
ESTUDIO DE SERIES TEMPORALES ORIENTADO A ESTRATEGIAS DE PAIR TRADING

TRABAJO FIN DE GRADO

Autor:

Álvaro Videgain Barranco

Tutor:

Miguel Ángel Sánchez Granero

GRADO EN MATEMÁTICAS



JUNIO, 2021
Universidad de Almería

Índice general

1	Introducción	1
2	Lenguaje de programación: Python	3
3	Definiciones y conceptos	5
3.1.	Conceptos matemáticos	5
3.2.	Conceptos financieros	10
4	Resultados: análisis de las series temporales	13
4.1.	Obtención de los datos	13
4.2.	Presentación de los métodos	15
	Método de la distancia, 16.— Método del exponente de Hurst generalizado, 18.— Método de correlación, 21.— Método de la regresión lineal, 23.	
4.3.	Comparación empírica	25
	Experimento con cinco parejas simultáneas, 27.— Experimento con diez parejas simultáneas, 31.— Experimento con veinte parejas simultáneas, 34.— Resultados globales del experimento, 41.	
5	Conclusiones	43
	Bibliografía	45

Abstract in English

In the present work the behaviour of the time series associated to the price of the stocks in the market is studied in order to be able to estimate which will be their behaviour in the future. With this estimations a pair trading strategy is developed. This strategy is focused on four different approaches comparing their efficiency through an empirical experiment.

The first two approaches are implementations of the distance and the generalized Hurst exponent methods. These methods are very common for this kind of study. In this work two additional methods are designed by implementing the Pearson correlation coefficient and by introducing a new measure of distance between series.

In order to compare the efficiency of the four methods, a set of simulations with real data are used, avoiding this way the possible dependence that could exist between the data and the environment of the study. It is specially interesting to compare the results of the newly proposed methods with the ones from the well established approaches.

Among the proposed methods, results show that the one based on the new definition of distance is specially effective, performing even better than the popular generalized Hurst exponent method in certain conditions. Moreover, this implementation is a proof of concept and there is still much room for improvement.

Resumen en español

En el presente trabajo se estudia el comportamiento de las series temporales asociadas a los precios de las acciones en bolsa con el fin de poder estimar cual será el comportamiento de estas en el futuro. Con estas estimaciones se desarrolla una estrategia de pair trading, que es una estrategia de arbitraje estadístico. Esta estrategia se aborda desde cuatro enfoques diferentes, comparando su eficiencia mediante un experimento empírico.

Los dos primeros enfoques son implementaciones de los métodos de la distancia y el exponente de Hurst generalizado, los cuales son bastante populares para este tipo de estudios y han motivado los otros dos métodos desarrollados. El primero de los métodos propuestos consiste en utilizar el coeficiente de correlación de Pearson como medida del movimiento conjunto entre series, mientras que el otro método presentado se nutre del método de la distancia, buscando una forma alternativa de definir la distancia entre dos series.

Para comparar la eficiencia de los cuatro métodos se han realizado simulaciones con datos reales en diferentes condiciones, eliminando así la dependencia que pudiera existir entre los datos y las condiciones del estudio. Es de especial de interés comparar los resultados de los métodos propuestos con las alternativas que se utilizan en el mercado.

Entre los dos métodos propuestos destaca la nueva forma de definir la distancia a partir de los coeficientes de regresión lineal, el cual ha presentado unos resultados bastante buenos a la altura del popular método del exponente de Hurst generalizado. Si bien es cierto que este último ha tenido un comportamiento ligeramente mejor, se trata de una primera implementación del método propuesto con bastante potencial de desarrollo.

Introducción

En este trabajo se plantea el estudio de las series temporales asociadas a los precios de algunas acciones en bolsa para el desarrollo de una estrategia de inversión del tipo pair trading mediante la implementación de cuatro métodos diferentes. El pair trading es un tipo de estrategia financiera que consiste en comprar y vender acciones con un alto grado de correlación, obteniendo beneficios de las pequeñas fluctuaciones entre sus precios. Gerry Bamberger fue un pionero y presentó la estrategia más tarde empleada por la multinacional financiera Morgan Stanley en la década de 1980 [1]. La estrategia pair trading es neutral al mercado ya que al comprar y vender acciones constantemente, esta se beneficia tanto de las subidas como de las caídas de los precios en el mercado, eliminando cualquier relación que pudiera haber entre el comportamiento del mercado y la eficiencia de la estrategia.

Los datos utilizados han sido obtenidos directamente de Yahoo finanzas [2] y están asociados a diferentes acciones del índice estadounidense **S&P500** [3].

Los métodos implementados han sido los siguientes:

- Método de la distancia: Uno de los métodos más populares históricamente para este tipo de estudios.
- Método del exponente de Hurst generalizado: Un enfoque relativamente moderno y muy utilizado en la actualidad.
- Método de la correlación: Basado en un criterio muy recurrente en el estudio de series temporales como es el coeficiente de correlación de Pearson.
- Método de la regresión lineal: Un nuevo enfoque propuesto en este trabajo como alternativa a los métodos existentes.

Con el fin de comparar la eficiencia de los diferentes métodos se ha realizado una simulación utilizando los datos registrados durante los 18 meses previos al inicio del estudio. Para cada método se ha desarrollado una estrategia de pair trading con cinco, diez y veinte pares de series de forma simultánea para comparar los resultados no solo en función del método sino también en función de la cantidad de pares de series con las que se ha desarrollado la estrategia simultáneamente. La simulación ha consistido en dos etapas:

1. Selección de parejas: Los primeros doce meses del experimento han sido dedicados íntegramente a monitorizar el desarrollo de los precios de las acciones con el fin de concluir qué parejas de acciones serán seleccionadas para la simulación.
2. Desarrollo de la estrategia: Consiste en simular el avance de los días realizando las pertinentes operaciones dictadas por cada método y registrando tanto las operaciones realizadas como los beneficios obtenidos durante un periodo de seis meses.

Toda la implementación de los métodos así como la obtención de los datos y posterior simulación ha sido realizada en el lenguaje de programación **Python** [4].

Lenguaje de programación: Python

Python [4] es un lenguaje de programación con un diseño muy centrado en que el código sea fácil de leer. Comenzó a desarrollarse en la década de 1980 por Guido van Rossum (1956) como alternativa al ya entonces obsoleto lenguaje de programación ABC. En 2019 ya había superado a otros lenguajes populares como Java o JavaScript en cantidad de usuarios debido a lo intuitivo y sencillo que es programar en él.

Una de las cosas a destacar de **Python** es el hecho de ser un lenguaje multiparadigma, lo que implica que soporta programación orientada a objetos, programación estructural, programación funcional e incluso programación orientada a aspectos. Además, con la ayuda de extensiones, puede llegar a funcionar con otros tipos de programación como puede ser la lógica.

Otro de los puntos fuertes de **Python** es la oferta de bibliotecas que existe, habilitando herramientas para una infinidad de tareas. Para este trabajo se han utilizado algunas de ellas relacionadas principalmente con el análisis y gestión de datos:

- Numpy [5]: Permite crear vectores y matrices de gran tamaño así como operar con ellos. También incluye una gran colección de funciones matemáticas que van desde las más elementales hasta algunas de alto nivel. Es imprescindible contar con esta biblioteca ya que, aunque **Python** incluye un modo matemático, este es bastante menos potente.
- Matplotlib [6]: Se trata de una biblioteca para la generación de gráficas, cuenta con una amplia personalización en el graficado y una interfaz que recuerda bastante a la de Matlab [7].
- SciPy [8]: Ofrece una amplia variedad de módulos relacionados con optimización, algebra lineal, interpolación, resolución de ecuaciones diferenciales ordinarias e incluso procesamiento de imágenes. Para este trabajo se ha hecho uso del módulo de funciones estadísticas así como del de optimización.
- Pandas [9]: Esta biblioteca introduce una herramienta para gestionar datos que resulta muy útil cuando se trata de series temporales, ya que permite almacenarlas en Dataframes especialmente diseñado con este propósito los cuales facilitan mucho la gestión y operación con este tipo de series. También cuenta con un módulo llamado *pandas datareader* el cual ha sido empleado en la obtención de datos.
- fBm [10]: Es una biblioteca dedicada mayormente a la generación artificial de movimientos Brownianos fraccionales.
- Tqdm [11]: Procede del Arabe taqadam, que significa progreso. Es una biblioteca auxiliar pues no es imprescindible en ningún código, pero resulta de mucha ayuda en la realización de bucles muy largos. Esta biblioteca da la posibilidad de mostrar una barra de progreso en la ejecución del código, permitiendo saber si todo funciona correctamente y el tiempo esperado de ejecución.

Definiciones y conceptos

Durante la realización de este trabajo se ha hecho uso de algunos conceptos matemáticos y financieros que requieren de una definición previa para una mejor comprensión del estudio.

3.1 Conceptos matemáticos

Definición 3.1. *Un proceso estocástico o proceso aleatorio $\mathbf{X} = \{X(t), t \in T\}$ es una familia de variables aleatorias. Esto es, para cada t en el conjunto de índices T , $X(t)$ es una variable aleatoria.*

Es común interpretar t como el tiempo y llamar $X(t)$ al estado del proceso en el tiempo t . Si el conjunto de índices T es un conjunto discreto, llamamos a \mathbf{X} proceso estocástico de tiempo discreto. En nuestro caso el conjunto T de índices está ordenado y corresponde a los instantes temporales asociados a cada día. Para cada uno de estos días está definida la variable aleatoria $X(t)$, de forma que los valores observados para estos instantes temporales constituyen una serie temporal.

Definición 3.2. *Una serie temporal es un conjunto de observaciones X_t tales que cada una de ellas ha sido realizada en un tiempo concreto t . En particular si, como ocurre en nuestro caso, el conjunto T de índices en los que se realizan las observaciones es discreto, se dice que la serie temporal es discreta. Generalmente, los valores del conjunto de índices T están equiespaciados, de donde se obtiene una cierta dependencia entre los datos.*

Se denota como $\{X_t\}_{t=1,\dots,n}$ a la serie temporal de longitud n .

Si bien es cierto que se puede obtener el valor medio de una serie temporal calculando su media aritmética, este valor puede perder el sentido para series suficientemente largas. A continuación se propone una alternativa.

Definición 3.3. *Se denota como $\bar{X}_{t,n}$ a la media de la serie temporal X_t en el tiempo t con ventana n , es decir, a la media aritmética calculada tomando únicamente los valores de la serie en $[t-n, t]$.*

De igual manera se define la desviación típica de la serie para periodos más cortos.

Definición 3.4. *Se denota como $\sigma(X_{t,n})$ a la desviación típica de la serie temporal X_t en el tiempo t con ventana n es decir, teniendo en cuenta solo los valores de la serie en $[t-n, t]$.*

Definición 3.5. *La tendencia de una serie temporal es la componente que indica la trayectoria que va a seguir a largo plazo. No tiene en cuenta las pequeñas fluctuaciones temporales, varía con el tiempo y puede ser estable, descendente, ascendente o una combinación de las dos últimas. Para que la tendencia de una serie tenga sentido, se deben utilizar suficientes observaciones durante su estudio.*

La tendencia es una característica muy importante de las series temporales, ya que en este trabajo se ha buscado trabajar con series que tengan una tendencia suficientemente estable como para tener una idea del comportamiento de la serie en el futuro.

Otra herramienta para estudiar el comportamiento de una serie temporal en el futuro es la estacionariedad.

Definición 3.6. *Una serie temporal se dice estacionaria si la distribución conjunta de cualesquiera momentos de cualquier orden (varianza, esperanza, o momentos de orden superior) junto con la serie nunca depende del tiempo.*

Esta definición es demasiado estricta para ser aplicada a cualquier modelo real, por lo que se hará una diferenciación entre dos tipos de estacionariedad.

Definición 3.7. *Una serie temporal se dice estrictamente estacionaria si la distribución conjunta de cualesquiera momentos de cualquier orden (varianza, esperanza, o momentos de orden superior) junto con la serie nunca depende del tiempo.*

Definición 3.8. *Una serie temporal X_t tiene tendencia estacionaria si es de la forma $X_t = f(t) + e_t$ siendo $f(t) : \mathbb{R} \rightarrow \mathbb{R}$ una función arbitraria y e_t una serie estrictamente estacionaria.*

De ahora en adelante se hará referencia a las series con tendencia estacionaria como series estacionarias.

Si una serie temporal tiene tendencia estacionaria, se puede transformar en estrictamente estacionaria restándole la tendencia, definida por $f(t)$.

Esta definición de estacionariedad nos da suficientes garantías para trabajar en modelos reales sin ser demasiado estricta. En las series estacionarias la media y la varianza permanecen estables, así como la tendencia.

Un caso muy común de proceso estacionario es el ruido blanco.

Definición 3.9. *Se denota ϵ_t al proceso estocástico estacionario llamado ruido blanco, cuyos valores son independientes e idénticamente distribuidos a lo largo del tiempo.*

Al no existir relación entre los puntos de un ruido blanco, la media y la esperanza son cero mientras que la varianza se mantiene estable. También se puede observar que la serie no tiene "memoria", pues los puntos no dependen de los anteriores.

Al contrario que en las series temporales estacionarias, en las no estacionarias se da una variación de la media la cual dibuja una tendencia a crecer o decrecer a largo plazo, por lo que la serie no oscila en torno a una media. Un buen modelo que encaja bien con sucesos estocásticos no estacionarios es el camino aleatorio.

Definición 3.10. *Se dice que una serie temporal X_t sigue un camino aleatorio si es de la forma*

$$X_t = X_{t-1} + \epsilon_t$$

siendo ϵ_t una serie de ruido blanco.

Gráficamente es bastante sencillo diferenciar entre una serie estacionaria y una que sigue un camino aleatorio. A continuación se muestran dos series temporales donde se puede entender gráficamente este concepto.

