



**Universitat
Pompeu Fabra
Barcelona**

**Facultad
de Economía
y Empresa**

Curso Académico: 2024 - 2025

Trabajo de Fin de Grado

Valoración relativa de empresas basada en análisis fundamental

Ayleen Shantal Abreu Matos (Grado en Administración y Dirección de Empresas)

Raquel Bas Saumoy (Grado en Administración y Dirección de Empresas)

Lucas Der Parsehian Bueno (Grado en Economía)

Noel Monzón Blasco (Grado en Economía)

Tutor: Hector Omar Gallegos Gonzalez

RESUMEN

El trabajo de fin de grado aborda la problemática de la valoración cuantitativa de acciones en un contexto real de mercado. El objetivo es generar un modelo a partir de Econometría y Machine Learning para poder evaluar el precio ideal de una empresa basándose en su análisis fundamental relativo a su sector. La idea detrás de esto es que el valor intrínseco de una empresa está dado por las características financieras en comparación a sus competidores, aislando efectos macroeconómicos y de mercado.

El modelo permite clasificar las desviaciones del precio de una acción respecto a su valor ideal para poder estudiar magnitudes, períodos y patrones de comportamiento. Además, en el modelo se genera un índice que mide la salud financiera de las empresas a lo largo del tiempo para contrastarlo con el rendimiento del precio de la acción.

Este modelo puede ser utilizado por inversores o fondos de inversión como herramienta de análisis fundamental para aquellos que creen que el valor de una empresa está dado por sus características financieras. Su estudio potencial permite realizar evaluaciones de reversión de la media, análisis de diferencias en diferencias y pruebas de portafolios para diferentes estrategias.

PALABRAS CLAVE

- **Análisis fundamental:** evaluación basada en los estados financieros, ratios y otros factores internos de la empresa.
- **Índice ponderado:** suma ponderada entre pesos y variables financieras.
- **Variables control:** variables que reflejan el ruido del mercado.
- **Valoración relativa:** método que compara empresas dentro de un mismo sector para identificar si están sobre o infravaloradas según sus variables financieras.
- **Valor intrínseco:** estimación del valor real de una empresa, independiente del precio de mercado. En el trabajo, fue basado en sus métricas financieras.
- **Residuos:** desviación entre la valoración intrínseca y la de mercado.
- **Precio ideal:** Serie acumulada de precio de mercado corregida por sus residuos.

ÍNDICE

Introducción.....	1
Objetivo del trabajo y motivación.....	2
Marco conceptual.....	2
Origen del contenido y de los datos.....	3
Metodología, desarrollo y aplicación.....	4
Transformación de los datos y creación del índice.....	5
Selección de variables.....	6
Regresión ols inicial.....	7
Algoritmo bayesiano.....	8
Segunda regresión ols: valoración.....	11
Resultados.....	12
Resultados del modelo.....	12
Resultados empíricos.....	14
Decisiones y potenciales mejoras.....	22
Conclusiones.....	23
Bibliografía.....	
Anexos.....	A1

INTRODUCCIÓN

La correcta valoración de activos financieros supone el foco principal para la toma de decisiones de inversión y planificación de estrategias.

El precio que le asigna el mercado a un activo suele estar sesgado por diversos factores externos. En consecuencia, se propuso desarrollar una herramienta que sea capaz de encontrar estos desajustes entre el precio mercado y el valor ideal (basado en las métricas financieras). En la actualidad se utilizan diferentes enfoques como descuentos de flujos de caja o valoración por múltiplos. Estos métodos no toman en cuenta el efecto del mercado en la valoración, utilizan estimaciones subjetivas y están muy orientadas a flujo y ganancias, por lo que no toman a la empresa como un conjunto. Según el punto de vista mostrado en el trabajo, compartido con varios analistas y economistas, el valor intrínseco de la empresa viene determinado por el diferencial de la salud financiera respecto a su sector. La idea general del trabajo es que existe un arbitraje provocado por factores externos que tiende a corregirse en el largo plazo, reflejando el valor de una empresa dadas sus características internas relativas a sus competidores.

Hay que aclarar que el enfoque del trabajo no es generalizable a todo el mercado, pero que el mismo estudio que se lleva a cabo para la muestra utilizada podría expandirse a otras cambiando los datos y variables control. El objeto de estudio fue el sector financiero estadounidense desde enero del año 2020 a diciembre de 2024. En concreto, se seleccionaron 18 empresas ([ver anexo](#)) que abarcan distintas áreas del sector, como banca tradicional, aseguradoras, servicios de inversión y medios de pago. Sobre estos datos se seleccionaron variables basándose en criterios similares a “Valuation Approaches and Metrics: A Survey of the Theory and Evidence” de Damodaran. Con estas variables realizamos una serie de regresiones y se implementó un algoritmo Bayesiano. El resultado es la construcción de un índice de salud financiera y un modelo que genera una serie de precios ideales derivados del valor intrínseco.

El trabajo se centra en explicar cómo se ha conseguido llegar al modelo y los resultados obtenidos. Este inicia aclarando de dónde se ha extraído la información para la base de datos y las teorías en las que se ha basado la tesis del trabajo. Más adelante, se explica el proceso del modelo y se muestran los resultados y las conclusiones extraídas con una breve reflexión.

OBJETIVO DEL TRABAJO Y MOTIVACIÓN

El mercado financiero es una caja negra en la cual los resultados que se observan muchas veces no tienen una explicación exacta. Los movimientos a corto plazo suelen estar dominados por noticias, sentimientos y tendencias de mercado. En cambio, en el largo plazo podemos intuir que hay conductores que hacen que determinadas empresas con determinadas características prevalezcan por encima de sus competidores. Nuestra motivación se basa en esto, en creer que gracias a las herramientas con las que se cuenta hoy en día podemos acercarnos a encontrar de forma más precisa estos conductores. Después de haber estudiado diferentes teorías y modelos financieros, haber adquirido herramientas econométricas y haber comprendido los componentes macroeconómicos, nos vimos capacitados para poder incursionarnos en este desafío.

MARCO CONCEPTUAL

El trabajo se basa en estudios y enfoques teóricos para la construcción del modelo. Diferentes analistas y académicos han abordado la valoración de empresas de distintas formas.

Benjamin Graham y David Dodd son unos de los principales defensores del análisis fundamental. Son coautores del libro “Security Analysis”, donde justificaban la idea de que el verdadero valor de una acción no tenía porqué ser igual al valor del mercado. Defendían que el valor intrínseco de la acción se justificaba mediante fundamentos financieros como los activos, ganancias y dividendos que la misma generaba. El objetivo es que una vez localizada alguna empresa con menor valor intrínseco, es decir, infravalorada por el mercado, es conveniente invertir a largo plazo en la empresa ya que el mercado corregirá la desviación. A su vez, crearon el concepto “margen de seguridad”, que se refería a la diferencia entre el precio de mercado actual de una acción y el valor intrínseco estudiado. Esta disparidad supone un nivel de protección para el accionista en caso de un posible error de cálculo o por circunstancias inesperadas en el mercado.

Entre los fieles seguidores de la teoría de Graham y Dodd, se encuentra el inversor Warren Buffet. El libro “The Warren Buffett Way” escrito por Robert G. Hagstrom expone que el valor de la empresa se encuentra en el rendimiento de la acción (ROE), en la relación entre la deuda y el capital, el margen de ganancias y en la ventaja competitiva del modelo de negocio.

En 2007, Damodaran publicó el artículo “Valuation Approaches and Metrics: A Survey of the Theory and Evidence” en The Journal of Finance. En él, se ofrece una revisión de los métodos de valoración fundamental: Explica que el valor intrínseco de una empresa debe valorarse según el flujo de caja descontado, ROE, ROA, deuda y márgenes operativos. También, Damodaran considera que el entorno es relevante a la hora de estudiar el valor de una empresa. Este considera que es importante tener en cuenta el contexto macroeconómico principalmente , ya que influye directamente en el rendimiento de una empresa.

Estos modelos justifican la construcción del índice utilizado en el trabajo, con variables fundamentales que ellos usan y consideran imprescindibles para el análisis (EPS, ROE, FCFMG...). A su vez, se incorporan variables macroeconómicas siguiendo la teoría de Damodaran ([Ver Tabla A1](#)).

ORIGEN DEL CONTENIDO Y DE LOS DATOS

Para poder desarrollar el modelo fue necesario recopilar una amplia variedad de información. Como punto de partida, se obtuvieron las ideas procedentes de diferentes estudios previos como el artículo “Valuation Approaches and Metrics: A Survey of the Theory and Evidence” de Damodaran, la teoría del valor intrínseco de Graham y Dodd, y algunas teorías de grandes inversionistas como Warren Buffet, que sirvieron de base para la realización del trabajo y orientaron la construcción de la herramienta de análisis.

Respecto a la base de datos, se partió del objetivo de basarnos en fuentes fiables y con cobertura histórica suficiente para el estudio. A pesar de eso, tuvimos diversas dificultades debido a que la mayoría de fuentes confiables tienen limitaciones de acceso. Se acabó encontrando Finchat, la cual posee herramientas muy interesantes. A pesar de esto, se comprobó que tenía inconsistencias en el manejo de algunos datos, razón por la cual se buscó otra fuente para complementarlos.

