] WALCZAK, S. (2001). *An Empirical Analysis of Data Requirements for Financial Forecasting with Neural Networks*. Journal of Management Information Systems, 17(4).
- [25] WOOLDRIDGE, J. (1990). *A Unified Approach to Robust, Regression-Based Specification Tests*. Econometric Theory, 6, pp. 17-43.
- [26] WOOLDRIDGE, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press Books, The MIT Press, edition 2, volume 1.