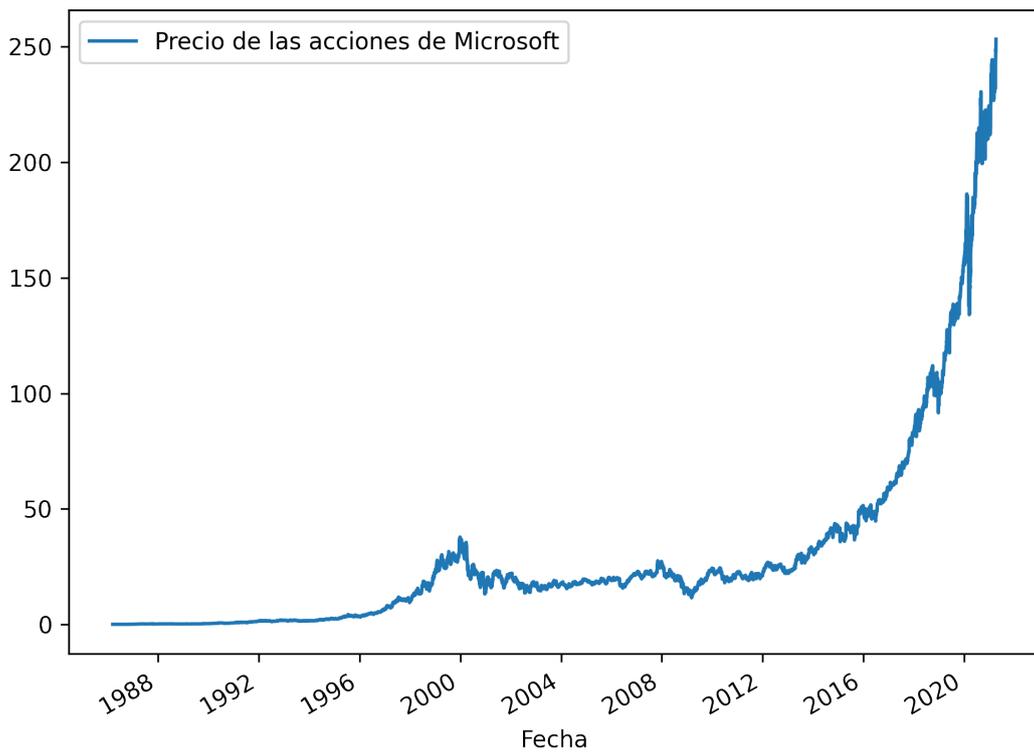


Figura 3.1: Histórico de precios de las acciones de Microsoft.

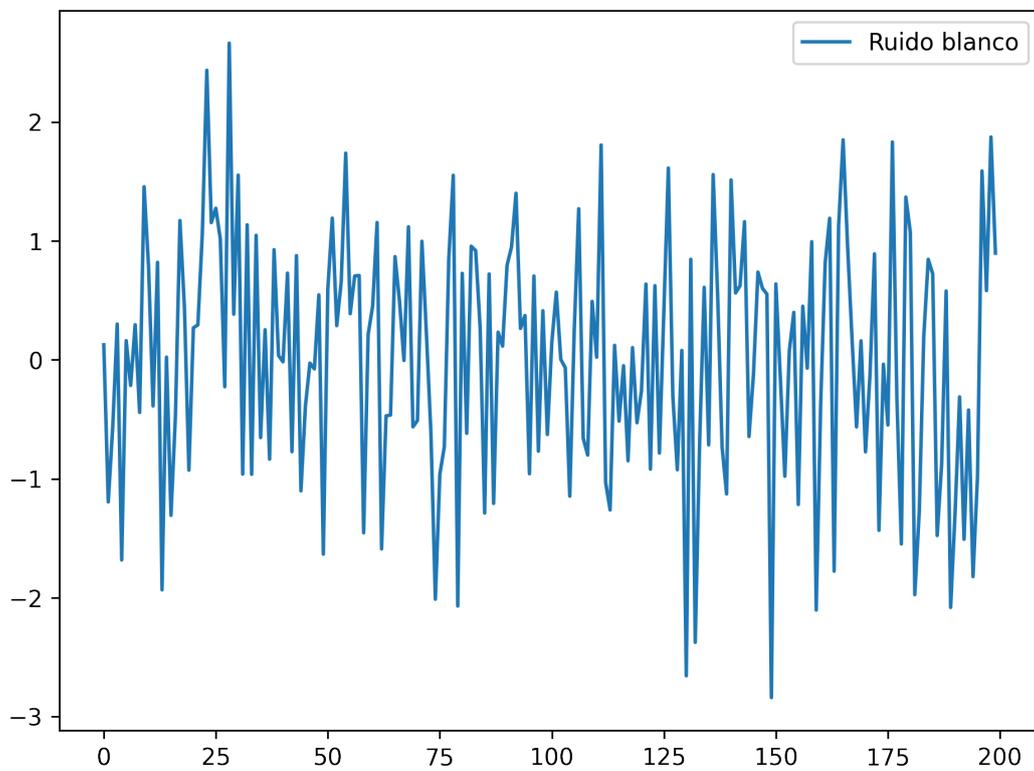


Figura 3.2: Ejemplo de ruido blanco.

La figura 3.1 muestra la evolución del precio de las acciones de la empresa Microsoft desde 1980 hasta nuestros días. Se puede observar como claramente la tendencia es ascendente y no oscila alrededor de ningún valor a excepción de algunos cortos periodos de tiempo, por lo que no es estacionaria.

Por otro lado, la figura 3.2 muestra un ejemplo de ruido blanco, en el que claramente los valores oscilan alrededor del cero dibujando una serie estacionaria.

Generalmente se busca trabajar con series temporales estacionarias ya que ofrecen muchas más posibilidades al tener una tendencia definida. Es por esto que se buscan formas de llevar las series no estacionarias a estacionarias mediante algunas transformaciones. Una de las transformaciones mas elementales es introducir lag a la serie.

Definición 3.11. Sea X_t una serie temporal no estacionaria, se define como X_t con "n" lags la serie $X_t^n = X_t - X_{t-n}$. La serie resultante X_t^n suele ser estacionaria, no obstante esta condición puede no darse.

Se pueden introducir lags en función de como sea la serie. Si tuviera valores diarios se podría introducir un lag de valor 7 para comparar la variación entre cada día y el respectivo de la semana anterior.

El concepto de lag permite introducir la autocorrelación, utilizada para medir la estacionariedad de una serie. La función de autocorrelación deriva directamente de aplicar el coeficiente de relación de Pearson a una serie con la propia serie después de haberle introducido lag.

Definición 3.12. Sean X e Y dos series, se define el coeficiente de correlación de Pearson entre ellas como:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Siendo $Cov(X,Y)$ la covarianza entre X e Y , σ_X la desviación típica de X y σ_Y la desviación típica de Y .

Definición 3.13. Sea X_t una serie temporal, se define la función de autocorrelación con m lags como:

$$R_{XX}(m) = \rho_{X_t, X_{t+m}}$$

Cuando $m = 0$ se obtendrá el coeficiente de orden 0, r_0 , que será uno al estar comparando una serie con ella misma. En general, para $m = k$, siendo n la cantidad total de observaciones de la serie, se obtendrán $n - k$ parejas con puntos separados una distancia m . Calculando el coeficiente de correlación de estas parejas se obtiene el coeficiente de autocorrelación de orden m , r_m .

Otra herramienta bastante importante en el estudio de la memoria de las series temporales es el exponente de Hurst, denotado también como H . Se utiliza para medir en particular la memoria a largo plazo de las series. Está relacionado con la autocorrelación de las series y con como esta disminuye al aumentar la cantidad de lag en la serie.

El exponente de Hurst es también llamado índice de dependencia o índice de dependencia a largo plazo ya que cuantifica la tendencia de una serie temporal por mostrar una fuerte regresión a la media o por moverse en una dirección sin atender a los valores anteriores de la serie.

El exponente de Hurst de una serie siempre es un valor entre 0 y 1. Una serie con valor de H entre 0 y 0,5 define un proceso anti-persistente, por lo que tendrá una reversión a la media. Esta tendencia seguirá a largo plazo, permitiendo estimar el futuro de la serie.

Por otro lado, si el exponente de Hurst para la serie se encuentra entre 0,5 y 1 la serie describe un proceso persistente. Esto significa que si la serie tiene valores altos, estos irán seguidos de otros valores altos y la tendencia de la serie será seguir subiendo.

Por último, en una serie con $H = 0,5$ no se tiene suficiente información para saber como se comportará en el futuro.

El exponente de Hurst puede definirse a partir de procesos estocásticos autosimilares [12].

Definición 3.14. Sean X_t e Y_t dos procesos estocásticos, se dice que ambos procesos tienen la misma función de distribución conjunta finito dimensional si la distribución de X_{t_1}, \dots, X_{t_n} es la misma que la de Y_{t_1}, \dots, Y_{t_n} para todo t_1, \dots, t_n y todo natural n . Esto se denota por $X_t \sim Y_t$.

Definición 3.15. Un proceso estocástico X es H -autosimilar (o autosimilar de parametro H) si existe $H > 0$ tal que $X_{at} \sim a^H X_t$ para todo $a > 0$ y cada $t \geq 0$.

El exponente de Hurst de una serie puede relacionarse con la esperanza de la variación entre las observaciones en función del lag entre estas, definida por $\mathbb{E}(|X_{t+m} - X_t|^2)$. Esta expresión se puede generalizar cambiando el exponente por un término mas general q .

Definición 3.16. Sea X_t una serie temporal define el exponente de hurst generalizado de orden $q > 0$, $H(q)$, a partir de:

$$\frac{\langle |X_{t+m} - X_t|^q \rangle}{\langle |X_t|^q \rangle} \propto m^{q*H(q)}$$

Siendo m el lag entre las observaciones, $\langle \cdot \rangle$ la media de la muestra en esa ventana y donde \propto representa proporcionalidad.

De ahora en adelante, cuando se hable del exponente de Hurst, será realmente el exponente de Hurst generalizado con $q = 1$ al que se estará haciendo referencia.

Otra manera de abordar las series no estacionarias sin perder mucha información es recurrir a la rentabilidad. La rentabilidad en el instante t mide cuan rentables han sido las series el día t , es decir, cuánto han variado desde el día anterior. Por haber trabajado durante la realización de este estudio en escala logarítmica, tiene sentido hablar de rentabilidad logarítmica así como de suma acumulada de la rentabilidad logarítmica. La suma acumulada de la rentabilidad logarítmica en el tiempo t indica el total de variaciones de todos los días previos al instante t medidos en escala logarítmica.

Definición 3.17. Sea X_t una serie temporal, se nota como $SAR_X(t)$ a la suma acumulada de la rentabilidad logarítmica en el instante t dada por:

$$SAR_X(t) = \sum_{i=1}^t \log\left(\frac{X_i}{X_{i-1}}\right) \text{ para } x > 0$$

y es definida como cero en el caso $t = 0$ por no existir aún ninguna rentabilidad.

3.2 Conceptos financieros

Debido a la naturaleza de este estudio es necesario introducir algunos sencillos conceptos financieros para facilitar la comprensión del mismo.

El precio de las acciones en bolsa se encuentra en constante cambio, pero con el fin de poder registrar esos datos, se suelen tomar como series discretas. Para cada acción se almacena cada día su precio a la hora de apertura del mercado así como a la hora de cierre, además del valor de su precio de cierre ajustado.

Mientras el precio de cierre solo representa el precio que se pagó en la última transacción antes del cierre del mercado, el precio ajustado tiene en cuenta factores como el desdoblamiento de acciones, dividendos o cualquier otro que afecte al precio de la acción después de la hora de cierre de los mercados.

Definición 3.18. *Un desdoblamiento de acciones es el aumento de participaciones de una compañía, es decir, un aumento en la cantidad de acciones de esa compañía en el mercado. Esto no cambia la capitalización total del mercado ni el porcentaje que cada usuario tuviera en propiedad de esa compañía.*

Por lo general cuando se dan estos movimientos aumenta la liquidez del mercado pues habrá más compradores para 100 acciones por valor de 1\$ que para 10 acciones por valor de 10\$. A pesar de esto, algunas compañías evitan este tipo de movimientos para mantener cierta exclusividad y por tanto unos precios muy altos. Cualquier tasa de desdoblamiento es válida, pero las más comunes son 2-*por-1*, 3-*por-1* y 3-*por-2*. El ajuste de los precios ante esta situación se realiza de la siguiente forma:

Definición 3.19. *Sea x -*por-y* la tasa de desdoblamiento de acciones realizada por una compañía. El precio ajustado respecto a este factor será $p * \frac{y}{x}$. Siendo p el precio de cierre de la acción registrado antes del desdoblamiento.*

Cabe aclarar que este ajuste se aplica a todos los valores de cierre registrados para días anteriores al del desdoblamiento, de forma que el desdoblamiento no suponga una variación significativa del precio entre dos días consecutivos.

Otro factor muy presente en el ajuste del precio son los dividendos.

Definición 3.20. *Un dividendo es la distribución de beneficios de una compañía entre sus propietarios. Todos los beneficios que no son repartidos se reinvierten en la propia compañía.*

Los dividendos pueden ser cobrados en efectivo o en acciones de la propia compañía y están definidos como una cantidad fija por acción, por lo que la cantidad obtenida por cada propietario es proporcional a la cantidad de acciones que posee. El ajuste del precio ante este factor se realiza a partir de los dividendo por acción.

Definición 3.21. *Siendo el total de dividendos, d , entre el número de acciones, a , la cantidad de dividendos por acción, (d/a) , el precio ajustado para los dividendos es el precio de cierre registrado antes del reparto de dividendos menos la cantidad de dividendos por acción, esto es:*

$$AdjClose = Close - \frac{d}{a}$$

Al igual que ocurre con el ajuste por desdoblamiento, este ajuste se realiza de manera retroactiva para mantener cierta coherencia en la serie de precios.

De ahora en adelante se hablará del precio de cierre ajustado como el precio de las acciones, pues es considerado más completo para el tipo de estudio que se ha realizado.

Una de las estrategias más utilizadas en el mercado financiero para obtener beneficio de la fluctuación de las acciones es el pair trading.

Definición 3.22. *El pair trading es una estrategia de intercambio de acciones basada en monitorizar el precio de dos acciones que históricamente han estado correlacionadas. Cuando la correlación entre ellas se pierde momentáneamente, es decir, una sube y/o la otra baja, se abre la posición invirtiendo en corto la que está por encima de su valor habitual y en largo la que está por debajo de este, esperando que la correlación del par vuelva a la normalidad. Una vez esto sucede se cierra la posición, invirtiendo en largo y corto las que habíamos invertido en corto y largo respectivamente.*

Invertir en corto o largo se puede ver de forma muy simplificada como vender y comprar acciones respectivamente.

Definición 3.23. *Se llama venta en corto o posición corta a la práctica de vender unos activos con la intención de comprar exactamente los mismos en el futuro a un precio menor. Al contrario de lo que pueda parecer por el nombre de esta práctica, el periodo de tiempo durante el que desarrolla no es necesariamente corto.*

Aunque vender en corto es una estrategia que se realiza buscando que el precio vaya a la baja, siempre existe el riesgo de que el precio suba, pudiendo la pérdida superar la cantidad invertida al no tener los precios de las acciones techo.

Definición 3.24. *Se llama compra en largo o posición larga a la estrategia de comprar un activo con la esperanza de que suba su precio para obtener beneficio a la hora de la venta.*

La posición larga suele ser la adoptada por la mayoría de personas cuando implementan una estrategia de inversión.

Cuando se está desarrollando una estrategia de pair trading con un par de acciones, las operaciones se realizan de forma conjunta. Esto significa que cuando se venda una cantidad de una de las acciones, se comprará proporcionalmente una cantidad de la otra acción. Esta operación de comprar y vender en conjunto es llamada abrir posición.