Con esta búsqueda, se encontró Stock Analysis, la cual, juntamente con Finchat, fue utilizada para registrar todos los datos fundamentales de las 18 empresas financieras previamente seleccionadas. Con la finalidad de mantener una base de datos formada con valores reales, se comprobó que los datos obtenidos de Finchat y Stock Analysis fueran correctos contrastándolos con los datos de la SEC (Securities and Exchange Commission) y los 10-Q de las empresas.

Para el precio histórico diario de las acciones analizadas de todas las empresas se utilizaron datos de [Investing.com](#), que fueron contrastados con los de Yahoo Finance para garantizar su consistencia. Adicionalmente, se consiguieron los gráficos de Target price gracias a [Marketscreener.com](#).

Desde [Macrotrends.net](#) y FRED (Federal Reserve Economic Data) se pudieron obtener los datos macroeconómicos utilizados en el modelo. Estas otorgaron información trimestral de la inflación, tipo de interés, crecimiento del PIB real, tipos de cambio y diferenciales de crédito (tanto real como esperado).

En cuanto a los datos financieros, Yahoo Finance fue la responsable de facilitar el histórico de índices relevantes como XLF (índice del sector financiero), VIX (volatilidad implícita del mercado americano) y SPY (S&P 500). También se incluyó el índice NFCI (índice de condiciones financieras nacionales) obtenido desde FRED aunque luego no fue utilizado.

Finalmente seleccionamos las “Feature variables”: *Volume* (volumen de negociación de mercado promedio trimestral desde Yahoo Finance) y *shares* (cambio neto de acciones que poseen los fondos de inversión sobre la empresa desde Dataroma). A continuación, con toda la información obtenida, se generó una base de datos en Excel que fue utilizada para la realización del modelo.

METODOLOGÍA, DESARROLLO Y APLICACIÓN

El trabajo tiene como objetivo desarrollar una serie de regresiones complementadas con un algoritmo bayesiano. Esto permite diseñar un modelo para conocer qué empresas presentan acciones sobrevaloradas e infravaloradas en un sector específico. Se seleccionaron 18 empresas estadounidenses del sector financiero elegidas por su homogeneidad, capitalización, volumen de negociación de mercado, transparencia de datos y alta correlación entre rendimientos ([ver figura A1](#)). Esto llevó a la creación de dos bases de datos relevantes para el desarrollo del modelo: Una con variables fundamentales y otra con variables control (Macroeconómicas, de sentimiento del mercado y *Smart Money (inversores institucionales)*). Damodaran utiliza el enfoque de DCF para entender cómo se evalúa una compañía y, a partir de esto, selecciona los múltiplos relacionados. En el trabajo se plantea algo similar, se hace una selección de los aspectos para comparar una empresa con sus competidores (Rentabilidad, Financiamiento y Eficacia) y se seleccionan las métricas financieras que demuestren esto para

cada empresa. El análisis comprende el período que comienza con el primer trimestre de 2020 hasta el último de 2024. En este, ocurren eventos relevantes como la pandemia de COVID-19 y su recuperación, inflación fluctuante, cambios en los tipos de interés y tensiones geopolíticas, todos con una fuerte relevancia en el sector. Una vez recopilada toda la información y organizada en formato trimestral, se utilizó un código en Python ([ver anexo](#)) para construir el modelo. La intención era detectar infravaloración o sobrevaloración relativa por período mediante un índice financiero denominado INDEX2. Este índice fue desarrollado con técnicas econométricas y machine learning. El desarrollo del modelo se dividió en tres etapas: transformación de datos, construcción del índice y aplicación del modelo de regresión y optimización , incluyendo medidas de validación como R cuadrado, MSE, test de Durbin y análisis de residuos. A continuación se pasa a explicar cada proceso de forma profunda.

TRANSFORMACIÓN DE LOS DATOS Y CREACIÓN DEL ÍNDICE

La primera fase del código consistió en la carga y preparación de la base de datos ([Ver Tabla A1](#)). Como primer paso se optó por combinar ambos conjuntos de datos en excel para generar un único archivo mediante la unión de las columnas TICKER (nombre de la empresa) y DATE (fecha de captura de datos) que se convirtieron en variables categóricas y se combinaron para asegurar una secuencia temporal acorde al tiempo.

Se prosiguió con la creación de la variable del crecimiento porcentual del precio medio (calculado como el cambio porcentual entre la media de un trimestre y su siguiente). Se utilizó el precio medio, ya que tomar precio inicial y final podría generar una dependencia muy fuerte a un período corto de tiempo, lo que podría generar un sesgo. A continuación, se realizó una estandarización de las variables fundamentales mediante el método Min-Max. Se eligió este tipo de estandarización ya que permite evaluar las empresas del 0 al 1 (como un puntaje) y es ideal para formar el índice.

Se realizó un primer cálculo de la correlación de las variables contra el crecimiento del precio medio ([ver figura A2](#)) para empezar a estudiar la magnitud del impacto y la dirección del efecto (positivo o negativo). Según las correlaciones, se puede ver cómo las variables de control están más ligadas al cambio en el precio que las fundamentales. Esto justifica la creación de un índice que junte el poder predictivo de los fundamentales sobre el cambio en el precio.

A continuación, se definió un diccionario de pesos y se estudió la matriz de correlación ([ver Tabla A2](#)) para todas las variables fundamentales. Como resultado, se obtuvo un índice

ponderado para cada empresa de forma trimestral basado en los pesos asignados. Estos pesos iniciales no son relevantes para los resultados. Se asignaron valores que parecían razonables para luego poder optimizarlos. Los valores asignados fueron: 0,2 para *Cash, EPS, FCFMG, Quick, Employees y Revenues*; 0,25 para *NETPROFITMG, OPERATINGMG*; 0,15 para *ROA y ROE*; y -1 para *Debt*.

Con estos pesos, el índice mostró una correlación contra el cambio en el precio de -0,0669 (muy baja y de signo negativo). Más adelante, se presenta una optimización de los pesos para mejorar el resultado.

Por otro lado, las variables macroeconómicas se estandarizaron mediante diferencias logarítmicas. Los indicadores y otras características también se estandarizaron con diferencias logarítmicas, excepto la variable de acciones en circulación (*Shares*). La particularidad de esta variable frente al resto es que primero se estandariza mediante diferencias logarítmicas, pero posteriormente se aplica una segunda estandarización utilizando el método Min-Max. Toda selección de método de estandarización fue estudiada según los valores de cada variable, su distribución y su impacto en la correlación, VIF y Lasso. Finalmente, se volvieron a calcular las correlaciones ([ver figura A3](#)).

Debido al cambio porcentual de estandarizar con el método de diferencias logarítmicas, la base de datos elimina el primer valor a falta de los datos del período anterior. Así pues, el código procede a eliminar las primeras filas de la base de datos, además de las filas que contengan celdas vacías.

SELECCIÓN DE VARIABLES

Una vez cargados y procesados los datos, el siguiente paso consiste en la selección de variables en función de las correlaciones, el proceso de estandarización y métodos econométricos.

A partir de aquí, se calculó el Factor de Inflación de la Varianza (VIF) entre las variables predictoras y cambio en el precio y se eliminaron las que generaban multicolinealidad. Esto resultó en el descarte de las variables del crecimiento del PIB real y la tasa de desempleo. Después de haber excluido estas variables, se volvió a recalcular el VIF y todas reflejaron tener un VIF menor a 4 ([ver figura A4](#)).

A continuación, se utilizó el método Lasso para seleccionar las variables más relevantes del modelo y penalizar aquellas con menos poder explicativo para reducir el riesgo de sobreajuste, pero ninguna variable fue eliminada en el proceso.

REGRESIÓN OLS INICIAL

Después de haber obtenido las variables óptimas, el siguiente paso fue estimar un modelo de regresión con errores robustos con crecimiento porcentual del precio (PRICE_Growth) como variable dependiente y el conjunto de predictores seleccionados como variables explicativas. Es importante destacar en todo momento que la intención del trabajo no es predecir el cambio en el precio, pero sí lo es ajustar el índice de forma que más se adecúe a la realidad. Así pues, esta regresión inicial se utiliza para eliminar el ruido que puede afectar al estimar la relación entre el índice y el cambio en el precio. Las estadísticas como R cuadrado y MSE son importantes para ver un buen ajuste, pero no necesariamente determinantes, ya que la finalidad del modelo no es predecir.

Este modelo fue la base de partida para ver qué se debía mejorar. Los resultados alcanzados fueron:

- R^2 : 0,427.
- Coeficientes significativos con signo apropiado:
 - Tasas de interés
 - Tipo de cambio
 - *Spreads de crédito* (diferenciales de crédito)
 - SPY
- Índice: coeficiente de 0.0132 no significativo
- Durbin Watson: 1.8 (con 2 como valor ideal)
- Error cuadrático medio // *Mean Squared Error* (MSE): 0.0075

Con estos resultados, se puede ver cómo se obtuvo un R cuadrado dentro del rango para modelos financieros, con un Durbin Watson aceptable y una media de residuos de cero, aunque el test de normalidad es negativo. Los resultados muestran mucho margen de mejora y

por eso se ajustó el algoritmo para optimizar esta relación entre los fundamentales y el crecimiento en el precio.

ALGORITMO BAYESIANO

Los algoritmos bayesianos operan bajo el principio del Teorema de Bayes, integrando conocimientos previos con nueva evidencia para refinar las estimaciones de manera eficiente. En problemas de optimización, utilizan un modelo probabilístico (típicamente un proceso gaussiano) para seleccionar los valores más prometedores de los parámetros, evitando la necesidad de evaluar exhaustivamente todas las combinaciones posibles.