Definición 3.25. *Durante el desarrollo de una estrategia pair trading se conoce como abrir posición al acto de invertir en corto una de las acciones del par e invertir en largo la otra con el fin de deshacer la operación en el futuro. Al abrir la posición se dice que se está vendiendo o comprando el par en función de cual de las acciones es la que se posiciona en corto.*

Definición 3.26. *Durante el desarrollo de una estrategia pair trading, se conoce como cerrar una posición la acción de terminar la venta en corto y largo de acciones que se había comenzado en el momento en que esta posición fue abierta.*

Definición 3.27. *Sea el par de acciones con el que se está desarrollando una estrategia de pair trading de la forma $A - \gamma B$,*

- *Se dice que se compra el par cuando se invierten en corto γ acciones de B por cada una que se invierte en largo de A.*

- *Se dice que se vende el par cuando se invierten en largo y acciones de B por cada una que se invierte en corto de A.*

Una cualidad muy importante de las estrategias del tipo pair trading es que son neutrales al mercado. Al vender y comprar acciones en cada posición abierta, estas estrategias pueden beneficiarse tanto de las subidas como bajadas del mercado, eliminando la correlación entre la estrategia y el comportamiento de este. Es por esto que no se deben comparar los resultados obtenidos con la evolución del mercado, ya que este tipo de estrategias entran dentro de las llamadas de *retorno absoluto*. Una consecuencia directa de esta característica es que las estrategias de esta forma funcionarán muy bien en conjunto con estrategias que dependan del mercado, pues al no haber prácticamente correlación entre ellas, el riesgo total será bastante bajo.

Cabe finalmente destacar que en el mercado financiero existen mecanismos que ofrecen la posibilidad de vender acciones sin tenerlas en propiedad, por lo que lo único necesario para comenzar a realizar pair trading con una pareja de acciones es monitorear sus precios y realizar una inversión.

Resultados: análisis de las series temporales

El principal objetivo de este estudio ha sido la obtención de parejas de acciones tales que las series temporales asociadas a su precio tuvieran una relación suficientemente alta como para poder ser implementadas en una estrategia de pair trading.

A pesar de la no estacionariedad de las series de acciones, si las dos series son suficientemente parecidas, la serie del par, definida generalmente como la diferencia entre las series, es estacionaria. Sucede algo parecido a la transformación de una serie en estacionaria mediante la introducción de lags. Por este motivo se ha trabajado mayormente con las series asociadas a los pares directamente.

4.1 Obtención de los datos

Para la realización de este estudio se ha trabajado con el índice **S&P500** [3] por ser considerado el índice más representativo de la situación real del mercado. Abarca aproximadamente el 80% del mercado de Estados Unidos. Además ha sido considerada una buena opción para este tipo de trabajo debido a la diversidad en cuanto a los tipos de acciones de los que se compone.

Se ha trabajado con todas las acciones de este índice a excepción de Brown-Forman Corporation (BF.B) y Berkshire Hathaway (BRK.B) debido a ciertas dificultades en la obtención de los datos y Carrier Global corporation (CARR), Lumen Technologies (LUMN), Otis Worldwide (OTIS), ViacomCBS (VIAC) y Vontier corporation (VNT) por no tener suficientes observaciones para un estudio adecuado debido a su entrada en este mercado a lo largo del año y medio previo a la realización del estudio.

Las fechas de las observaciones tomadas para la realización del estudio han sido desde el 27 de Agosto de 2019 hasta el 22 de Febrero de 2021. Teniendo en cuenta que un año en el mercado financiero se compone de aproximadamente 250 días, se ha trabajado con datos que abarcan un año y medio. El primer año de observaciones ha sido destinado íntegramente al estudio de las parejas con el fin de elegir que pares eran más indicados para ser utilizadas en un estrategia de pair trading. Una vez elegidos los pares, se ha desarrollado la estrategia mencionada durante un periodo de seis meses.

Para la obtención de datos hemos utilizado la biblioteca *pandasdatareader*. El proceso ha consistido en obtener todos los símbolos asociados a las acciones mediante la lectura de una tabla [3] con el siguiente código.

```
from pandas_datareader import data
import pandas as pd

table=pd.read_html(
'https://en.wikipedia.org/wiki/List\_of\_S%26P\_500\_companies')
df = table[0]
df.to_csv('S&P500-Info.csv')
df.to_csv("S&P500-Symbols.csv", columns=['Symbol'])
```

Para después tomar el histórico de precios de cada acción desde Yahoo finanzas [2] en el tiempo que se ha empleado para la realización del estudio y almacenarlo en un dataframe mediante el siguiente código:

```
import pandas as pd
from pandas_datareader import data
import datetime

acciones={}
for x in tqdm(range(505)):
    try:
        accion=df['Symbol'][x]
        acciones[accion]=data.DataReader(accion,
            'yahoo', datetime.datetime(2019, 8, 27),
            datetime.datetime(2021, 2, 22))
    except:
        print('problemas al obtener datos', accion)
```

Este proceso genera un dataframe con los datos del precio máximo y mínimo, el precio de apertura, de cierre y de cierre ajustado para cada día, así como el volumen de acciones movidas (Cuadro 4.1), pero al ser únicamente de interés para este estudio los datos del cierre ajustado, se ha reducido a otro dataframe que contiene exclusivamente esos datos (Cuadro 4.2) mediante el código:

```
import pandas as pd

cierres={x:y['Adj Close'] for x,y in acciones.items()}
df=pd.DataFrame(cierres)
df.to_pickle(
    '/content/gdrive/My Drive/uai/gcolab/'+
    'videgain/datosExperimento.pkl')
```

La última línea de este código almacena el dataframe generado en un archivo de forma local para evitar tener que generar el archivo nuevamente en cada sesión de trabajo.

Fecha	Máximo	Mínimo	Apertura	Cierre	Volumen	Cierre Ajustado
2019-08-27	136.7200	134.6600	136.3899	135.7400	23102100.0	133.5539
2019-08-28	135.7599	133.5500	134.8800	135.5599	17393300.0	133.3768
2019-08-29	138.4400	136.9100	137.2500	138.1199	20168700.0	135.8956
2019-08-30	139.1799	136.2700	139.1499	137.8600	23940100.0	135.6398
...
2021-02-17	244.3099	240.9400	241.3200	244.1999	21653500.0	244.1999
2021-02-18	243.9299	240.8600	241.8000	243.7899	16925600.0	243.7899
2021-02-19	243.8600	240.1799	243.7500	240.9700	25262600.0	240.9700
2021-02-22	237.9299	232.3999	237.4199	234.5099	36446900.0	234.5099

Cuadro 4.1: Dataframe con histórico de todos los precios de las acciones de Microsoft.

Fecha	Cierre Ajustado
2019-08-27	133.5539
2019-08-28	133.3768
2019-08-29	135.8956
2019-08-30	135.6398
...	...
2021-02-17	244.1999
2021-02-18	243.7899
2021-02-19	240.9700
2021-02-22	234.5099

Cuadro 4.2: Dataframe con histórico de precios ajustados de las acciones de Microsoft.

4.2 Presentación de los métodos

Existen diversos métodos para implementar la estrategia de pair trading, en la realización de este estudio se ha trabajado con dos métodos propuestos anteriormente llamados método de la distancia [13] y [14] y método del exponente de Hurst generalizado [15] y [16] así como con dos métodos propuestos basados en el coeficiente de correlación de Pearson y en la regresión lineal mediante mínimos cuadrados ordinarios.

Durante la realización de este trabajo se han aplicado los cuatro métodos mencionados anteriormente, monitorizando durante un año las diferentes acciones con el fin de elegir las parejas más adecuadas para a continuación simular una operación en el mercado financiero desarrollando una estrategia de pair trading siguiendo los criterios de cada método.

Cabe señalar que durante el estudio se ha trabajado mayormente con el logaritmo del precio ajustado de las acciones, pues aporta una visión mas clara de la evolución de los precios. La principal ventaja de la escala logarítmica respecto a la escala aritmética resultante de trabajar con los precios ajustados directamente es que en la escala logarítmica los cambios que representen el mismo porcentaje abarcan la misma distancia, mientras que en la escala aritmética la distancia que un cambio representa es proporcional a la diferencia entre los valores. Si una acción pasa de valer 8\$ a 9\$, al ser un cambio del 12,5%, en una gráfica con escala logarítmica estará representado igual que un cambio de 10000\$ a 11250\$. Esto resulta muy importante si las variaciones de los precios cubren distintos porcentajes. Si el precio de una acción baja de 150\$ a 8\$ y luego subiera a 9\$, en una escala aritmética, debido a que la escala abarca hasta 150\$, la diferencia entre 8\$ y 9\$ apenas se verá representada gráficamente, mientras que en una escala logarítmica, al suponer el cambio un 12,5%, se seguirá viendo representado el segundo cambio a pesar de que el primero fuera mucho mas significativo.

Método de la distancia

Este método ha sido implementado como se propone en [13], dividiéndolo en dos fases claramente diferenciadas. La primera fase consiste en seleccionar las series más convenientes para la estrategia a desarrollar.

Bajo el planteamiento de este método, el movimiento conjunto de una pareja de series viene dado por la "distancia", entendiendo la distancia como la suma del valor absoluto de la diferencia entre la suma acumulada de la rentabilidad logarítmica (Definición 3.17) para cada instante de tiempo t .

Definición 4.1. Sean X_t, Y_t dos series temporales de igual longitud, se define $D(X_t, Y_t)$ como la distancia entre ambas series, dada por:

$$D(X_t, Y_t) = \sum_{i=1}^n |SAR_Y(i) - SAR_X(i)| \quad (4.1)$$

Siendo n el número de valores observados en cada serie y $SAR_X(t)$ y $SAR_Y(t)$ la suma acumulada de las rentabilidades logarítmicas en el instante t de las series X e Y respectivamente.

La implementación del cálculo de distancia entre series ha sido realizada mediante el siguiente código:

```
import pandas as pd
import numpy as np

def distanceMethod(par):
    accion1=np.log(par.iloc[:,0])
    accion2=np.log(par.iloc[:,1])
    accion1=accion1-accion1.iloc[0]
    accion2=accion2-accion2.iloc[0]
    distancia=np.sum(np.abs(accion1-accion2))
    return (distancia)
```

La entrada requerida par es un dataframe que incluye dos columnas, representando cada una los valores de una de las series. El método no podrá funcionar correctamente si la longitud de ambas series no coincide o si para algún día t no hay registro del valor de alguna de las dos acciones.

El primer paso que realiza el código es almacenar como $accion1$ los logaritmos de los precios registrados para la primera acción para a continuación hacer lo propio para $accion2$. Lo siguiente es restar a todos los valores de la serie obtenida el primer valor de esta. En ese paso se está transformando la serie de los logaritmos del precio en la de la suma acumulada de la rentabilidad logarítmica, ya que en el cálculo de la suma acumulada de la rentabilidad logarítmica se simplifican todos los valores a excepción del primero y el último.

$$\begin{aligned} \log(X_t) - \log(X_0) &= \log(X_t) - \log(X_{t-1}) + \log(X_{t-1}) - \log(X_{t-2}) + \dots \\ &+ \log(X_2) - \log(X_1) + \log(X_1) - \log(X_0) = \sum_{i=1}^t \log\left(\frac{X_i}{X_{i-1}}\right) = SAR(t) \end{aligned}$$

Por último se calcula la distancia entre las series siguiendo la definición anterior (Ecuación 4.1).

Mediante la aplicación de ese método a todas las parejas candidatas y un posterior examen de los resultado se concluyen las parejas óptimas para el desarrollo de la estrategia pair trading.

La segunda fase consiste en establecer los criterios por los cuales se rige el método para desarrollar la estrategia. Para estos criterios se trabaja con lo que se denomina la serie del par.

Definición 4.2. En el método de la distancia, se denomina serie del par X_t, Y_t a la diferencia entre la suma acumulada de la rentabilidad logarítmica de los precios de las acciones para cada instante de tiempo t nombrada como $(XY)_t^D$, es decir:

$$(XY)_t^D = SAR_X(t) - SAR_Y(t), \forall t$$

El criterio utilizado en este método para abrir posición es que la serie del par en el instante de tiempo t sea mayor que la suma de su media y dos veces su desviación típica o menor que la suma de su media menos dos veces su desviación típica, entendiendo la media (Definición 3.3) y la desviación típica (Definición 3.4) como los respectivos valores en el tiempo t con una ventana de longitud cuarenta. Esto es:

$$(XY)_t^D > \overline{(XY)}_{t,40}^D + 2\sigma((XY)_{t,40}^D) \quad (4.2)$$

ó

$$(XY)_t^D < \overline{(XY)}_{t,40}^D - 2\sigma((XY)_{t,40}^D) \quad (4.3)$$

En la apertura de posición, si el valor de la acción X_t está por encima del de Y_t , es decir, se satisface la Ecuación 4.2, se vende el par. En este método, dada la forma en la que se define la serie del par, se dedicará una mitad del presupuesto a invertir en corto acciones de X_t y la otra mitad a invertir en largo acciones de Y_t .

Por otro lado, si el valor de Y_t es superior al de X_t en el momento de abrir posición, se cumple la Ecuación 4.3. En ese caso se compra el par, destinando la mitad del presupuesto a invertir en corto acciones de Y_t mientras se emplea la otra mitad en invertir en largo acciones de X_t .

Una vez abierta la posición, no se opera hasta que no se cumpla alguno de los criterios de cierre de posición. Estos criterios varían en función de algunos parámetros como los beneficios potenciales o el riesgo a correr. Para la realización de este estudio se ha optado por establecer unos criterios de cierre con una ambición moderada, priorizando que el riesgo sea bajo.

El principal criterio para cerrar una operación es que la serie del par haya vuelto a la normalidad, siguiendo la reversión a la media propia de las series estacionarias. Esto se ve representado de diferente forma si se ha comprado o vendido el par.

- En el caso en el que se ha vendido el par se consideraba suficiente reversión a la media una vez la serie del par es menor que la suma de la media más la desviación típica, esto es:

$$(XY)_t^D < \overline{(XY)}_{t,40}^D + \sigma((XY)_{t,40}^D)$$

- Si por el contrario se ha comprado el par, la reversión a la media se considera suficiente cuando la serie del par está por encima de la media menos la desviación típica, es decir:

$$(XY)_t^D > \overline{(XY)}_{t,40}^D - \sigma((XY)_{t,40}^D)$$

- El último criterio de cierre se define para el caso en el que el movimiento conjunto de las acciones con las que se ha trabajado se rompe. Las series dejan de moverse juntas, al menos temporalmente. En ese momento se cierra la posición, generalmente con pérdidas, para evitar que la pérdida sea aún mayor. Para la realización de este trabajo se ha considerado como perdido el movimiento conjunto de dos acciones cuando la serie del par dista más de cuatro veces la desviación típica de la media de esta. Esto se refleja como

$$(XY)_t^D > \overline{(XY)}_{t,40}^D + 4\sigma((XY)_{t,40}^D)$$

en caso de que se haya vendido el par y como

$$(XY)_t^D < \overline{(XY)}_{t,40}^D - 4\sigma((XY)_{t,40}^D)$$

si el par ha sido comprado.