Así pues, el algoritmo inicia optimizando los pesos del índice inicial pero también generando otras 40 combinaciones aleatorias como base. Para mantener la coherencia financiera en el índice, se delimitaron los pesos para que DEBT solo pudiera ir de -1 a 0 y las demás de 0 a 1. El algoritmo itera sobre la Regresión OLS inicial probando diferentes pesos, para posteriormente reconstruir el índice y finalizar generando la regresión de nuevo. La métrica a optimizar es:

```
metric = alpha * mse_score + (1 - alpha) * correlation - penalty - l2_penalty
```

“MSE score” es una métrica inversa del MSE de la regresión inicial. Es decir, cuanto más grande es, menor es el MSE, y en consecuencia, mejor el ajuste. “Correlation” es la correlación entre el índice y el cambio en el precio. “Alpha” es un parámetro que pondera a qué se le da más importancia.

La idea es maximizar la correlación entre el índice y el *price growth* y, a su vez, maximizar la relación entre índice y cambio en el precio. Es decir, obtener un índice final que se ajuste bien en una regresión que predice el precio pero que también por su cuenta tenga una relación positiva con el cambio del precio. De este modo, se consigue un índice que se ajuste a la realidad y que también contiene relación con el crecimiento del precio de la acción para el siguiente trimestre. El índice no predice precio por su cuenta, pero la idea es que los pesos contengan la importancia de las variables con su efecto limpio del ruido de mercado (esto se logra con una regresión que ajuste bien precio con variables de control).

Se acabó estableciendo Alpha en 0,25 ya que el cambio marginal en el MSE era mucho menor comparado al cambio marginal en la parte de correlación (se realizaron diferentes pruebas para corroborarlo). A su vez, se agregó una penalización si los pesos no se encuentran dentro del rango establecido y si son muy parecidos entre sí para obtener resultados más interpretables.

Se probaron diferentes puntos iniciales, número de iteraciones, alphas y límites de los pesos para el índice ([Ver tabla A3](#)). Se decidió la mejor combinación basada en tres criterios: Relación de INDEX2 con el crecimiento del precio, interpretabilidad del INDEX2 y análisis de residuos de valoración.

Finalmente, se generaron 40 pesos iniciales aleatorios y se realizaron 400 iteraciones de los pesos (se probaron diferentes combinaciones y esta es la más eficiente a nivel tiempo de ejecución y resultados). Los pesos optimizados obtenidos fueron los siguientes:

Tabla 1. Pesos Optimizados del Índice Fundamental y correlación con Índice.

Variable	Pesos	Corr con Índice
CASH	0,225	0,00
DEBT	-0,191	-0,66
EMPLOYEES	0,009	0,30
EPS	0,009	0,25
FCFMG	0,009	0,49
NETPROFITMG	0,009	0,57
OPERATINGMG	0,009	0,18
QUICK	0,859	0,52
REVENUES	0,046	0,35
ROA	0,009	0,16
ROE	0,009	0,32

Tal y como se muestra en la tabla, las variables Debt, Cash, Quick ratio y Revenues fueron las más significativas. A pesar de su alto peso, la correlación entre Quick e Índice no destaca en exceso por sobre las demás (lo cual es importante ya que no opaca a las demás variables). Se volvió a realizar la regresión y a evaluar las estadísticas:

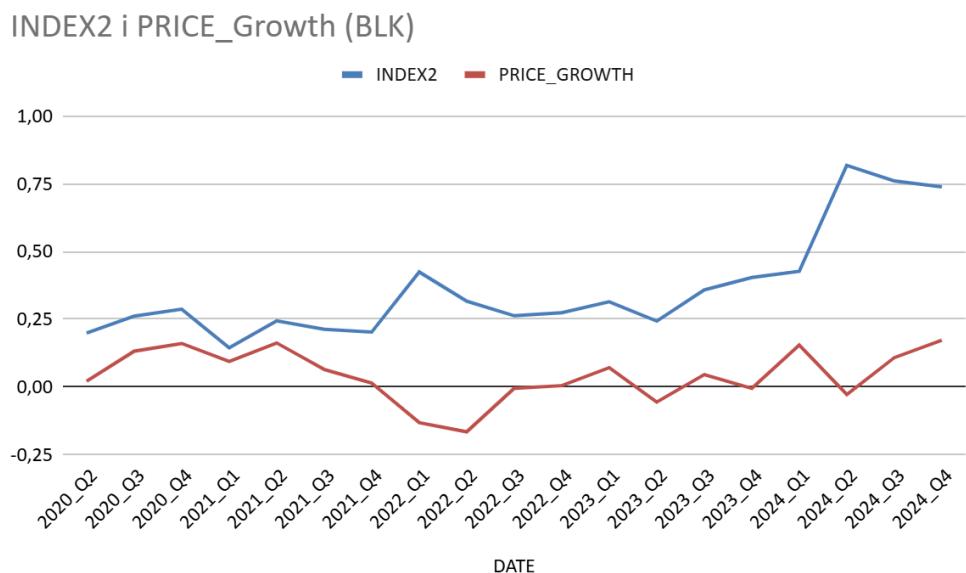
- R^2 : 0,427 (se mantiene)
- Significativos:
 - Tasas de interés
 - Tipo de cambio
 - Spread de crédito
 - SPY
- Coeficiente de índice: 0.0191 (aumenta 44%) aunque sigue sin ser significativo

- MSE: 0.00744501 (disminuye)
- Correlación (índice y cambio en el precio): 0.1522 (Pasa a ser positiva y mejora mucho)
- Validación-cruzada MSE (Media) // *Cross-Validated MSE (Mean)*: 0.0077482
- Validación-cruzada MSE (Desviación estándar) : 0.0023037
- Test MSE: 0.0063485
- Test RMSE: 0.0796777
- Test MAE: 0.0644130
- Los VIF son prácticamente iguales

Se puede observar cómo los conductores más grandes siguen siendo las variables macroeconómicas, pero la mejora tanto en la dirección como en la influencia en la predicción pasan a ser mucho mejores luego de la optimización.

Podemos ver cómo en algunos gráficos, como por ejemplo, en el caso de BLK, se observa una relación marcada, a pesar de que esto no es así para todas las empresas ([ver a anexo partir de Figura A7](#)).

Figura 1. Relación entre INDEX2 y Price_Growth para BlackRock.



Es importante recordar que el índice utilizado no busca predecir el cambio en el precio de forma directa, sino que solamente optimizarlo de tal forma que la relación sea lo más interpretable posible.

SEGUNDA REGRESIÓN OLS: VALORACIÓN

Esta segunda regresión es entre Índice 2 y cambio en el precio. La idea es ajustar cómo la empresa media del sector financiero tendría que haber rendido teóricamente solamente según sus variables fundamentales (sin intención predictiva). A través de los residuos, se intentó analizar si existía una infravaloración o sobrevaloración (siempre relativa a sus competidores) persistente en alguna empresa. Y, posteriormente, si se encontraba alguna forma de comprobar que esto se revirtiera. Cabe destacar que el período de cinco años puede no ser suficiente, ya que el cambio puede resultar más evidente o más significativo en el largo plazo. Para comprobar que la regresión fuera coherente, se miró que la relación entre el índice y el cambio en el precio fuera positiva y que los residuos fueran lo más interpretables posibles. A continuación, se adjuntan los resultados:

1. Coeficiente: 0.0505 (significativo y mayor a 0.013 de antes)
2. Durbin-Watson: 1.676
 - a. Media de residuos: 0.0000000 (Ideal: cercano a 0)
 - b. Desviación estándar de los residuos: 0.1134818
 - c. Asimetría de los residuos: -0.1392 (Ideal: cercano a 0)
 - d. Curtosis de residuos: 1.3311 (Ideal: cercano a 3 para distribución normal)
 - e. Estadístico Durbin-Watson: 1.6759 (Ideal: cercano a 2, indica no autocorrelación)
 - f. Distancia máxima de Cook': 0.0342243 (Los valores > 1 sugieren puntos influyentes)
 - g. Valor máximo apalancamiento (Valor máximo de apalancamiento): 0.0327459
(Valores altos indican potenciales valores atípicos)

Se obtuvo un análisis de residuos dentro del rango aceptado (aunque no siguen una distribución normal).

RESULTADOS

Los resultados pueden ser divididos y analizados en dos partes:

1. Resultados del modelo (eficacia).
2. Resultados empíricos (interpretación).

RESULTADOS DEL MODELO

Los resultados del modelo fueron:

- Baja multicolinealidad.
- Variables influyentes comprobadas con Lasso.
- Signo positivo de Index 2 en ambas regresiones.
- Coeficiente de Index 2 significativo en la segunda regresión.
- Significancia estadística y de variables de control.
- Mejora del modelo con Index 2 (menor MSE, ajuste y coeficientes).
- Mejora de correlación de Index 2.
- R cuadrado de 42% (alto para un modelo explicativo del sector financiero).
- El coeficiente de Index 2 en la primera regresión no es significativo (mejorable).