Con el fin de mejorar la estrategia se contempla aplicar dos restricciones más a la hora de operar.

1. La primera restricción consiste en establecer un tiempo máximo de operación. Si en el tiempo establecido no se ha cumplido uno de los criterios de cierre, se cierra la posición al alcanzarse el límite de tiempo.
2. La segunda restricción es una consecuencia directa de la primera. Si una operación ha sido cerrada antes de que se de una reversión a la media en la serie del par, ese par queda excluido de cualquier operación hasta que la serie del par revierta a la media. Si las dos acciones no se vuelven a mover juntas no se volverá a operar con ellas.

Para la realización del estudio se ha optado por no incluir estas dos restricciones adicionales, pues se ha considerado que el tercer criterio de cierre de operación cumple las funciones de estas restricciones.

Método del exponente de Hurst generalizado

El exponente de Hurst fue introducido por primera en 1951 por el hidrólogo inglés Harold Edwin Hurst (1880-1978) [16] con el fin de lidiar con los problemas que generaba la espontaneidad del caudal del río Nilo durante la temporada de lluvias.

Abreviado generalmente como método de Hurst, se ha implementado según lo propuesto en [15].

La implementación del exponente de Hurst está motivada por el fin de buscar una alternativa a la co-integración que de resultados parecidos. La co-integración mide muy bien el movimiento conjunto entre dos series, pero es muy exigente en cuanto a encontrar pares válidos. Teóricamente la co-integración funciona muy bien, pero la realidad es que si se toman periodos de tiempo extensos va a ser muy difícil encontrar

series co-integradas. Con la idea de medir el movimiento conjunto con una alta precisión pero sin que haya tanto problema para encontrar parejas de series válidas se ha implementado el método de Hurst.

El algoritmo seleccionado para trabajar con el exponente de Hurst durante este estudio ha sido el GHE (del inglés *generalized Hurst exponent*), basado en que la distribución de los incrementos $|X_{t+m} - X_t|$ es la misma que la de $m^H |X_{t+1} - X_t|$. Este algoritmo se nutre del exponente de Hurst generalizado presentado anteriormente (Definición 3.16) al estar este relacionado con la dependencia a largo plazo de las series temporales. Para aplicar este algoritmo primero consideramos el estadístico de la definición:

$$\frac{\langle |X_{t+m} - X_t|^q \rangle}{\langle |X_t|^q \rangle} \propto m^{q*H(q)} \quad (4.4)$$

Teniendo en cuenta que m puede variar entre uno y un cuarto de la longitud de la serie.

El exponente de Hurst generalizado se calcula como la media de observaciones obtenidas de la Expresión 4.4, pero también se puede obtener a partir de un estadístico alternativo:

$$\langle |X_{t+m} - X_t|^q \rangle = \sum_{t=1}^{n-m} \frac{|X_{t+m} - X_t|^q}{n - m + 1} \propto m^{q*H(q)}$$

Siendo n la longitud de la serie X_t .

Aunque se ha implementado este método para cualquier valor de q , durante la realización de este estudio se ha trabajado con $q = 1$, lo cual reporta un exponente de Hurst generalizado bastante similar al exponente de Hurst básico.

A diferencia de otros métodos, el método de Hurst admite una única serie como entrada, por lo que se ha trabajado directamente con la serie del par. Para este método se ha definido la serie del par de forma diferente a la anterior, pues se ha añadido un parámetro adicional respecto al método de la distancia.

Definición 4.3. En el método de Hurst, se denomina serie del par X_t, Y_t , nombrada como $(XY)_t^H$:

$$(XY)_t^H = \log(X_t) - b * \log(Y_t), \forall t$$

siendo b el cociente entre las desviaciones estándar de las rentabilidades logarítmicas de X_t e Y_t respectivamente.

$$b = \frac{std(R_X(t))}{std(R_Y(t))}$$

Donde $R_X(t)$ y $R_Y(t)$ representan la rentabilidad logarítmica de X_t e Y_t respectivamente.

A pesar de que el algoritmo para el cálculo del exponente de Hurst solo admita una serie como entrada, la implementación del método de Hurst en Python se ha realizado de forma que la entrada sea un dataframe que incluya dos columnas, representando cada una los valores de una de las series. El código admite cualquier valor para q , tomando $q = 1$ en caso de que no se proporcione ningún valor.

Lo primero que hace el código a partir de un par es calcular el parámetro b para poder formar la serie del par. A continuación genera la serie del par y comprueba que los datos introducidos están en el formato deseado, transformándolos de no ser así. El siguiente paso es elegir los valores para los que se toman las observaciones. Se han elegido los valores de la forma 2^k para que estén equiespaciados a pesar de estar trabajando con logaritmos.

A continuación se ejecuta un bucle que realiza una serie de cálculos para $T = 2^k$ con k entre uno y el máximo posible tal que 2^k sea menor que la longitud total de la serie. Calcula el momento absoluto de orden q de los incrementos con T lags, para después hacer tanto el logaritmo de esa media, como de T . Al estar este algoritmo basado en la premisa de que $\langle |X_{t+T} - X_t| \rangle$ es proporcional a $T^{qH} \langle |X_{t+1} - X_t| \rangle$, tiene sentido tomar logaritmos obteniendo

$$\log(mT) = qH \log(T) + \log(m1) + c$$

Siendo mT la media en t de los incrementos con T lag y $m1$ el incremento con un lag y c una constante.

Al haberse realizado estos cálculos para cada valor de T , se tiene en realidad una lista de parejas mT, T , por lo que se puede obtener $H(q)$ como la pendiente de la regresión de la lista de $\log(mT)$ sobre $\log(T)$ dividida por q .

```
import pandas as pd
import numpy as np
from scipy import stats

def genHurstExponent(par, q=1):
    b=(np.log(par.iloc[:,0]).diff().dropna().std())/
    (np.log(par.iloc[:,1]).diff().dropna().std())
    serie=np.log(par.iloc[:,0])-b*np.log(par.iloc[:,1])
    serie=serie.values if hasattr(serie, 'values') else serie
    kmax=int(np.log2(len(serie))-2)
    regx=[]
    regy=[]
    for k in range(kmax):
        T=2**k
        mT=np.mean(np.abs(serie[T:]-serie[:-T])**q)
        regx.append(np.log(T))
        regy.append(np.log(mT))
    stats.linregress(regx, regy)
    H=stats.linregress(regx, regy).slope/q
    return (H,b)
```

Cabe apuntar que el código no sólo devuelve el valor del exponente de Hurst calculado, sino que devuelve también la b empleada en la formación de la serie del par. Esto se debe a que todos los criterios de apertura y cierre de posición se definen en función de la serie del par tal y cómo esta ha sido definida, es decir, haciendo uso del parámetro b .

Los criterios de apertura y cierre de posición son bastante parecidos a los definidos para el método anterior, siendo en esta ocasión la serie del par algo diferente, implementando el parámetro b como dicta la Definición 4.3.

La posición se abre cuando la serie del par en el instante t dista más de dos veces la desviación típica de la media. Esto es:

$$(XY)_t^H > \overline{(XY)}_{t,40}^H + 2\sigma((XY)_{t,40}^H) \quad (4.5)$$

$$(XY)_t^H < \overline{(XY)}_{t,40}^H - 2\sigma((XY)_{t,40}^H) \quad (4.6)$$

Siendo $(XY)_t^H$ la serie del par en el instante t y $\overline{(XY)}_{t,40}^H$ y $\sigma((XY)_{t,40}^H)$ la media y la desviación típica respectivamente de la serie X_t en el instante t con una ventana de longitud 40.

El par se vende cuando se satisface la Ecuación 4.5 y se compra si se satisface la Ecuación 4.6. Una vez abierta la posición, se espera a que la serie del par vuelva a oscilar alrededor de la media. Se considera que esta oscilando alrededor de la media cuando

$$(XY)_t^H < \overline{(XY)}_{t,40}^H + \sigma((XY)_{t,40}^H)$$

En caso de que se hubiera vendido el par. Y

$$(XY)_t^H > \overline{(XY)}_{t,40}^H - \sigma((XY)_{t,40}^H)$$

Cuando el par se ha comprado.

En los métodos para los cuales la serie asociada a cada par sea la resta de la suma acumulada de la rentabilidad logarítmica de la primera menos la de la segunda, el presupuesto destinado a operar con cada acción cuando se compra o vende el par es la mitad del presupuesto total, pero esto cambia al construir la serie del par de distinta forma. Para este método en particular, dada la estructura de la serie del par, los presupuestos se reparten de la siguiente forma:

- Cuando se vende el par se invierte $\frac{P}{1+b}$ en posicionar en corto la primera acción y $\frac{bP}{1+b}$ en posicionar en largo la segunda, siendo P el presupuesto total y b el coeficiente del método.
- Cuando se compra el par se invierte $\frac{bP}{1+b}$ en corto en la segunda acción y $\frac{P}{1+b}$ en largo la primera, siendo P el presupuesto total y b el coeficiente del método.

Existe un tercer criterio de cierre de posición que cubre los casos en los que el movimiento conjunto de las acciones se rompe. Si los precios de las acciones se han separado demasiado, se cierra la posición automáticamente. Esto se puede expresar como

$$(XY)_t^H > \overline{(XY)}_{t,40}^H + 4\sigma((XY)_{t,40}^H)$$

en caso de que se haya vendido el par y de la forma

$$(XY)_t^H < \overline{(XY)}_{t,40}^H - 4\sigma((XY)_{t,40}^H)$$

si el par ha sido comprado.

Como se ha introducido en la sección anterior, una serie con exponente de Hurst menor que 0,5 está asociada a un proceso anti-persistente, es decir, con reversión a la media. Por tanto para el empleo de este método en una estrategia pair trading, se han formado todas las parejas posibles y seleccionado aquellas con un valor de H más bajo.

Método de correlación

La relación entre el movimiento conjunto de dos series y su coeficiente de correlación es un tema muy recurrente en el estudio de series temporales en general, en particular para el pair trading. Esto es debido a que generalmente las parejas de series

con una buena proyección para el desarrollo de una estrategia de pair trading suelen presentar un coeficiente de correlación alto. Esta relación viene dada generalmente por los valores en el mercado a las que las dos series están asociadas. Si ambas acciones están asociadas a empresas del mismo sector, generalmente los eventos que afecten al sector tendrán repercusión sobre el precio de las dos acciones. Existen también parejas de series asociadas a empresas de sectores completamente independientes que están de igual manera correlacionadas, aunque el coeficiente de correlación del par suele ser menor.

La principal desventaja de utilizar la correlación como factor para desarrollar una estrategia de pair trading es que es más complicado detectar cuando las series se separan y cuando revierten a la media, dado que estos movimientos son bastante discretos, debido a la baja volatilidad entre las series.

El uso de la correlación en la realización de este trabajo ha consistido en la aplicación directa del coeficiente de correlación de Pearson, definido anteriormente como:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Siendo $\text{cov}(X, Y)$ la covarianza entre X e Y , σ_X la varianza de X y σ_Y la varianza de Y .

Al ser el coeficiente de correlación de Pearson simétrico, ha sido suficiente con calcularlo una única vez para cada par de series. Por estar trabajando con escala logarítmica, se ha calculado la correlación para los logaritmos de los precios.

```
import pandas as pd
import numpy as np

def correlacion(par):
    accion1=np.log(par.iloc[:,0])
    accion2=np.log(par.iloc[:,1])
    correlacion=np.corrcoef(accion1,accion2)
    return(correlacion[0,1])
```

Cabe destacar que el método de **numpy** llamado *corrcoef* admite una entrada de n series y devuelve una matriz simétrica ($n \times n$) con todos los coeficientes de correlación de Pearson entre las n series de entrada dos a dos, es por eso que la salida del código se especifica que se devuelva solo el valor de la fila uno y segunda columna.

Al tratarse de un método que se nutre directamente del coeficiente de correlación de Pearson, la serie de un par de acciones se define de la misma forma que para el método de la distancia.

Definición 4.4. En el método de la correlación, se denomina serie del par X_t, Y_t a la diferencia entre la suma acumulada de la rentabilidad logarítmica de los precios de las acciones para cada instante de tiempo t nombrada como $(XY)_t^\rho$, es decir:

$$(XY)_t^\rho = SAR_X(t) - SAR_Y(t), \forall t$$

Los criterios de apertura y cierre se definen siguiendo un perfil bajo de riesgo al igual que en los métodos anteriores:

- El par se vende si se satisface:

$$(XY)_t^\rho > \overline{(XY)^\rho}_{t,40} + 2\sigma((XY)^\rho_{t,40})$$

- Se compra el par si se satisface:

$$(XY)_t^\rho < \overline{(XY)}_{t,40}^\rho - 2\sigma((XY)_{t,40}^\rho)$$

- Se cierra posición si se ha vendido el par y se satisface:

$$(XY)_t^\rho < \overline{(XY)}_{t,40}^\rho + \sigma((XY)_{t,40}^\rho)$$

- En caso de haber comprado el par, se cierra al satisfacerse:

$$(XY)_t^\rho > \overline{(XY)}_{t,40}^\rho - \sigma((XY)_{t,40}^\rho)$$

- Por último, se cierra posición si se ha vendido el par y:

$$(XY)_t^\rho > \overline{(XY)}_{t,40}^\rho + 4\sigma((XY)_{t,40}^\rho)$$

o en el caso de que se haya comprado y se satisfaga:

$$(XY)_t^\rho < \overline{(XY)}_{t,40}^\rho - 4\sigma((XY)_{t,40}^\rho)$$

Siendo $\overline{(XY)}_{t,40}^\rho$ y $\sigma((XY)_{t,40}^\rho)$ la media y desviación típica de la serie del par en el instante t con una ventana de longitud 40 respectivamente.

Método de la regresión lineal

El método que se propone basado en la regresión lineal nace motivado por la definición de esta. Se puede aproximar una serie de valores a otra utilizando únicamente un valor para la pendiente y otro para el intercepto. Intuitivamente cabe pensar que si la regresión lineal de Y sobre X genera unos valores m y n tales que $Y = mX + n$, cuanto más se parezcan m y n a 1 y 0 respectivamente, más se parecerán Y y X .

Partiendo de esta motivación se ha definido una función regresión con el fin de cuantificar el parecido entre dos series temporales a partir de los coeficientes de su regresión lineal ya que este parecido estará estrechamente relacionado con su movimiento conjunto.

Para la definición de esta función se ha decidido trabajar con la rentabilidad logarítmica de las acciones por ser su serie asociada bastante más lineal que la correspondiente serie asociada a los precios de las acciones.

El siguiente paso es definir el peso que tienen en una regresión lineal la pendiente y el intercepto. Michael Hochman [17] propone una métrica para calcular la distancia entre similaridades de la forma $\varphi(x) = ax + b$.

Definición 4.5. Se define la distancia $d(\cdot, \cdot)$ entre dos similaridades $\varphi(x) = mx + n$ y $\varphi'(x) = m'x + n'$ como

$$d(\varphi(x), \varphi'(x)) = |n - n'| + |\log m - \log m'|$$

Con base en esa definición de métrica se ha adaptado la aplicación de esa función a la distancia entre $Y = mX + n$ e $Y = X$, ya que lo que se quiere medir es cuanto difiere una serie de la otra, y en caso de ser iguales la regresión sería de la forma $Y = X$.

Uno de los inconvenientes que surgen de la definición propuesta es que no contempla la posibilidad de que m sea negativo, lo cuál es algo que sucede en las regresiones lineales. Esto no supone un problema al estar buscando parejas de series que se muevan de forma parecida. Cuando la pendiente de una regresión lineal es negativa, la relación entre las series es inversa, por lo que no existe un gran parecido entre el comportamiento de estas. Por tanto esta definición de distancia se ha implementado únicamente para pares de series cuya regresión lineal tenga una pendiente positiva, descartando directamente las que no cumplan esta condición.

Otro factor a tener en cuenta es la obtención de resultados diferentes en la realización de una regresión de X sobre Y que en una de Y sobre X , por lo que se ha elegido el coeficiente que mide su parecido como el que menos similitud entre las series represente.

Definición 4.6. Se define el coeficiente de separación por regresión lineal $r(X,Y)$ entre dos series cuya regresión lineal tiene pendiente positiva como el máximo entre

$$|n_X| + |\log(m_X)|$$

y

$$|n_Y| + |\log(m_Y)|$$

Siendo m_X y m_Y las pendientes producto de hacer una regresión lineal de X sobre Y y de Y sobre X respectivamente y n_X y n_Y sus respectivos interceptos.

Cabe señalar que esta función no está definida para el caso en el que alguna de las dos pendientes obtenidas valga cero, pero no se ha contemplado ese caso pues ninguna de las series asociadas a una acción presenta una perfecta gráfica paralela al eje X .

La implementación de este método se ha realizado mediante el siguiente código.

```
def RegLin(par):
    accion1=np.log(par.iloc[:,0])
    accion2=np.log(par.iloc[:,1])
    accion1=accion1.diff().dropna()
    accion2=accion2.diff().dropna()
    (slope1,intercept1,rvalue1,pvalue1,stderr1)=stats.linregress(
        accion1,accion2)
    if slope1 > 0:
        f1=np.abs(intercept1)+np.abs(np.log(slope1))
    else:
        f1=np.inf
    (slope2,intercept2,rvalue2,pvalue2,stderr2)=stats.linregress(
        accion2,accion1)
    if slope2 > 0:
        f2=np.abs(intercept2)+np.abs(np.log(slope2))
    else:
        f2=np.inf
    return(max(f1,f2))
```

Al igual que los otros métodos presentados, está definido de tal forma que tiene una única entrada de dos columnas donde cada columna representa la serie asociada a los precios de una de las acciones.

En la implementación se ha definido la distancia como infinito en el caso de que la pendiente de la regresión fuera negativa para un mejor funcionamiento.

Los criterios de apertura y cierre establecidos para este método son similares a los de los métodos introducidos anteriormente.

Definición 4.7. En el método de la regresión lineal se define la serie del par X_t , Y_t como la diferencia entre la suma acumulada de la rentabilidad logarítmica de los precios de las acciones para cada instante de tiempo t .

$$(XY)_t^r = SAR_X(t) - SAR_Y(t), \forall t$$

- El par se vende si se satisface:

$$(XY)_t^r > \overline{(XY)}_{t,40}^r + 2\sigma((XY)_{t,40}^r)$$

- Se compra el par si se satisface:

$$(XY)_t^r < \overline{(XY)}_{t,40}^r - 2\sigma((XY)_{t,40}^r)$$

- Se cierra posición si se ha vendido el par y se satisface:

$$(XY)_t^r < \overline{(XY)}_{t,40}^r + \sigma((XY)_{t,40}^r)$$

- En caso de haber comprado el par, se cierra al satisfacerse:

$$(XY)_t^r > \overline{(XY)}_{t,40}^r - \sigma((XY)_{t,40}^r)$$

- Por último, se cierra posición si se ha vendido el par y:

$$(XY)_t^r > \overline{(XY)}_{t,40}^r + 4\sigma((XY)_{t,40}^r)$$

o en el caso de que se haya comprado y se satisfaga:

$$(XY)_t^r < \overline{(XY)}_{t,40}^r - 4\sigma((XY)_{t,40}^r)$$

Siendo $\overline{(XY)}_{t,40}^r$ y $\sigma((XY)_{t,40}^r)$ la media y desviación típica de la serie del par en el instante t con una ventana de longitud 40 respectivamente.

4.3 Comparación empírica

Con el fin de comprobar que método de los propuestos funciona mejor bajo unas condiciones dadas, se ha planteado un experimento que consiste en simular el desarrollo de una estrategia de pair trading con cada uno de los métodos durante seis meses. Como se ha mencionado anteriormente, los datos utilizados para la realización de este estudio han sido los recogidos desde el 27 de Agosto de 2019 hasta el 22 de Febrero de 2021, siendo el periodo de selección de parejas desde el principio hasta el 24 de Agosto de 2020, donde se inicia la simulación de intercambio de acciones hasta el 22 de Febrero de 2021.

El experimento ha consistido en asignar a cada método una cartera de cien mil dólares, por estar operando en el mercado estadounidense, y simular su desarrollo para n parejas de series en paralelo. Los criterios de apertura y cierre de posición han sido los definidos en la presentación de cada método y en los casos en los que una posición estuviera abierta al término del periodo experimental, se ha cerrado en ese momento respetando los precios que estaban vigentes en el momento que fue abierta, de forma que esa última posición no se ha tenido en cuenta. En este caso n ha tomado los valores $n = 5, 10, 20$ con el fin de ver si la eficacia de los métodos está relacionada con la cantidad de parejas elegidas para la estrategia.

En primer lugar se han hecho todas las parejas posibles dentro del conjunto de series con el que vamos a trabajar. Para cada pareja se ha comprobado que haya datos registrados de todos los días del periodo de formación de pares y en caso afirmativo se le han aplicado los cuatro métodos.

```
resultados=[]
for x in range(503):
    for t in range(x+1,503):
        symbols=[df.columns[x],df.columns[t]]
        par=df[symbols].iloc[-375:-125].dropna()
        if len(par.iloc[:,0])==250 and len(par.iloc[:,1])==250:
            resultados.append([symbols,distanceMethod(par),
                genHurstExponent(par),correlacion(par),RegLin(par)])
```

Cabe señalar que el segundo bucle *for* abarca solo desde $x+1$ en adelante para evitar que se tengan en cuenta las parejas de una serie con ella misma así como que las parejas se formen dos veces. Se han evitado las parejas duplicadas ya que todos los métodos se han programado de forma simétrica, es decir, obteniendo el mismo resultado al aplicarlo para A y B que para B y A .

Hay que tener en cuenta también que para cada pareja de acciones se han almacenado cinco valores numéricos a pesar de haberles aplicado solo cuatro métodos. Esto es debido a que el método del exponente generalizado de Hurst devuelve, además del resultado, la b utilizada para formar la serie del par.

Una vez obtenidos los datos necesarios para todas las posibles parejas, se eligen las mejores n parejas para el desarrollo de cada método. En los métodos de la distancia, regresión lineal y exponente de Hurst generalizado, se toman los n valores con el respectivo coeficiente más bajo, mientras que en el de correlación se seleccionan las n series con mayor coeficiente de correlación.

Una vez elegidos los pares se da paso a la simulación de la estrategia pair trading. La simulación se ha realizado de forma independiente para cada método, asignando a cada una de los n pares con los que se ha trabajado $\frac{100,000}{n}$ dólares como cartera inicial. A partir de este punto el desarrollo del experimento es cíclico, desarrollándose para cada pareja de acciones el siguiente proceso:

Para cada día con la posición cerrada se comprueban los criterios de apertura de posición, si algunos de ellos se cumple se compra o vende el par según indique el criterio. Si se vende el par formado por las acciones con series temporales asociadas X_t e Y_t , se destinará la mitad de la cartera disponible a invertir en largo tantas acciones como se pueda de Y_t y la otra mitad a invertir en corto tantas como el presupuesto permita de X_t . En el caso de comprar el par se invertirán en largo y en corto acciones de X_t y Y_t respectivamente. Cabe aclarar que el presupuesto se reparte de esa forma

para tres de los cuatro métodos presentados, pero que en el método del exponente generalizado de Hurst el presupuesto se reparte como se ha definido en la introducción del método.

En el momento que se realiza cualquiera de estas dos operaciones se actualiza la cartera, restando el dinero invertido en posicionar acciones en largo y sumando el ingreso obtenido al invertir en corto las otras acciones. Además en la realización de este estudio se han almacenado los precios a los que se ha abierto posición para los casos en los que la posición no se hubiera cerrado al terminar el tiempo de experimento.

Se actualiza también la cantidad de acciones en la cartera. Las acciones vendidas en corto se añaden en negativo pues se deben volver a comprar en el futuro mientras que las compradas en largo se añaden en positivo ya que serán vendidas a la hora de cerrar la posición.

Una vez se tiene la posición abierta se comprueban cada día las condiciones de cierre de posición en función de si la posición se ha abierto mediante la compra o venta del par. Una vez se satisface alguna de esas condiciones se terminan las posiciones en largo y corto abiertas, vendiendo y comprando las acciones que se habían comprado y vendido respectivamente. La cartera se actualiza añadiendo el beneficio de cerrar la posición en largo y restando el gasto de comprar las acciones vendidas en corto y las acciones se actualizan valiendo cero en ambos casos pues ya no hay acciones posicionadas. En el caso de que el criterio satisfecho para el cierre de la posición haya sido considerar que el movimiento conjunto se ha perdido temporalmente, la apertura de posiciones en esa pareja queda bloqueada hasta que se experimente una reversión a la media en la serie del par.

Al encontrarse nuevamente la posición cerrada se volvería al punto inicial del desarrollo de este experimento, comprobando los criterios de apertura de posición cada día.

Una vez terminado el periodo experimental se eliminan las posiciones que no han podido cerrarse a tiempo utilizando los valores registrados en el momento que fueron abiertas, simulando así nunca haber sido iniciadas.

Este experimento ha sido realizado para cinco, diez y veinte parejas diferentes.

Experimento con cinco parejas simultáneas

En la realización del experimento con únicamente cinco pares por método se ha trabajado con los siguientes pares para los métodos de la distancia, Hurst, correlación y regresión lineal respectivamente.

	Par de acciones	Distancia
1	[GOOGL, GOOG]	0.793
2	[FOXA, FOX]	2.324
3	[UAA, UA]	2.574
4	[NWSA, NWS]	3.413
5	[MA, V]	4.419

Cuadro 4.3: Mejores cinco series para el método de la distancia.

	Par de acciones	Exponente de Hurst	Valor de b
1	[PEAK, PAYX]	0.2505	1.214
2	[UAA, UA]	0.2535	1.041
3	[GOOGL, GOOG]	0.2565	1.006
4	[MSCI, NFLX]	0.2680	1.211
5	[RTX, VNO]	0.2684	0.877

Cuadro 4.4: Mejores cinco series para el método de Hurst.

	Par de acciones	Coefficiente de correlación
1	[GOOGL, GOOG]	0.999768
2	[UAA, UA]	0.999588
3	[FOXA, FOX]	0.997708
4	[DISCA, DISCK]	0.997354
5	[NWSA, NWS]	0.996668

Cuadro 4.5: Mejores cinco series para el método de la correlación.

	Par de acciones	Coefficiente de regresión lineal
1	[GOOGL, GOOG]	0.008
2	[FOXA, FOX]	0.024
3	[DISCA, DISCK]	0.028
4	[CFG, KEY]	0.037
5	[NWSA, NWS]	0.051

Cuadro 4.6: Mejores cinco series para el método de la regresión lineal.

Llama la atención la presencia de el par $[GOOGL, GOOG]$ entre las seleccionadas de todos los métodos, pero tiene sentido pues las dos acciones son de *Google*. Cuando una compañía comienza a cotizar en bolsa y una cantidad suficiente de sus acciones es vendida, los propietarios originales pueden perder el control sobre ella. Para evitar esto se suele dividir la compañía en dos y aunque ambas salen al mercado, de una de ellas no se pueden obtener más del cincuenta por ciento de las acciones.

Dado que la naturaleza de las acciones es la misma tiene sentido su alto grado de movimiento conjunto ya que el precio oscilará en función de los mismos factores externos. Algo similar pasa con las parejas $[FOXA, FOX]$, $[UAA, UA]$, $[NWSA, NWS]$ y $[DISCA, DISCK]$. El hecho de que pertenecieran originalmente a la misma compañía no supone un problema a la hora de operar con ellas, pero dado su alto grado de relación se darán menos oportunidades de abrir posición. En la siguiente gráfica se ve claramente el alto movimiento conjunto que presenta la pareja $[GOOGL, GOOG]$.

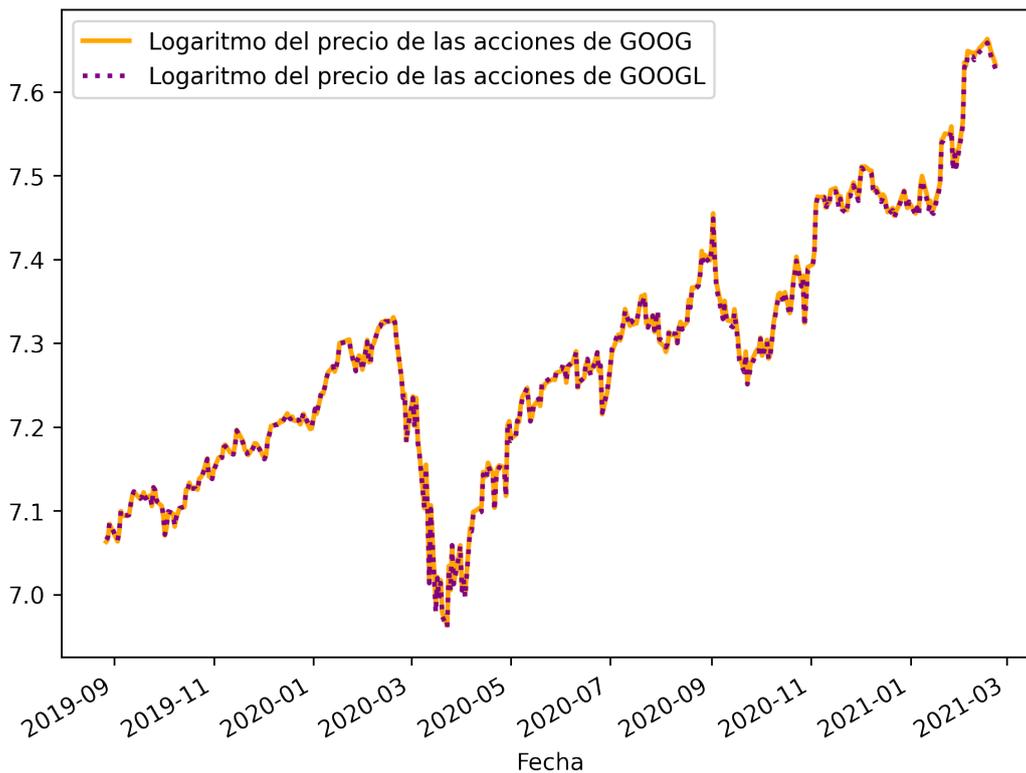


Figura 4.1: Logaritmo del precio de las acciones *GOOGL* y *GOOG*.