Tabla 2. *Resultados de eficiencia del modelo.*

Métrica	Variable / Resultado	Valor
Correlación con PRICE_Growth	INDEX	-0.066980
	Interest rates	-0.207161
	Real GDP Growth	0.392984
	Exchange Rates	-0.155098
	Unemployment rate	-0.368246
	VIX	-0.241274
	XLF	0.341343
	SPY	0.600907

VIF (antes de limpieza)	Real GDP Growth	17.398
	Unemployment rate	13.560
VIF (tras limpieza)	SPY	3.507
	Credit spreads	3.245
	Interest rates	2.158
	INDEX	2.078
	Exchange Rates	2.081
	XLF	2.064
	VIX	2.355
	VOLUME	1.379
	SHARES	1.947
OLS Coef. (modelo base)	SPY	1.600590
	Exchange Rates	-0.767111
	Credit spreads	0.222541
	INDEX	0.013204
OLS Coef. (modelo optimizado)	SPY	1.575900
	Exchange Rates	-0.736893
	Credit spreads	0.213680
	INDEX2	0.019100
R²		0.427
MSE		0.007445
Correlación optimizada	INDEX2 y PRICE_Growth	0.152245
Mejores pesos (optimizados)	QUICK	0.8587
	CASH	0.2249
	DEBT	-0.1907
	REVENUES	0.0463
	EMPLOYEES, EPS, FCFMG, NETPROFITMG, OPERATINGMG, ROA, ROE	0.0087 cada uno

Tabla 3. Análisis de la regresión del índice con crecimiento porcentual del precio.

Indicador	Valor	Comentario
Index 2	0.0508	p = 0.028 (significativo al 5%)
R-cuadrado	0.013	El modelo explica el 1.3% de la varianza (no es relevante ya que no busca predecir)
P-valor del F-statistic	0.0286	Modelo globalmente significativo al 5%

RESULTADOS EMPÍRICOS

Analizar los resultados empíricos del modelo es una tarea complicada, ya que, dependiendo del enfoque que se le dé, se pueden obtener diversos resultados.

La hipótesis inicial parte de que la valoración errónea indicada por el modelo debe corregirse en el largo plazo, pero definir qué es largo plazo puede ser subjetivo, teniendo en cuenta también que para este estudio solo se cuenta con 5 años de información. De todas formas, se realizó una evaluación y análisis de los datos mediante diferentes métodos y enfoques.

Lo primero que se realizó fue el cálculo de una serie temporal que refleja el precio ideal de cada acción. Los residuos del modelo indican infravaloración o sobrevaloración del trimestre basado en sus métricas financieras respecto a la de los demás. Sin embargo, para aproximar el valor ideal de la acción se asumió que además de corregir por ese trimestre el precio corrige por el acumulado mal valorado anteriormente. La idea detrás de esto es no solamente tener lo que debería valer ese trimestre (que es poco interpretable ya que no se sabe cuándo puede corregir), sino sumar lo que debería haber cambiado el período anterior. Esto permite generar una serie actualizada que en cada momento refleja la valoración real de la empresa. Esto es muy importante porque una empresa puede pasar mucho tiempo infravalorada, pero en el último trimestre estar sobrevalorada, y esto no permite concluir en qué estado se encuentra la acción. En la Tabla 4 (a continuación), podemos observar una ilustración de esta idea para una de las empresas objeto del estudio.

Tabla 4. Ejemplo del análisis de resultados empíricos para la empresa Ameriprise Financial, Inc.

TICKER	DATE	Real	Ideal	Explicación	Corregido	Explicación	Residuo
AMP	2020_Q1	148,16					
AMP	2020_Q2	128,21	153,74	(Precio del trimestre) * (1- Residual)	153,74	(Precio del trimestre) * (1- Residual)	-19,92%
AMP	2020_Q3	154,00	157,97	(Precio del trimestre) * (1- Residual) + (Valor ideal anterior - Valor real anterior)	132,44	(Precio del trimestre) * (1- Residual)	14,00%
AMP	2020_Q4	178,99	165,02	IDEIM	161,05	IDEIM	10,02%
AMP	2021_Q1	216,23	165,10	IDEIM	179,07	IDEIM	17,19%
AMP	2021_Q2	252,07	174,88	IDEIM	226,01	IDEIM	10,34%
AMP	2021_Q3	261,59	189,53	IDEIM	266,73	IDEIM	-1,97%
AMP	2021_Q4	295,77	203,77	IDEIM	275,82	IDEIM	6,75%
AMP	2022_Q1	302,51	222,49	IDEIM	314,50	IDEIM	-3,96%
AMP	2022_Q2	267,88	235,41	IDEIM	315,43	IDEIM	-17,75%
AMP	2022_Q3	263,12	251,94	IDEIM	284,41	IDEIM	-8,09%
AMP	2022_Q4	305,01	264,27	IDEIM	275,44	IDEIM	9,69%
AMP	2023_Q1	327,63	282,49	IDEIM	323,23	IDEIM	1,34%
AMP	2023_Q2	307,35	300,25	IDEIM	345,38	IDEIM	-12,38%
AMP	2023_Q3	341,28	317,70	IDEIM	324,80	IDEIM	4,83%
AMP	2023_Q4	342,90	339,17	IDEIM	362,74	IDEIM	-5,79%
AMP	2024_Q1	400,39	348,23	IDEIM	351,96	IDEIM	12,10%
AMP	2024_Q2	427,19	373,34	IDEIM	425,50	IDEIM	0,39%
AMP	2024_Q3	433,86	400,46	IDEIM	454,31	IDEIM	-4,71%
AMP	2024_Q4	533,69	409,77	IDEIM	443,16	IDEIM	16,96%

La columna “real” muestra el precio de mercado promedio del trimestre. La columna “corregido” muestra el valor que debería haber tenido ese trimestre. La columna ideal muestra el valor “corregido” más el rendimiento ideal del trimestre anterior.

Así, las empresas que se encuentran sobre o infravaloradas al final de la serie son las mostradas en la siguiente tabla.

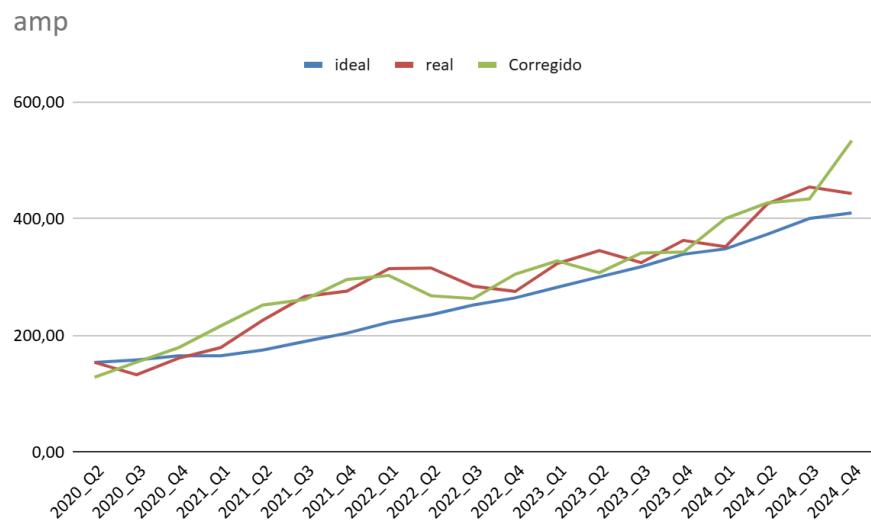
Tabla 5. Resultados de sobrevaloración o infravaloración de las empresas estudiadas.

TICKER	DATE	Valoración Acumulada)
BRK	2024_Q4	Infravalorada
C	2024_Q4	Infravalorada
SPGI	2024_Q4	Infravalorada
TFC	2024_Q4	Infravalorada
USB	2024_Q4	Infravalorada
V	2024_Q4	Infravalorada
AMP	2024_Q4	Sobrevalorada
AXP	2024_Q4	Sobrevalorada
BAC	2024_Q4	Sobrevalorada
BLK	2024_Q4	Sobrevalorada
CB	2024_Q4	Sobrevalorada
GS	2024_Q4	Sobrevalorada
JPM	2024_Q4	Sobrevalorada
MA	2024_Q4	Sobrevalorada
MMC	2024_Q4	Sobrevalorada
PNC	2024_Q4	Sobrevalorada
SCHW	2024_Q4	Sobrevalorada
WFC	2024_Q4	Sobrevalorada

Al ser un valor acumulado, la desviación de la valoración es interpretable para todos los períodos y es determinante (no como la trimestral que solo refleja un error para ese período).

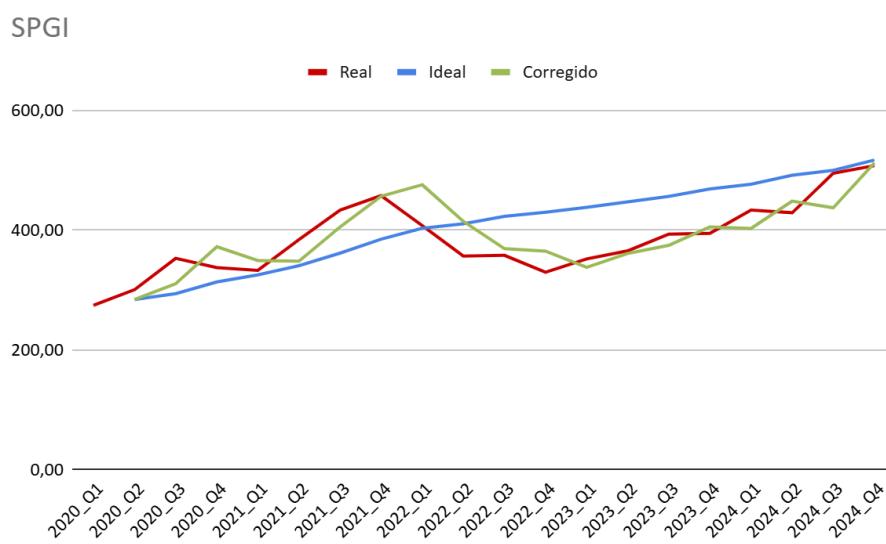
A partir de esta serie, que refleja el precio ideal y la serie de valores corregidos, se pueden observar tendencias y patrones muy interesantes, como se muestra en la Figura 2.

Figura 2. Comparación del precio ideal, real y corregido de Ameriprise Financial, Inc.



En este gráfico de la empresa AMP, se puede ver cómo el precio ideal se mantiene por debajo del real desde el año 2021 en adelante, mostrando una sobrevaloración persistente. Es extremadamente notable cómo la serie actúa como soporte del precio real, en especial del período de 2022 hasta 2024, y cómo la acción cada vez que diverge vuelve a buscar un soporte en el precio ideal. Por otro lado, la serie de Corregido alterna con la de precio real y es visible cómo el precio corregido anticipa el movimiento del precio real para el movimiento de los siguientes (aunque no se puede afirmar que tenga impacto predictivo debido a la alta autocorrelación).

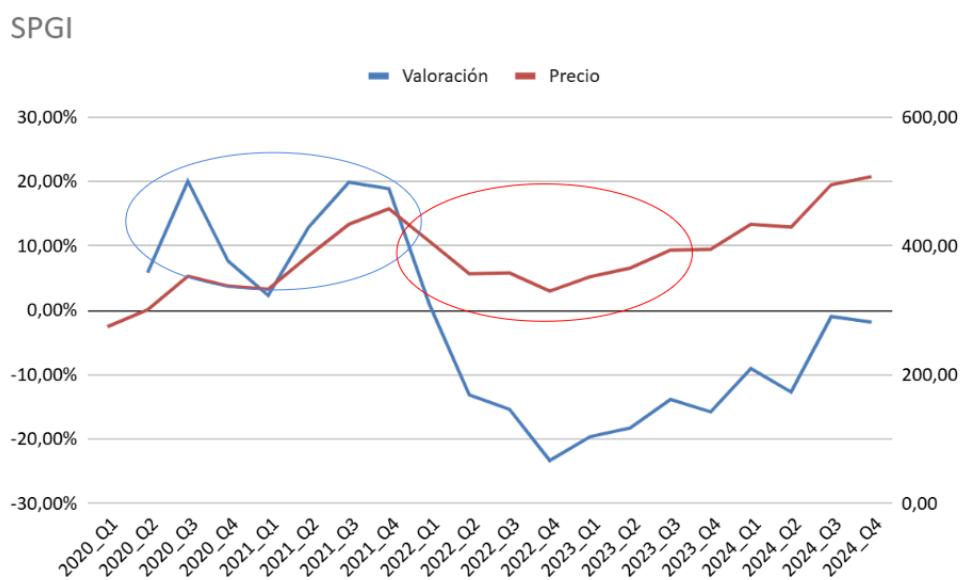
Figura 3. Comparación del precio ideal, real y corregido SPGI.



El gráfico de SPGI es consistente con el de AMP, aunque en este caso empieza sobrevalorada (actuando de soporte de la misma forma que el anterior) hasta 2022, donde pasa a estar infravalorada y al final converge. Estos gráficos pueden ser utilizados para estudiar el comportamiento de la acción respecto a su valor teórico y ver qué escenarios pueden suceder luego de largos períodos o períodos puntuales en los cuales la desviación de la valoración fue muy elevada. El estudio del comportamiento de estas variables se extrae a todos los gráficos de todas las empresas ([ver a partir de la Figura A7](#)).

En el siguiente gráfico, se puede analizar el historial de valoración (si es positivo indica sobrevaloración y si es negativo indica infravaloración) y observar qué sucede en un período posterior a esta. En el gráfico de S&P Global, se puede ver cómo la sobrevaloración inicial persiste dos años y luego de este período la acción presenta un decrecimiento considerable de su precio.

Figura 4. Resultado de la empresa S&P Global.



En el gráfico de MasterCard, se puede observar la situación inversa. Luego de un período largo de infravaloración y muy grande (más de 20%), se refleja un crecimiento posterior al período del precio de la acción mayor al previo.

Figura 5. Resultado de la empresa MasterCard.



A continuación se hizo un estudio que separaba los datos en dos períodos:

- Q1_2020 a Q4_2021
- Q1_2022 a Q4_2024

Para el primer período, se calculó la media de la desviación de la valoración de las empresas (positiva es sobrevaloración y negativa es infravaloración) y luego el crecimiento del precio promedio para la segunda mitad. Se puede comparar y vemos que las empresas infravaloradas la primera mitad tienen un rendimiento superior a las sobrevaloradas, lo cual se alinea con el diseño del proyecto, que busca realizar una valoración relativa.

Tabla 6. Comparación de crecimiento.

Media de crecimiento de sobrevaloradas (en la primera mitad) para la segunda mitad	1,94%
Media de crecimiento de Infravaloradas (en la primera mitad) para la segunda mitad	3,01%

Otro resultado interesante se obtiene al ver los períodos con más diferencia entre valor real y valor ideal e intentar vincularlos a eventos. El año 2021 fue un año marcado por la recuperación económica post-covid en donde las empresas crecieron entre un 20 y 30% más de lo que sus fundamentales justificaban. De la misma manera, es observable cómo 2020 durante el Covid-19 fue el período con más infravaloración (-10 a -15%), explicada por factores externos a las empresas. Otro período marcado como infravaloración es 2023, año en el cual quiebra

Silicon Valley Bank (y colapsos posteriores), subidas agresivas de tipos de interés y sentimiento de mercado negativo.

Tabla 7. Media de desviación de la valoración.

DATE	Media de desviación de la valoración
Mayor sobrevaloración	
2021_Q2	27,44%
2021_Q4	26,70%
2021_Q3	25,07%
2024_Q4	15,61%
2021_Q1	14,75%
Mayor infravaloración	
2023_Q3	-7,62%
2023_Q4	-10,03%
2023_Q2	-10,65%
2020_Q3	-11,93%
2020_Q2	-15,22%

A continuación, se intentó averiguar si un residuo de determinado signo conlleva a un cambio en el precio del siguiente trimestre del signo opuesto (es decir, si la infravaloración en un trimestre lleva a un crecimiento del precio en el siguiente y viceversa). Se encontró que esta relación se cumple un 42,9% de las veces (tabla 8). Por otro lado, también se comparó el signo del residuo de un trimestre contra su crecimiento del precio del siguiente trimestre (si fue mayor a la media o no). Esta relación se cumple el 48,15% de las veces (Tabla 9). La relación entre trimestre parece no ser clara, por eso planteamos un horizonte largo plazo).

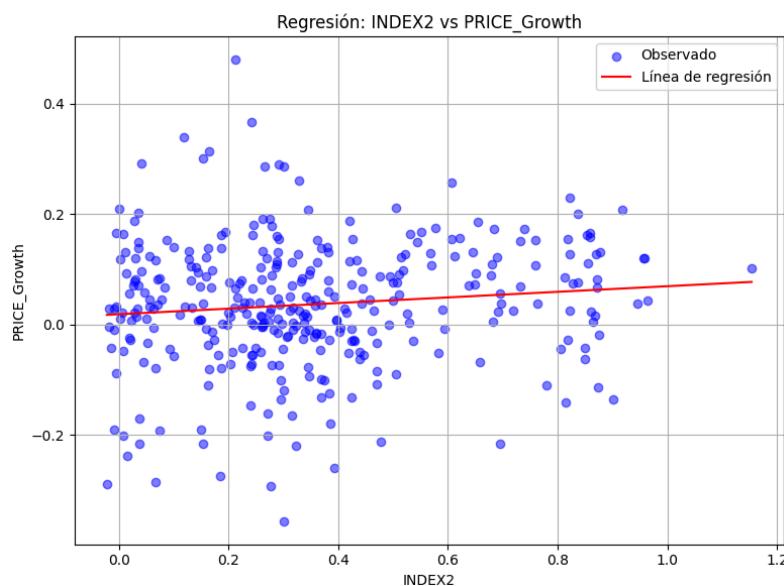
Tabla 8. Efecto del residuo en el crecimiento del precio en el siguiente trimestre.