Las cartera asignada a cada par en este caso ha sido de 20000 dólares, obteniendo los siguientes resultados.

Método	Porcentaje de beneficios	Posiciones cerradas
Distancia	1.156%	20
Hurst	5.571%	20
Correlación	0.231%	17
Regresión lineal	1.246%	17

Cuadro 4.7: Resultados de los cuatro métodos para cinco parejas.

Se puede observar que los resultado, aunque positivos, son bastante bajos debido al bajo riesgo que se ha elegido correr a la hora de establecer los criterios de apertura y cierre. El método de Hurst funciona bastante mejor que el resto, siendo el de correlación el peor de ellos. A continuación se muestran los resultados separados en los pares elegidos para cada método.

Cabe señalar que en este desglose por pares se han contabilizado como operaciones tanto la apertura como el cierre de posición de forma que si el número de operaciones realizadas n es par, se habrán completado $\frac{n}{2}$ posiciones completas, pero si son impares, la última posición se habrá abierto pero por no poder cerrarse en tiempo no se habrá tenido en cuenta.

	Par de acciones	Cartera final	Número de operaciones realizadas
1	[GOOGL, GOOG]	20055.432	4
2	[FOXA, FOX]	20121.943	8
3	[UAA, UA]	20026.419	9
4	[NWSA, NWS]	20353.154	12
5	[MA, V]	20599.277	8

Cuadro 4.8: Resultados del método de la distancia para cinco pares.

Como cabía de esperar, la pareja [GOOGL, GOOG] es muy segura pero genera pocos beneficios habiéndose presentado a lo largo de seis meses únicamente dos oportunidades de abrir posición. La pareja que más beneficios ha generado es [MA, V], la cuál está formada por las acciones de Mastercard y Visa respectivamente. Esto se debe en parte a que a pesar de pertenecer al mismo sector del mercado, al ser las dos acciones pertenecientes a diferentes compañías, se generarán mas oportunidades para abrir posición.

	Par de acciones	Cartera final	Número de operaciones realizadas
1	[PEAK, PAYX]	21685.006	12
2	[UAA, UA]	20218.299	9
3	[GOOGL, GOOG]	20053.230	4
4	[MSCI, NFLX]	22377.322	12
5	[RTX, VNO]	21237.194	4

Cuadro 4.9: Resultados del método de Hurst para cinco pares.

Se puede observar que el método de Hurst funciona bastante mejor que los otros métodos, al menos para cinco series. Las series identificadas han tenido bastante buen rendimiento a pesar de que ninguna de ellas presentaba un exponente de Hurst menor a $\frac{1}{4}$. Al igual que en el resto de métodos se repite el *efecto Google*, cerrando únicamente dos posiciones y con unos beneficios muy bajos, pero en esta ocasión el resto de series cierra más posiciones y lo hace con bastante buenos resultados. Destaca la quinta pareja, la cual ha generado un 6,19% de beneficios en tan sólo dos posiciones, las mismas que la pareja de *Google*.

	Par de acciones	Cartera final	Número de operaciones realizadas
1	[GOOGL, GOOG]	20055.432	4
2	[UAA, UA]	20026.419	9
3	[FOXA, FOX]	20121.943	8
4	[DISCA, DISCK]	19673.850	2
5	[NWSA, NWS]	20353.154	12

Cuadro 4.10: Resultados del método de la correlación para cinco pares.

El método de la correlación ha funcionado bastante mal a pesar de haber elegido todos los pares con una correlación superior al 99%. Estando los pares tan correla-

cionados cabía esperar que el *efecto Google* se extendiese a los demás pares generando pocos beneficios, pero no solo ocurre esto si no que además la cuarta pareja presenta en tan solo una posición unas pérdidas elevadas en relación con los beneficios de los demás pares.

	Par de acciones	Cartera final	Número de operaciones realizadas
1	[GOOGL, GOOG]	20055.432	4
2	[FOXA, FOX]	20121.943	8
3	[DISCA, DISCK]	19673.850	2
4	[CFG, KEY]	21041.774	8
5	[NWSA, NWS]	20353.154	12

Cuadro 4.11: Resultados del método de la regresión lineal para cinco pares.

Los resultados del método propuesto son los mejores después del método de Hurst cuando se trata de cinco series. Los resultados no distan demasiado de los del método de correlación pues ambos métodos coinciden en cuatro de los cinco pares seleccionados, pero en este caso la otra pareja ha generado unos beneficios del 5,21 % en cuatro posiciones mientras que la correspondiente pareja elegida por el método de correlación presentaba unos beneficios del 0,13 %.

Experimento con diez parejas simultáneas

Para el caso del experimento en el que cada método trabaja con diez pares de acciones simultáneamente se presentan los pares seleccionados junto a los resultados obtenidos directamente. Al igual que en el caso anterior se han contabilizado como operaciones tanto la apertura como el cierre de posición para de esta forma tener también información sobre si quedó una posición abierta la cual no se cerró antes de terminar el periodo experimental. A continuación se muestran los pares con los que ha trabajado cada método.

	Par de acciones	Distancia	Cartera final	Operaciones realizadas
1	[GOOGL, GOOG]	0.793	10027.756	4
2	[FOXA, FOX]	2.324	10060.978	8
3	[UAA, UA]	2.574	10013.252	9
4	[NWSA, NWS]	3.413	10176.610	12
5	[MA, V]	4.419	10299.580	8
6	[BAC, JPM]	4.440	9780.962	6
7	[ATO, CMS]	5.437	9875.715	5
8	[AWK, XEL]	5.567	9619.558	8
9	[LNT, CMS]	5.580	10453.600	8
10	[DISCA, DISCK]	5.594	9836.716	2

Cuadro 4.12: Resultados del método de la distancia para diez pares de acciones.

4. RESULTADOS: ANÁLISIS DE LAS SERIES TEMPORALES

Puede observarse que de las cinco parejas añadidas respecto al caso anterior solo una ha presentado beneficios mientras que las otras cuatro han presentado pérdidas. Al ser series con una distancia mayor, su movimiento conjunto es menos estricto por lo que es mayor el riesgo, pero también es mayor el beneficio potencial que se puede obtener.

A pesar de las pérdidas en cuatro de las diez parejas, el balance general ha sido positivo aunque los beneficios obtenidos son muy bajos.

	Par de acciones	Exponente de Hurst	Cartera final	Operaciones realizadas
1	[PEAK, PAYX]	0.2505	10842.939	12
2	[UAA, UA]	0.2535	10107.209	9
3	[GOOGL, GOOG]	0.2565	10022.660	4
4	[MSCI, NFLX]	0.2680	11166.955	12
5	[RTX, VNO]	0.2684	10616.475	4
6	[AMAT, NDAQ]	0.2704	9424.410	7
7	[NDAQ, TFX]	0.2754	10412.304	8
8	[CFG, RCL]	0.2777	9310.867	6
9	[EW, JNPR]	0.2795	9789.668	8
10	[SRE, SBUX]	0.2799	9900.667	8

Cuadro 4.13: Resultados del método de Hurst para diez pares de acciones.

Las cinco parejas añadidas al método de Hurst respecto al caso anterior también han presentado resultados negativos, reportando cuatro de ellas pérdidas. El único par añadido que ha generado beneficios tampoco ha sido brillante pues tres de los cinco primeros pares con un valor menor del exponente de Hurst han presentado más beneficios. El riesgo ha sido mayor mientras que los resultados han empeorado notablemente.

	Par de acciones	Coefficiente de correlación	Cartera final	Operaciones realizadas
1	[GOOGL, GOOG]	0.999768	10027.756	4
2	[UAA, UA]	0.999588	10013.252	9
3	[FOXA, FOX]	0.997708	10060.978	8
4	[DISCA, DISCK]	0.997354	9836.716	2
5	[NWSA, NWS]	0.996668	10176.610	12
6	[CMA, HBAN]	0.993864	10278.945	6
7	[FRT, REG]	0.993693	10399.014	10
8	[CCL, NCLH]	0.993552	10320.180	7
9	[DFS, SYF]	0.993240	9796.706	8
10	[CFG, FITB]	0.992993	10171.928	8

Cuadro 4.14: Resultados del método de correlación para diez pares de acciones.

El método de la correlación, al contrario que los dos métodos anteriores, ha mejorado su rendimiento con el aumento de pares a trabajar simultáneamente. Cuatro de los cinco pares incorporados respecto al caso anterior presentan beneficios y lo hacen en una proporción considerablemente mayor que los cinco primeros. Mientras que los cinco primeros pares presentaban un beneficio conjunto del 0,23%, los cinco siguientes en conjunto presentan un beneficio del 1,93%. Esto se debe a que, aunque las oportunidades de abrir posición han sido equivalentes, en las series menos correlacionadas, las posiciones abiertas presentan mayor potencial de beneficios por tener las series del par un movimiento conjunto menor.

	Par de acciones	Coefficiente de regresión lineal	Cartera final	Operaciones realizadas
1	[GOOGL, GOOG]	0.008	10027.756	4
2	[FOXA, FOX]	0.024	10060.978	8
3	[DISCA, DISCK]	0.028	9836.716	2
4	[CFG, KEY]	0.037	10520.912	8
5	[NWSA, NWS]	0.051	10176.610	12
6	[UAA, UA]	0.058	10013.252	9
7	[FITB, KEY]	0.059	9909.526	6
8	[CFG, FITB]	0.065	10171.928	8
9	[ESS, UDR]	0.073	10686.354	10
10	[CMS, XEL]	0.077	10191.563	4

Cuadro 4.15: Resultados del método de la regresión lineal para diez pares de acciones.

El método propuesto a partir de las regresiones lineales también ha mejorado su rendimiento con el aumento de series a desarrollar en paralelo, aunque en menor medida que el método de la correlación. Cuatro de los cinco pares añadidos han presentado beneficios, siendo los beneficios de los cinco pares añadidos ligeramente mayores que los de los cinco primeros, aunque la diferencia es bastante baja.

Se presentan a continuación los resultados resumidos para cada método, representando como posiciones cerradas el número de posiciones que se han abierto y cerrado durante el periodo experimental, a diferencia del número de operaciones utilizado anteriormente en el desarrollo de los resultados para cada método. Llama la atención que mientras entre los pares seleccionados en el caso anterior por cada método se repetían bastante las mismas parejas, entre los cinco pares anexados por cada método existe mucha más diversidad.

Método	Porcentaje de beneficios	Posiciones cerradas
Distancia	0.145%	34
Hurst	1.594%	36
Correlación	1.082%	36
Regresión lineal	1.596%	35

Cuadro 4.16: Resultados de los cuatro métodos para diez parejas.

Para trabajar con diez parejas simultáneamente funcionan mejor los métodos de regresión lineal y el exponente de Hurst generalizado, aunque este último ha sufrido una pérdida de efectividad bastante considerable con el aumento de pares a desarrollar. El otro gran perdedor de este aumento de parejas ha sido el método de la distancia, el cual ha pasado de presentar un beneficio superior al uno por ciento a presentar un beneficio prácticamente nulo. Por último, el método de la correlación el cual presentaba unos beneficios muy bajos cuando operaba con cinco pares, ha mejorado sus beneficios situándolos por encima del uno por ciento.

Experimento con veinte parejas simultáneas

Se presentan a continuación los resultados obtenidos en el experimento realizado tomando para cada método veinte pares diferentes con los que desarrollar la estrategia. Dada la cantidad de parejas seleccionadas, se puede observar que mientras en los casos anteriores las series de las parejas seleccionadas se movían de forma muy parecida, ahora habrá parejas cuyo movimiento no sea tan parecido. Estos pares presentarán muchas mas opciones de abrir posición así como mayores beneficios potenciales en estas posiciones, pero también un riesgo bastante más elevado.

	Par de acciones	Distancia	Cartera final	Operaciones realizadas
1	[GOOGL, GOOG]	0.793	5013.832	4
2	[FOXA, FOX]	2.324	5030.435	8
3	[UAA, UA]	2.574	5006.624	9
4	[NWSA, NWS]	3.413	5088.286	12
5	[MA, V]	4.419	5149.756	8
6	[BAC, JPM]	4.440	4890.500	6
7	[ATO, CMS]	5.437	4937.886	5
8	[AWK, XEL]	5.567	4809.836	8
9	[LNT, CMS]	5.580	5226.757	8
10	[DISCA, DISCK]	5.594	4918.149	2
11	[CMS, MDLZ]	5.610	5120.549	8
12	[AME, NSC]	5.865	5136.485	6
13	[LNT, AEP]	5.953	5075.574	4
14	[AEP, ATO]	5.959	4881.788	5
15	[PEP, PG]	6.006	4909.084	2
16	[CSX, UNP]	6.061	5167.008	15
17	[MDT, SRE]	6.103	4943.608	10
18	[RF, TFC]	6.164	5157.736	12
19	[ETR, SRE]	6.317	5013.814	5
20	[CMI, PCAR]	6.415	5279.956	12

Cuadro 4.17: Resultados del método de la distancia para veinte pares de acciones.

Vemos que de las diez parejas que se han añadido respecto al caso anterior para el método de la distancia, siete de ellas presentan beneficios, aunque solo los beneficios de cinco de ella son relevantes.

Llama la atención en particular la pareja $[CSX, UNP]$, en la cual se han cerrado correctamente siete posiciones y abierto una más la cual no se ha podido cerrar antes de que terminase el tiempo de simulación del experimento. Esto está relacionado con lo que se ha comentado anteriormente al respecto del movimiento conjunto entre las series y la cantidad de oportunidades de abrir posición que presenta el par. Este fenómeno se puede ver reflejado claramente en las últimas cinco parejas del experimento, las cuales tienen una mayor distancia entre las series que el resto de las seleccionadas. Solo uno de esos cinco pares ha cerrado menos de cinco posiciones satisfactoriamente, mientras que en todas las demás series del experimento, únicamente una pareja ha cerrado más de cinco posiciones.