TICKER	DATE	Residuals	Price growth (del siguiente cuatrimestre)	Crecimiento
AMP	2020_Q2	-19,92%	0,201207667	Sí
AXP	2020_Q2	-27,05%	0,06968353182	Sí
BAC	2020_Q2	-26,47%	0,0537319245	Sí

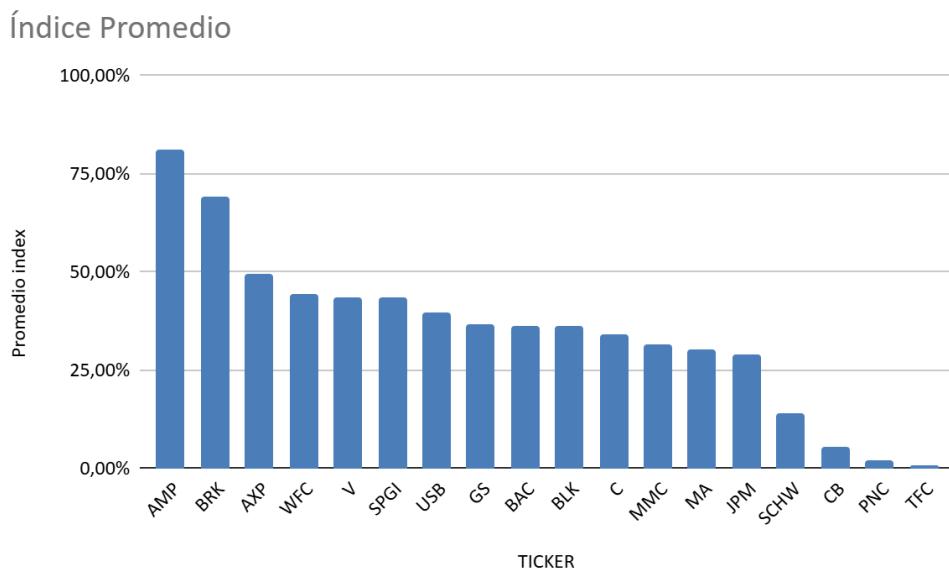
Tabla 9. Crecimiento de precio superior a la media tras un residuo de signo opuesto.

TICKER	DATE	Residuals	Price growth (del siguiente cuatrimestre)	Media del trimestre	Mayor a media
AMP	2020_Q2	-19,92%	0,201	6,64%	SI
AXP	2020_Q2	-27,05%	0,069	6,64%	SI
BAC	2020_Q2	-25,47%	0,053	6,64%	NO

Finalmente, se estudia la relación entre el índice, el crecimiento en el precio y su coherencia con la realidad. La mayoría de rendimientos trimestrales se centran entre -10% y 10%, y la relación positiva entre índice y rendimiento es observable. La intensidad de resultados se centra alrededor de la regresión, lo que refleja que el mercado sigue de cierta forma al valor intrínseco, pero hay desviaciones grandes que pueden ser explicadas por factores externos.

Figura 5. Regresión entre INDEX2 con PRICE_Growth.

Las empresas con un mayor índice promedio a lo largo del período son AMP, BRK y AXP y las peores son TFC, PNC y CB (primer gráfico). El primer grupo tiene un historial financiero de márgenes altos, ingresos crecientes y elevados y gestión eficiente de deuda. El segundo grupo ha presentado un deterioro general de su salud financiera desde el año 2020. En el gráfico posterior se puede ver también la diferencia extraordinaria de rendimiento del precio de la acción para el período. La relación entre (índice -Salud financiera- Crecimiento) del precio se ve explícitamente.

Figura 6. Índice promedio de las empresas estudiadas.**Tabla 10.** Cambio en el precio de las empresas con mejor y peor índice promedio.

TICKER	Cambio en el precio (Q4_2024/Q1_2020)
AMP	220,36%
BRK	100,44%
AXP	142,40%
CB	75,00%
PNC	30,00%
TFC	-15,54%

DECISIONES Y POTENCIALES MEJORAS

Durante el proceso de creación del modelo, se encontraron diversos problemas que posteriormente se solucionaron mientras que otros no se pudieron. A continuación se detallan los más significativos.

1. Fiabilidad de los datos (solucionada con SEC y 10-Q).
2. Multicolinealidad (solucionada con VIF).
3. Transformaciones de datos (solucionada probando diferentes formas).
 - a. Min-Max para fundamentales y log-diff para control.
4. Añadir más variables fundamentales, empresas y períodos (Usar PCA - *Principal component analysis* para su selección).

5. Añadir más variables que disminuyan el ruido de mercado.
6. Mejorar la relación de índice con el cambio en el precio (agregar dividendos).
7. Mitigar la autocorrelación con modelos ARIMA.
8. Probar relaciones no lineales de las variables, añadir lags e interacción entre variables.
9. Agregar análisis técnico y ratios de valoración.
10. Utilizar diferentes algoritmos (uso del Bayesiano debido a mejor ajuste de hiperparámetros).
11. Analizar burbujas de la bolsa.
12. Selección de parámetros de algoritmo (Balance entre coherencia y ajuste).
13. Añadir variación del índice como predictor.
14. Usar estimaciones para incorporar valoración futura (estimación de EPS, estimaciones de ingresos, estimaciones de flujo de caja, etc).

CONCLUSIONES

A partir del análisis de correlaciones y de la primera regresión, se puede llegar a la primera conclusión del trabajo: El sector financiero parece ser mayoritariamente dependiente de la macroeconomía y del comportamiento general del mercado. Esto se puede observar dados los coeficientes significativos y elevados de la primera regresión ([ver figura A5](#)). Esto significa que el trabajo de aislar el efecto de las variables fundamentales puede ser más complicado pero más necesario a la vez.

Una conclusión intermedia es que la estandarización Min-Max es adecuada para la comparación relativa de variables financieras y para generar un índice de salud financiera.

Otra cosa que se puede concluir es que la creación de un índice para evaluar la salud financiera supera al análisis de las variables de forma independiente ya que se puede concentrar su efecto obteniendo coeficientes más elevados y correlaciones con cambio en el precio con un mayor peso.

Como cuarta conclusión, se puede decir que el enfoque de regresiones con algoritmo bayesiano permite optimizar el índice mejorando la base inicial y obteniendo métricas más coherentes y precisas lo que puede ayudar a evaluar la salud financiera de una empresa atada

a su rendimiento de mercado. Se puede concluir que a pesar de las limitaciones el enfoque es representativo y con más tiempo y recursos podría seguir perfeccionándose.

El objetivo de mejora sería principalmente lograr un coeficiente de índice positivo significativo en la primera regresión sin perder interpretabilidad del índice. Sin asignar límites a los pesos se puede conseguir esto pero se obtiene un índice que tiene pesos incoherentes (Por ejemplo peso del Roe de -0,99 lo cual no tiene sentido). Agregar más variables fundamentales también podría ayudar a solucionar esto. Los resultados econométricos están en un rango aceptable pero mejorable expandiendo el horizonte temporal, la cantidad de variables y añadiendo formas de evaluar e interpretar los resultados.

La creación de una serie “Precio ideal” refleja la idea del trabajo de buscar analizar infravaloración o sobrevaloración sostenida y contiene el pensamiento de que el error en la valoración tiende a corregirse. A partir de esta serie se pueden analizar muchas cosas y permite realizar análisis futuros que pueden ser útiles para tomar decisiones. Por otro lado, “Precio Corregido” posee una alineación sorprendente con Precio del siguiente trimestre (Correlación de 0,96) pero no se ha podido demostrar que tenga poder predictivo real fuera de la muestra.

El precio ideal podría tener uso en el análisis técnico de acciones, usado como *benchmark* (referencia) para determinar compras y ventas para determinados períodos de tiempo o generar métricas relacionadas a su desviación, magnitud de la desviación y tiempo de la desviación.

En varios gráficos se puede observar que períodos largos de infravaloración están seguidos por períodos de crecimiento de precios y viceversa. Se puede ver la notoria diferencia de crecimiento entre las sobre e infravaloradas para períodos posteriores, lo cual se alinea con el planteo establecido de valoración relativa.

Finalmente se puede concluir que, en caso de ser perfeccionado, el enfoque teórico es adecuado y el enfoque práctico puede tener aplicaciones en el mundo financiero.

El diseño del modelo aísla el efecto de las características fundamentales de las empresas del sector financiero para poder obtener un índice que refleje la salud financiera atado al crecimiento del precio de la acción. El problema que resuelve es poder encontrar la mejor forma de medir la salud financiera relativa de las empresas a lo largo del período de forma pura (sin efecto externo). Esto es así ya que evaluar solamente las variables fundamentales contra el precio de la acción lleva a que el peso de las variables y su impacto se vea sesgado por factores

externos (Mercado, Macro, Sentimiento, etc). De esta manera se puede hallar un precio limpio que refleja solamente el valor fundamental de una empresa basado en la diferencia de su rentabilidad, eficiencia y financiamiento respecto a sus competidores.

BIBLIOGRAFÍA

- Damodaran, A. (2007). *Valuation approaches and metrics: A survey of the theory and evidence.* Now Publishers Inc.
- Graham, B., & Dodd, D. (1996). *Security Analysis.* McGraw-Hill.
- Hagstrom, R. G. (2023). *The Warren Buffett way (3rd ed.).* John Wiley & Sons.
- (2025), *Investing.com - stock market quotes & financial news.* Consultada el 8 de mayo de 2025. en <http://investing.com>.
- (2025), *Stock analysis - free online stock information for investors.* Consultada el 8 de mayo de 2025. en <https://stockanalysis.com/>.
- (2025), *FinChat.io - The complete AI powered stock research platform.* Consultada el 8 de mayo de 2025. en <https://finchat.io>.
- (2025), *MarketScreener - Financial News & Stock Market Quotes.* Consultada el 9 de mayo de 2025. en [https://www.marketscreener.com/](https://www.marketscreener.com).
- (2025), *Macrotrends - The Long Term Perspective on Markets.* Consultada el 9 de mayo de 2025. en <https://macrotrends.net>.
- (2025), *Federal Reserve Economic Data, FRED, St. Louis Fed.* Consultada el 9 de mayo de 2025. en <https://fred.stlouisfed.org>.
- (2025), *Yahoo Finance.* Consultada el 12 de mayo de 2025. en <https://finance.yahoo.com>.

ANEXOS

Aclaración: Todas las tablas, gráficas y código de python son elaboración propia.

Diccionario 1. Empresas seleccionadas para el estudio.

AMP (Ameriprise Financial, Inc.), AXP (American Express), BAC (Bank of America Corp), BLK (BlackRock), BRK (Berkshire Hathaway Inc. (BRK-B)), C (Citigroup), CB (Chubb Ltd), GS (Goldman Sachs Group Inc), JPM (JPMorgan Chase & Co), MA (MasterCard), MMC (Marsh & McLennan Companies Inc), PNC (PNC Financial Services), SCHW (Charles Schwab Corporation Common Stock), SPGI (S&P Global), TFC (Truist Financial Corp), USB (US Bancorp), V (Visa) y WFC (Wells Fargo & Co).

Figura A1. Matriz de correlaciones entre las empresas analizadas.

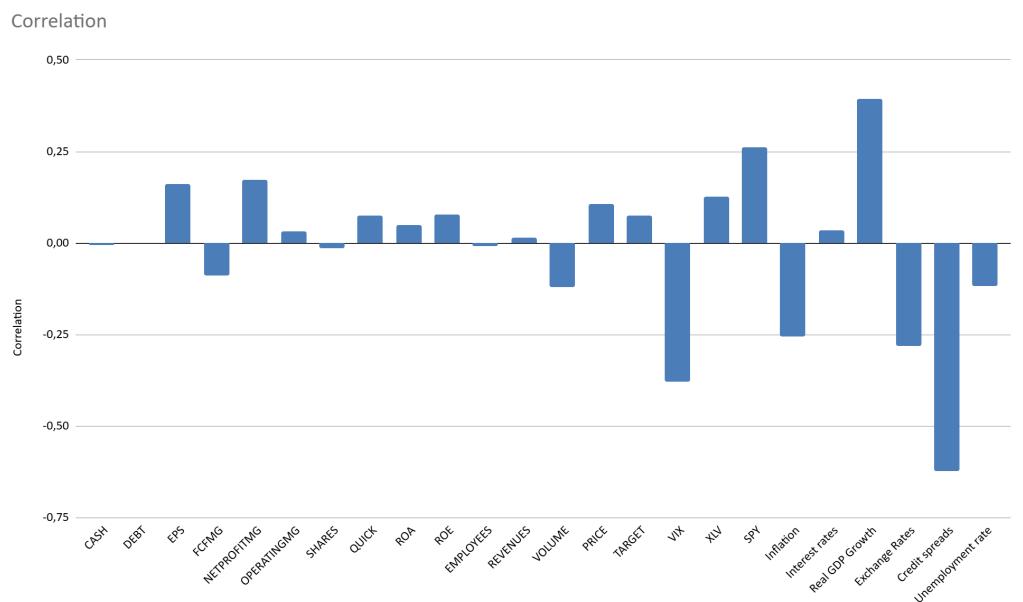
Ticker	V	MA	JPM	BLK	SPGI	SCHW	GS	C	BAC	XLF	WFC	MMC	USB	AXP	AMP
V	1.00	0.91	0.51	0.51	0.63	0.32	0.49	0.55	0.50	0.60	0.50	0.53	0.46	0.70	0.52
MA	0.91	1.00	0.51	0.49	0.65	0.34	0.52	0.59	0.52	0.62	0.48	0.51	0.47	0.70	0.53
JPM	0.51	0.51	1.00	0.70	0.41	0.71	0.85	0.85	0.91	0.87	0.80	0.49	0.80	0.71	0.78
BLK	0.51	0.49	0.70	1.00	0.63	0.58	0.68	0.67	0.71	0.71	0.51	0.53	0.63	0.57	0.70
SPGI	0.63	0.65	0.41	0.63	1.00	0.33	0.51	0.46	0.48	0.54	0.33	0.68	0.40	0.57	0.52
SCHW	0.32	0.34	0.71	0.58	0.33	1.00	0.64	0.63	0.78	0.71	0.71	0.30	0.70	0.57	0.68
GS	0.49	0.52	0.85	0.68	0.51	0.64	1.00	0.82	0.87	0.81	0.75	0.55	0.72	0.69	0.75
C	0.55	0.59	0.85	0.67	0.46	0.63	0.82	1.00	0.87	0.84	0.78	0.48	0.77	0.77	0.76
BAC	0.50	0.52	0.91	0.71	0.48	0.78	0.87	0.87	1.00	0.88	0.83	0.52	0.83	0.74	0.83
XLF	0.60	0.62	0.87	0.71	0.54	0.71	0.81	0.84	0.88	1.00	0.76	0.62	0.80	0.77	0.81
WFC	0.50	0.48	0.80	0.51	0.33	0.71	0.75	0.78	0.83	0.76	1.00	0.46	0.76	0.76	0.69
MMC	0.53	0.51	0.49	0.53	0.68	0.30	0.55	0.48	0.52	0.62	0.46	1.00	0.44	0.51	0.60
USB	0.46	0.47	0.80	0.63	0.40	0.70	0.72	0.77	0.83	0.80	0.76	0.44	1.00	0.67	0.71
AXP	0.70	0.70	0.71	0.57	0.57	0.57	0.69	0.77	0.74	0.77	0.76	0.51	0.67	1.00	0.70
AMP	0.52	0.53	0.78	0.70	0.52	0.68	0.75	0.76	0.83	0.81	0.69	0.60	0.71	0.70	1.00

Tabla A1. Descripción de variables utilizadas en el análisis.

Nombre de la Variable	Representa	Transformación	Período Inicial Tras Transformación	Primer Valor (t, t+1, t+2)
CASH	Reservas de efectivo de la empresa	Min-Max Scaling por DATE, luego shift(1) por TICKER	Q1_2020	t
DEBT	Ratio de deuda	Min-Max Scaling por DATE, luego shift(1) por TICKER	Q1_2020	t

EPS	Ganancias por acción	Min-Max Scaling por DATE, luego shift(1) por TICKER	Q1_2020	t
FCFMG	Margen de flujo de caja libre	Min-Max Scaling por DATE, luego shift(1) por TICKER	Q1_2020	t
NETPROFITMG	Margen de ganancia neta	Min-Max Scaling por DATE, luego shift(1) por TICKER	Q1_2020	t
OPERATINGMG	Margen operativo	Min-Max Scaling por DATE, luego shift(1) por TICKER	Q1_2020	t
QUICK	Ratio rápido (medida de liquidez)	Min-Max Scaling por DATE, luego shift(1) por TICKER	Q1_2020	t
ROA	Retorno sobre activos	Min-Max Scaling por DATE, luego shift(1) por TICKER	Q1_2020	t
ROE	Retorno sobre equidad	Min-Max Scaling por DATE, luego shift(1) por TICKER	Q1_2020	t
EMPLOYEES	Número de empleados	Min-Max Scaling por DATE, luego shift(1) por TICKER	Q1_2020	t
REVENUES	Ingresos totales	Min-Max Scaling por DATE, luego shift(1) por TICKER	Q1_2020	t
Interest rates	Tasas de interés del mercado	Cambio porcentual (pct_change) por TICKER	Q2_2020	t+1
Real GDP Growth	Tasa de crecimiento del PIB	Ninguna	Q2_2020	t+1
Exchange Rates	Tasas de cambio de divisas	Diferencia logarítmica por TICKER	Q2_2020	t+1
Credit spreads	Diferenciales de crédito	Diferencia logarítmica por TICKER	Q2_2020	t+1

Unemployment rate	Tasa de desempleo nacional	Diferencia logarítmica por TICKER	Q2_2020	t+1
VIX	Índice de volatilidad del mercado	Diferencia logarítmica (serie completa)	Q2_2020	t+1
XLF	Precio del ETF(Exchange Traded Fund) del sector financiero	Diferencia logarítmica (serie completa)	Q2_2020	t+1
SPY	Precio del ETF S&P 500	Diferencia logarítmica (serie completa)	Q2_2020	t+1
VOLUME	Volumen de negociación del valor	Diferencia logarítmica por TICKER	Q2_2020	t+1
SHARES	Cambio neto en las acciones que poseen los fondos de inversión	Diferencia logarítmica por TICKER, luego Min-Max Scaling por DATE	Q2_2020	t+1
PRICE_Growth	Cambio porcentual trimestral de PRICE	Cambio porcentual (pct_change) por TICKER	Q2_2020	t+1
INDEX	Índice ponderado de fundamentales	Suma ponderada de variables fundamentales escaladas	Q2_2020	t+1
INDEX2	Índice ponderado optimizado de fundamentales	Suma ponderada con pesos optimizados	Q2_2020	t+1

Figura A2. Correlación de las variables con el rendimiento de la acción.**Tabla A2.** Matriz de correlación de variables fundamentales

	CASH	DEBT	EPS	FCFM G	NETP ROFIT MG	OPE. MG	QUICK	ROA	ROE	EMPL OYEES	REVE NUES
CASH	1.000	0.291	-0.116	-0.068	-0.083	0.139	-0.127	-0.268	-0.217	0.701	-0.060
DEBT	0.291	1.000	0.081	-0.125	-0.003	-0.070	0.100	-0.074	0.063	0.128	-0.162
EPS	-0.116	0.081	1.000	0.016	0.355	0.225	0.306	0.195	0.163	0.013	-0.075
FCFM G	-0.068	-0.125	0.016	1.000	0.100	0.068	-0.042	0.097	0.088	-0.113	0.079
NETP ROFIT MG	-0.083	-0.003	0.355	0.100	1.000	0.447	0.105	0.658	0.549	-0.267	-0.107

OPERATING MG	0.139	-0.070	0.225	0.068	0.447	1.000	-0.110	0.401	0.281	0.032	0.007
QUICK	-0.127	0.100	0.306	-0.042	0.105	-0.110	1.000	0.141	0.206	0.095	0.036
ROA	-0.268	-0.074	0.195	0.097	0.658	0.401	0.141	1.000	0.808	-0.261	0.107
ROE	-0.217	0.063	0.163	0.088	0.549	0.281	0.206	0.808	1.000	-0.292	-0.034
EMPL OYEEES	0.701	0.128	0.013	-0.113	-0.267	0.032	0.095	-0.261	-0.292	1.000	0.188
REVENUES	-0.060	-0.162	-0.075	0.079	-0.107	0.007	0.036	0.107	-0.034	0.188	1.000

Figura A3. Correlación de las variables con el rendimiento de la acción después de estandarizar.