En general el método ha tenido un mejor rendimiento que en el caso para diez series por método, pero peor que en el caso para cinco series por método. Mientras que los diez pares añadidos presentan una mejora respecto al caso anterior, pesan demasiado las parejas de la seis a la diez.

	Par de acciones	Exponente de Hurst	Cartera final	Operaciones realizadas
1	[PEAK, PAYX]	0.2505	5420.399	12
2	[UAA, UA]	0.2535	5053.500	9
3	[GOOGL, GOOG]	0.2565	5010.190	4
4	[MSCI, NFLX]	0.2680	5565.319	12
5	[RTX, VNO]	0.2684	5309.106	4
6	[AMAT, NDAQ]	0.2704	4720.940	7
7	[NDAQ, TFX]	0.2754	5181.135	8
8	[CFG, RCL]	0.2777	4654.054	6
9	[EW, JNPR]	0.2795	4895.442	8
10	[SRE, SBUX]	0.2799	4954.100	8
11	[CVS, KEYS]	0.2802	5482.683	8
12	[EA, TMO]	0.2806	4829.123	6
13	[C, RTX]	0.2811	4879.387	4
14	[APTV, SNA]	0.2812	4820.178	4
15	[FIS, NVR]	0.2821	5782.282	10
16	[GOOG, BLK]	0.2825	5111.338	7
17	[GOOGL, BLK]	0.2828	5114.488	7
18	[SBUX, WMB]	0.2833	5169.595	6
19	[CVS, DISCA]	0.2840	5057.296	5
20	[APH, EMR]	0.2864	5280.866	8

Cuadro 4.18: Resultados del método de Hurst para veinte pares de acciones.

El método de Hurst también ha experimentado una mejora considerable respecto al caso anterior, pero sin alcanzar el rendimiento que presentaba para el caso en el que solo se consideraban cinco pares por método. Al igual que en el método de la distancia, siete de los diez pares de acciones añadidos desde el caso anterior presentan beneficios, siendo bastante representativos para seis de esos pares.

Llaman la atención las parejas en las posiciones dieciséis y diecisiete, $[GOOG, BLK]$ y $[GOOGL, BLK]$. Como consecuencia de lo que se ha llamado en este trabajo *efecto*

Google, es normal que si algún método selecciona una pareja de acciones en las cuales una de ellas es *GOOGL* o *GOOG*, inmediatamente después seleccione la correspondiente pareja pero utilizando esta vez la otra acción perteneciente a *Google*. Dado que las dos acciones pertenecientes a esta compañía se mueven de forma extremadamente parecida (Figura 4.1), si un método considera que una de ellas está relacionada con alguna acción, esa acción estará relacionada de igual manera con la otra acción de *Google*. Esto puede ocurrir para cualquier par de acciones que tenga un movimiento conjunto suficientemente alto, pero entre las acciones utilizadas para la realización de este estudio solo en este caso concreto se da con tanta claridad.

Otra pareja que llama la atención es $[FIS, NVR]$, la cual ha generado un beneficio del 15,65% cerrando un total de cinco posiciones. A continuación se muestra la gráfica de la serie del par frente a sus criterios para vender y comprar el par así como para cerrar las posiciones una vez abiertas. Se puede observar que tan solo se ha vendido el par una vez mientras que se ha comprado en cuatro ocasiones, obteniéndose el grueso de los beneficios en las tres últimas posiciones.

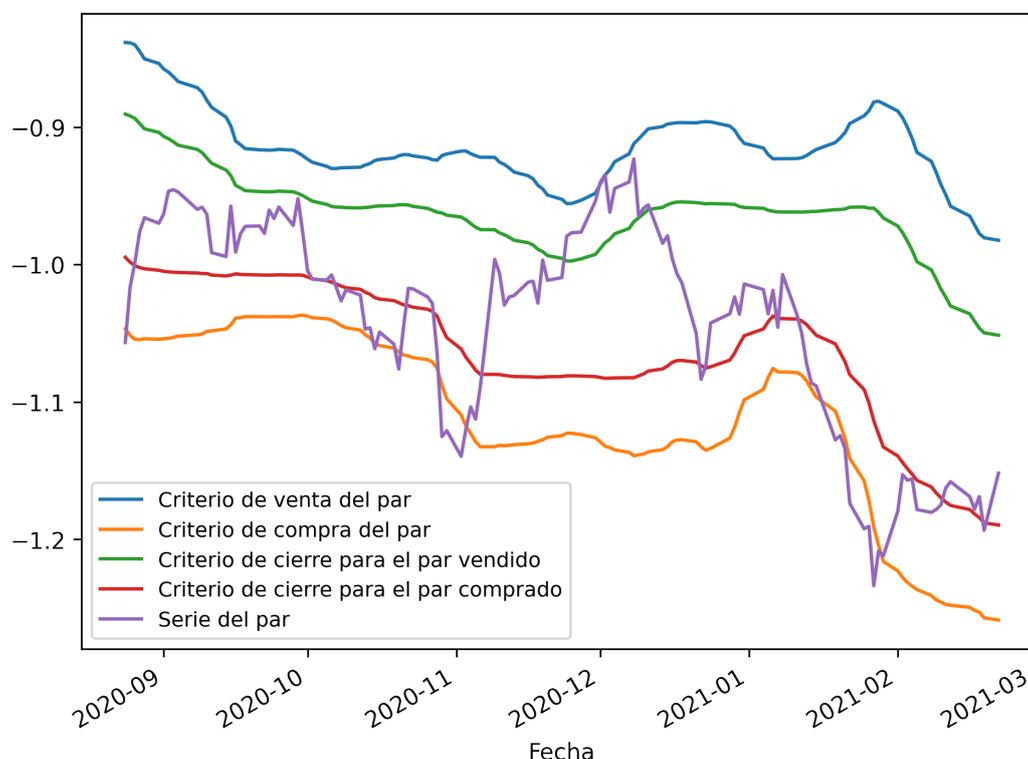


Figura 4.2: Serie del par $[FIS, NVR]$ para el método de Hurst frente a los criterios de apertura y cierre de posición.

El método de la correlación ha mejorado considerablemente respecto al caso anterior, que a su vez ya suponía una mejora respecto al caso de cinco parejas por método. Al igual que los métodos anteriores, siete de las diez parejas añadidas presentan beneficios, siendo los beneficios de dos de esos pares muy altos. Las parejas $[DVN, MRO]$ y $[FRT, SPG]$ han generado unos beneficios del 8,64% y 14,7% respectivamente. Además, la segunda de estas parejas es la única de todo el experimento con la que se han cerrado ocho posiciones completas durante el periodo experimental. En la Figura 4.3

puede verse bastante bien la irregularidad de la serie del par, dibujando movimientos drásticos en el precio de las acciones permitiendo a la estrategia de pair trading cerrar una alta cantidad de posiciones satisfactoriamente.

	Par de acciones	Coefficiente de correlación	Cartera final	Operaciones realizadas
1	[GOOGL, GOOG]	0.999768	5013.832	4
2	[UAA, UA]	0.999588	5006.624	9
3	[FOXA, FOX]	0.997708	5030.435	8
4	[DISCA, DISCK]	0.997354	4918.149	2
5	[NWSA, NWS]	0.996668	5088.286	12
6	[CMA, HBAN]	0.993864	5139.460	6
7	[FRT, REG]	0.993693	5199.503	10
8	[CCL, NCLH]	0.993552	5159.967	7
9	[DFS, SYF]	0.993240	4898.335	8
10	[CFG, FITB]	0.992993	5085.932	8
11	[C, KEY]	0.992711	5051.604	8
12	[FITB, RF]	0.992683	5124.929	10
13	[C, CFG]	0.992184	4729.738	8
14	[RF, TFC]	0.991611	5157.736	12
15	[BAC, JPM]	0.991551	4890.500	6
16	[FITB, TFC]	0.991533	4979.565	8
17	[KEY, RF]	0.991421	5132.426	10
18	[FRT, SPG]	0.990901	5734.723	16
19	[AVB, UDR]	0.990260	5201.056	10
20	[DVN, MRO]	0.990241	5432.028	9

Cuadro 4.19: Resultados del método de correlación para veinte pares de acciones.

También cabe destacar que entre las series añadidas en este método que pasa algo similar a lo que ocurría en el método de Hurst con la transitividad entre las acciones de *Google*, aunque en esta ocasión sin la presencia de esta compañía. Las acciones *FITB*, *RF* y *TFC* están muy relacionadas entre ellas, de forma que se construyen los pares $[FITB, RF]$, $[RF, TFC]$ y $[FITB, TFC]$. En la Figura 4.4 puede verse claramente cómo las series recorren la misma trayectoria en rangos distintos de precio. A pesar de todo esto, en general esta transitividad entre acciones de la que se habla es bastante débil, pues depende de una fuerte relación entre las series que normalmente no es satisfecha. Es muy común que se de el caso en el que una serie *A* esté relacionada con *B* y *C* pero *B* y *C* no estén lo suficientemente relacionada como para formar una pareja con ellas.

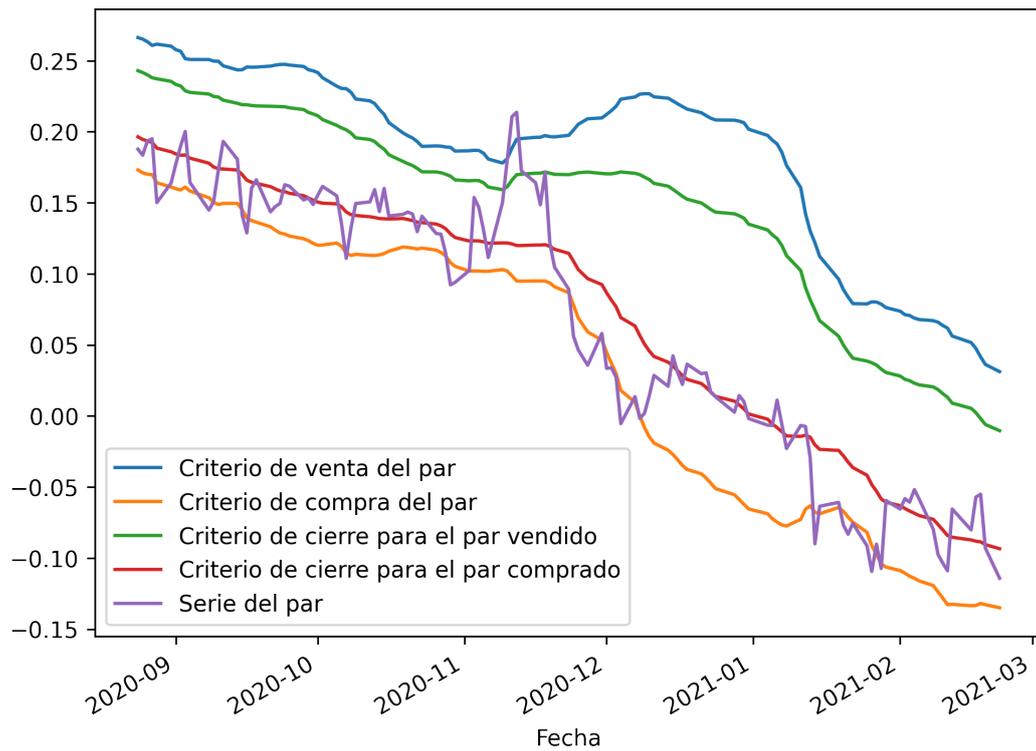


Figura 4.3: Serie del par $[FRT, SPG]$ para el método de la correlación frente a los criterios de apertura y cierre de posición.

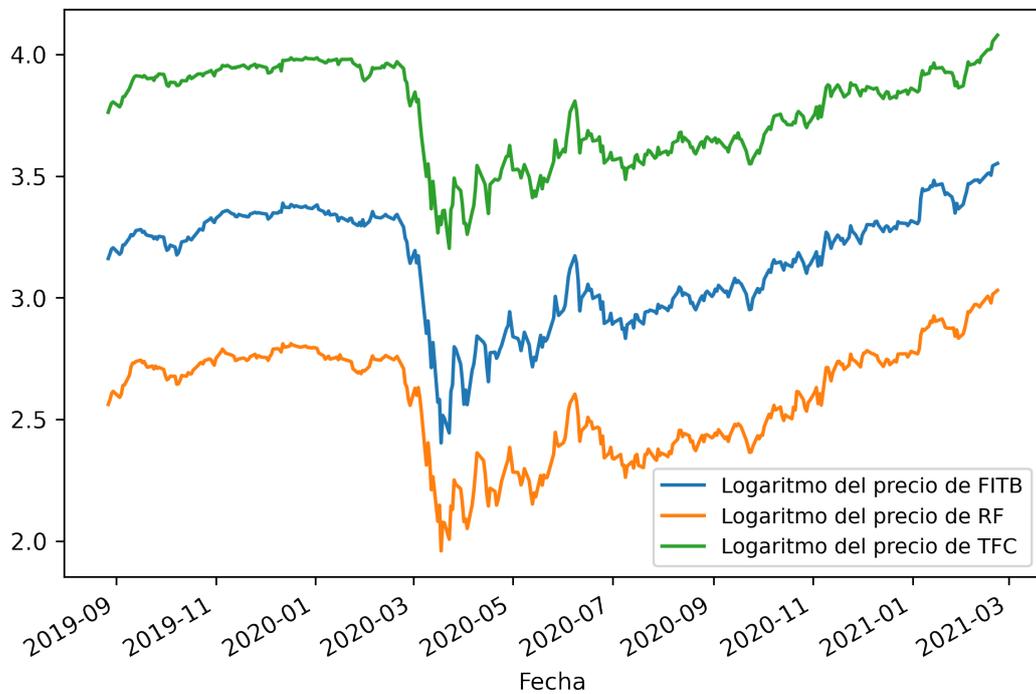


Figura 4.4: Series asociadas a los logaritmos de los precios de las acciones $FITB$, RF y TFC

Por último, el método que se propone derivado de la regresión lineal también ha presentado una mejora respecto al caso anterior del experimento, en el cual ya presentaba una mejora respecto al caso de cinco pares por método. A diferencia de los métodos anteriores, solo dos de las series seleccionadas por este método para completar la veintena presentan pérdidas, mientras que las otras ocho representan beneficios. Si bien es verdad que las pérdidas de esas dos son superiores a las pérdidas de otros pares en los métodos anteriores, los beneficios también son, en general, más elevados a pesar de que no haya ningún par con unos beneficios próximos al quince por ciento como había en los métodos anteriores. Se puede decir que, en general, hay más movimiento de beneficios y pérdidas, representando quizás más riesgo que otros métodos.

Al igual que en el método de la correlación, en este también se encuentra una terna de acciones relacionadas entre ellas que forman las posibles parejas entre ellas. En este caso la terna está formada por *PNC*, *USB* y *BAC* y en este caso sus respectivos pares se presentan juntos en la tabla. Como podemos ver en la Figura 4.5 lo que sucede en este caso es muy parecido a lo que ocurría con las dos acciones de *Google*. En este caso *PNC* y *USB* ocupan el papel de las dos acciones de la compañía y al estar la tercera acción *BAC* relacionada con una de ellas, lo está también con la otra. Es por esto que el coeficiente derivado de la regresión lineal entre *PNC* y *USB* es menor que los de [*BAC*, *USB*] y [*BAC*, *PNC*], los cuales distan muy poco entre ellos.

	Par de acciones	Coeficiente de regresión lineal	Cartera final	Operaciones realizadas
1	[GOOGL, GOOG]	0.008	5013.832	4
2	[FOXA, FOX]	0.024	5030.435	8
3	[DISCA, DISCK]	0.028	4918.149	2
4	[CFG, KEY]	0.037	5260.432	8
5	[NWSA, NWS]	0.051	5088.286	12
6	[UAA, UA]	0.058	5006.624	9
7	[FITB, KEY]	0.059	4954.714	6
8	[CFG, FITB]	0.065	5085.932	8
9	[ESS, UDR]	0.073	5342.996	10
10	[CMS, XEL]	0.077	5095.755	4
11	[PNC, USB]	0.081	5094.595	8
12	[BAC, USB]	0.085	4810.770	4
13	[BAC, PNC]	0.085	5226.918	8
14	[RSG, WM]	0.089	5096.399	6
15	[EQR, UDR]	0.090	5247.588	10
16	[MTB, USB]	0.090	4852.249	6
17	[LNT, CMS]	0.091	5226.757	8
18	[CMA, FITB]	0.095	5424.986	12
19	[COF, FITB]	0.097	5118.256	6
20	[CFG, CMA]	0.098	5098.352	4

Cuadro 4.20: Resultados del método de la regresión lineal para veinte pares de acciones.

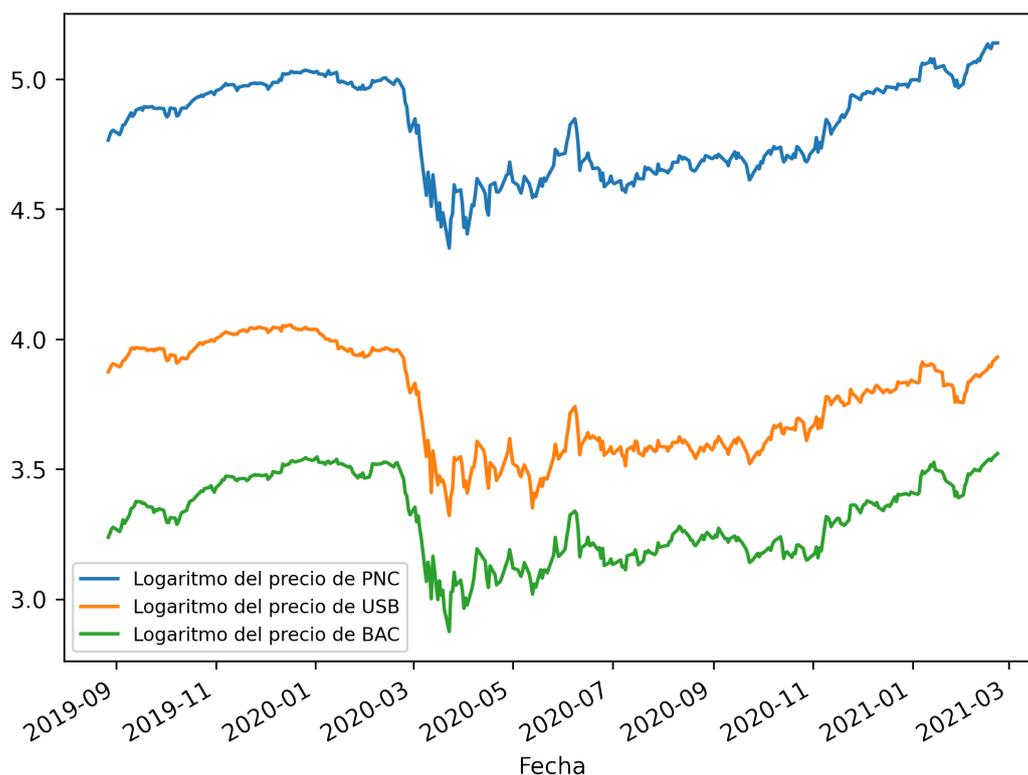


Figura 4.5: Series asociadas a los logaritmos de los precios de las acciones *PNC*, *USB* y *BAC*

A continuación se muestran los resultados del experimento desarrollando la estrategia en veinte pares de acciones para cada método. Se puede ver que el método de Hurst sigue a la cabeza como el método con mejores resultados, pero los métodos de correlación y regresión lineal no difieren mucho en resultados, habiendo mejorado bastante su rendimiento. Por otro lado el método de la distancia ha mejorado sus beneficios, aunque siguen siendo muy bajos en comparación con cualquiera de los otros métodos. Destaca la cantidad de posiciones cerradas en el método de la correlación, el cual seleccionó el par con el que más posiciones se han cerrado en todo el experimento. Se puede deducir que el método de correlación cierra muchas posiciones con poco beneficio mientras que el método de Hurst que es el que menos posiciones ha cerrado lo hace con un beneficio mayor por posición.

Método	Porcentaje de beneficios	Posiciones cerradas
Distancia	0.758 %	72
Hurst	2.291 %	69
Correlación	1.975 %	84
Regresión lineal	1.994 %	71

Cuadro 4.21: Resultados de los cuatro métodos para veinte parejas.

Resultados globales del experimento

Se presentan a continuación los resultados globales para los cuatro métodos una vez finalizado el experimento.

El método de la distancia no ha mostrado buenos resultados para ninguno de los tres casos estudiados. Es cierto que en el caso de los cinco pares simultáneos ha mostrado unos resultados aceptables, pero en los otros dos casos los resultados mostrados han sido bastante decepcionantes. De hecho es muy probable que de haber tenido en cuenta en el modelo las comisiones implícitas en las operaciones, el método de la distancia hubiese presentado pérdidas en el caso en el que se ha desarrollado con diez pares de acciones.

Por otro lado es fácil ver que el método de Hurst ha funcionado mejor con pocas series, bajando considerablemente su rendimiento al pasar de cinco a diez series y recuperando un poco al subir a veinte. Se puede decir que el método del exponente generalizado de Hurst funciona muy bien encontrando las series más eficientes pero que, al haber pocas de estas parejas, cuando se aumenta el número de pares a utilizar rellena con parejas de menor rentabilidad. A pesar de esto ha sido el método más rentable con bastante diferencia, solo en el caso de diez pares ha sido superado ligeramente por el método de la regresión lineal.

Método	Pares utilizados	Porcentaje de beneficios	Posiciones cerradas
Distancia	5	1.156 %	20
	10	0.145 %	34
	20	0.758 %	72
Hurst	5	5.571 %	20
	10	1.594 %	36
	20	2.291 %	69
Correlación	5	0.231 %	17
	10	1.082 %	36
	20	1.975 %	84
Regresión lineal	5	1.246 %	17
	10	1.596 %	35
	20	1.994 %	71

Cuadro 4.22: Resultados globales del experimento para los cuatro métodos propuestos.

El método de la correlación se proponía como algo sencillo pues es la más directa aplicación del concepto de correlación. Los resultados mostrados no están a la altura de los del método de Hurst pero tampoco han sido demasiado malos. Es cierto que este método ha funcionado francamente mal en presencia de únicamente cinco series, pero en cuanto el conjunto de pares ha aumentado, el rendimiento del método ha aumentado considerablemente. Esto se puede deber a la implementación tan directa que se ha hecho de la correlación, y es que si se eligen primero los pares con mayor correlación, estos no van a presentar prácticamente oportunidades donde la estrategia de pair trading puede brillar, mientras que al aumentar la cantidad de pares con los que trabajar

la correlación media entre las parejas será más baja, presentando mas oportunidades para que la estrategia funcione bien.

Queda por último examinar el nuevo método propuesto en la realización de este trabajo ya que, mientras el método de la correlación, a pesar de no haber sido extraído directamente de trabajos previos, no es más que la implementación directa de un concepto muy recurrente en todos los trabajos relacionados con las estrategias de pair trading. El método del coeficiente derivado a partir de la regresión lineal sí es presentado como una nueva forma de encontrar pares de acciones con un alto grado de movimiento conjunto. Este método ha mostrado un rendimiento bastante bueno, aumentando progresivamente sus beneficios en función de la cantidad de pares con los que estaba trabajando y llegando incluso a presentar un mejor funcionamiento que el método del exponente generalizado de Hurst en el caso de diez pares simultáneos.

En general es claro que el método de la distancia, a pesar de ser históricamente muy popular, deja bastante que desear con respecto a los otros métodos, siendo el método de Hurst el claro ganador en este experimento con un rendimiento muy superior al del resto de métodos a excepción del caso en el que se ha trabajado con diez pares, en el que la ventaja del método de la regresión lineal tampoco ha sido muy alta. Este último método también ha funcionado bastante bien, superando tanto al método de la correlación como al de la distancia.

Conclusiones

Durante este trabajo se han implementado cuatro métodos para el estudio de pares de series temporales orientadas a las estrategias del tipo pair trading, realizando una revisión bibliográfica y proponiendo un nuevo enfoque.

Todo el trabajo ha sido realizado con datos reales extraídos directamente de Yahoo finanzas [2], dando la posibilidad de estudiar y analizar las principales características de series temporales reales. De esta forma se ha entrado directamente en el mundo de la información real, aplicando conceptos teóricos y conociendo la dificultad de adaptarlos a este tipo de datos. Esto se puede ver claramente en los métodos de la distancia y del exponente de Hurst generalizado.

En el primer método, a excepción de la pareja perteneciente a *Google*, no se ha encontrado ninguna pareja de series con una distancia entre ellas menor a dos unidades. Incluso la pareja de *Google* presenta una distancia de casi 0,8 unidades a pesar de la clara similitud entre sus series (Figura 4.1).

En el método de Hurst, a pesar de que la teoría dicta que H puede tomar valores entre 0 y 1, no se ha encontrado ninguna pareja con un valor para este exponente menor a 0,25.

Si bien se ha podido llevar a cabo satisfactoriamente el objetivo planteado inicialmente, implementar una serie de métodos para una estrategia de pair trading y comparar su eficacia, los resultados obtenidos deben de ser entendidos únicamente como una primera aproximación. Para probar su consistencia se requiere expandir este estudio a otros periodos temporales, otros mercados e incluso diferentes pruebas.

A pesar de lo mencionado anteriormente, estudiando los resultados obtenidos sorprende el comportamiento del método propuesto basado en la regresión lineal, pues siendo un enfoque que no se ha desarrollado en gran medida ha funcionado francamente bien en comparación con otros métodos mucho más evolucionados. Una posible vía de desarrollo de los métodos implementados es introducir ciertos parámetros en la definición de la serie del par, como se hace en el método del exponente de Hurst generalizado, acercando más la serie del par a los conceptos teóricos en los que se basa cada método.

Otra forma de mejorar el experimento es aplicando un modelo más elaborado, introduciendo variables como las comisiones implícitas en cada operación de compra o venta de acciones. Probablemente en estas condiciones más próximas a la realidad se encuentren parejas de acciones con las que, a pesar de tener un alto grado de movimiento conjunto, no sea rentable desarrollar la estrategia generando beneficios insignificantes o incluso pérdidas. Para evitar esto habría que introducir algún tipo de criterio de riesgo en la selección de pares, eligiendo las mejores parejas de acciones para cada método pero imponiendo que su movimiento conjunto no sea demasiado alto.

En definitiva, este estudio ha resultado satisfactorio como un primer acercamiento a las estrategias de pair trading y las series temporales, permitiendo obtener algunas conclusiones y quedando abierto para una potencial continuación.

Bibliografía

- [1] Richard Bookstaber. *A Demon of Our Own Design: Markets, Hedge Funds, and the Perils of Financial Innovation*. John Wiley & Sons, 2007.
- [2] Yahoo finanzas. URL: <https://es.finance.yahoo.com/>.
- [3] Wikipedia SP500. URL: https://en.wikipedia.org/wiki/List_of_S%26P_500_companies.
- [4] Lenguaje de programación Python v.3.7.10. URL: <https://www.python.org/>.
- [5] Biblioteca de Python Numpy v.1.19.5. URL: <https://numpy.org/>.
- [6] Biblioteca de Python Matplotlib v.3.2.2. URL: <https://matplotlib.org/>.
- [7] Matlab. URL: <https://es.mathworks.com/products/matlab.html>.
- [8] Biblioteca de Python SciPy v.1.4.1. URL: <https://www.scipy.org/>.
- [9] Biblioteca de Python Pandas v.1.1.5. URL: <https://pandas.pydata.org/>.
- [10] Biblioteca de Python fbm v.0.3.0. URL: <https://pypi.org/project/fbm/>.
- [11] Biblioteca de Python tqdm v.4.41.1. URL: <https://pypi.org/project/tqdm/>.
- [12] M Fernández-Martínez y col. “A note on power-law cross-correlated processes”. En: *Chaos, Solitons & Fractals* 138 (2020), pág. 109914.
- [13] William Goetzmann, K Geert Rouwenhorst y col. *Pairs trading: Performance of a relative value arbitrage rule*. Inf. téc. Yale School of Management, 1998.
- [14] Evan Gatev, William N Goetzmann y K Geert Rouwenhorst. “Pairs trading: Performance of a relative-value arbitrage rule”. En: *The Review of Financial Studies* 19.3 (2006), págs. 797-827.
- [15] José Pedro Ramos-Requena, JE Trinidad-Segovia y MA Sánchez-Granero. “Introducing Hurst exponent in pair trading”. En: *Physica A: statistical mechanics and its applications* 488 (2017), págs. 39-45.
- [16] Harold Edwin Hurst. “Long-term storage capacity of reservoirs”. En: *Transactions of the American society of civil engineers* 116.1 (1951), págs. 770-799.
- [17] Michael Hochman. “Dimension theory of self-similar sets and measures”. En: *Proceedings of the International Congress of Mathematicians*. Vol. 2. World Scientific. 2018, págs. 1943-1966.
- [18] Jan Beran y col. *Long-Memory Processes*. Springer, 2016.
- [19] Douglas S Ehrman. *The handbook of pairs trading: strategies using equities, options, and futures*. Vol. 240. John Wiley & Sons, 2006.