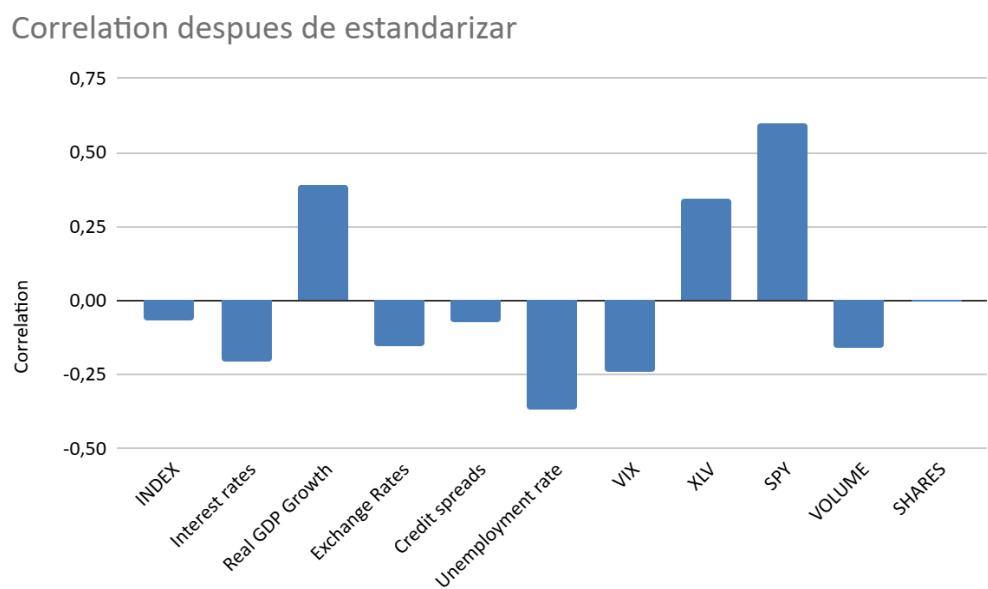
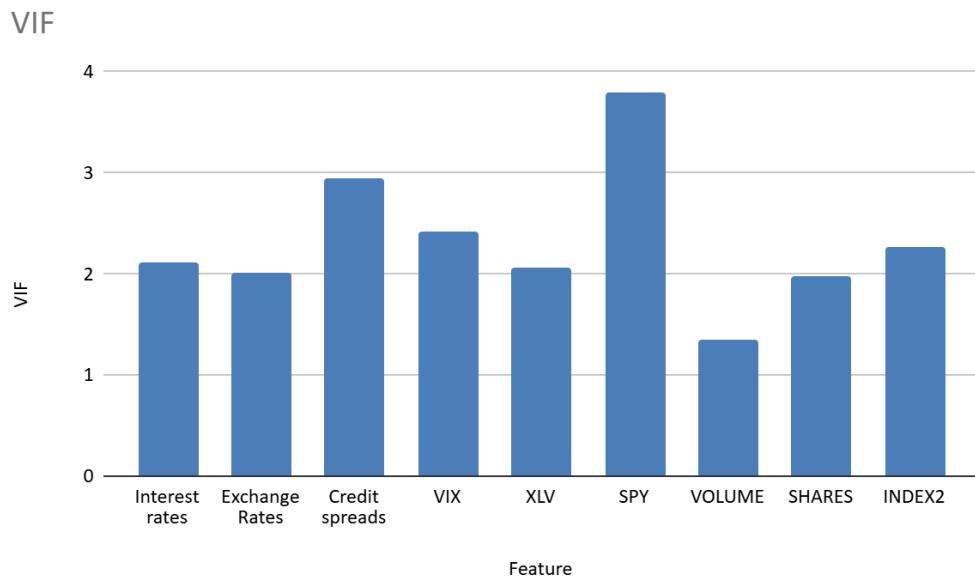
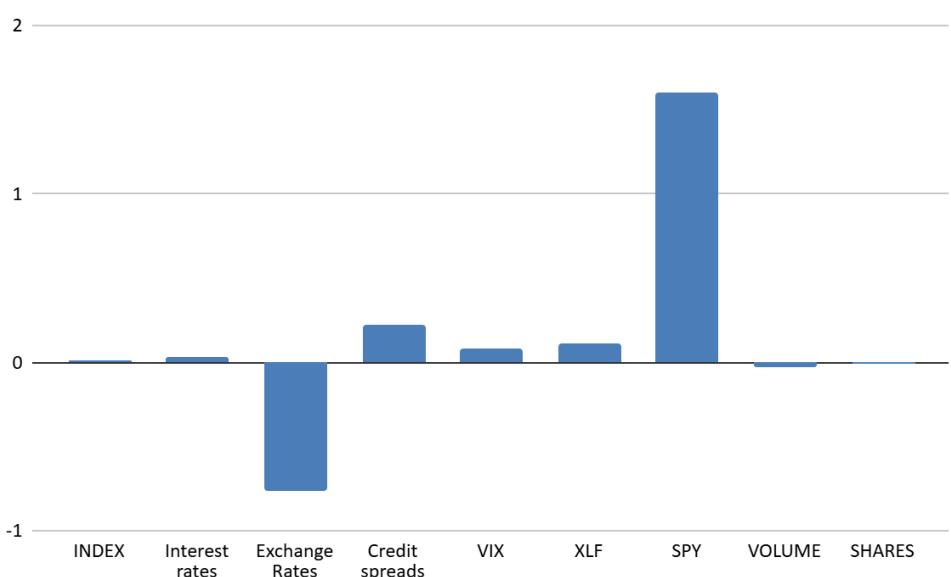


Figura A4. Valores VIF tras la depuración de variables.**Figura A5.** Coeficientes Relevantes y Significativos en la Regresión Inicial.**Tabla A3.** Selección de parámetros para el algoritmo bayesiano.

Criterio / Métrica	Resultado 1	Resultado 2	Resultado 3	Resultado 4	Resultado 5
Puntos iniciales	30	100	70	70	40
Iteraciones	600	1.000	700	700	400
Alpha	0,5	0,5	0,25	0,75	0,25

Límite de pesos	Sin límites	Sin límites	Sin límites	Sin límites	Penalización por pesos extremos
Criterio 1. Poder predictivo de INDEX2 sobre el crecimiento del precio.					
R ² primera regresión	0.4262	0.4262	0.4262	0.4264	0.427
MSE primera regresión	0.00746192	0.00746174	0.00746180	0.00745942	0.007445
Coef. INDEX2 (significancia)	-0.004780 (no sig.)	-0.004749 (no sig.)	-0.003816 (no sig.)	-0.004688 (no sig.)	0.019100 (No sig.)
Coef. SPY (significancia)	1.606210 (sig.)	1.606183 (sig.)	1.610841 (sig.)	1.608306 (sig.)	1.575900 (sig)
Multicolinealidad	Possible	Possible	Possible	Possible	Menor
Evaluación	Regular	Regular	Regular	Regular	Bueno
Criterio 2. Interpretabilidad del INDEX2.					
Peso DEBT	-0.394060 (correcto)	-0.391507 (correcto)	-1.000000 (correcto, extremo)	-0.996195 (correcto)	-0.191 (correcto)
Pesos positivos esperados	ROE: 0.873, NPM: 0.760	NPM: 1, FCFMG: 0.327	NPM, QUICK, REVENUES: 1	ROE, NPM, FCFMG: positivos	CASH: 0.225 Quick: 0.859 Revenues: 0.046 otros: 0.009
Pesos negativos incoherentes	EPS, OPERATINGMG, ROA	EPS, OPERATINGMG, ROA, ROE	EPS, OPERATINGMG, ROA, ROE	EPS, OPERATINGMG, ROA	Ninguno
Pesos extremos (1 o -1)	CASH, QUICK: 1	CASH, QUICK: 1	Varios: 1 o -1	QUICK, CASH, REVENUES, ROE: 1	No tiene
Evaluación	Malo	Malo	Malo	Malo	Excelente
Criterio 3. Análisis de residuos de valoración.					
R ² última regresión	0.047	0.049	0.053	0.045	0.013
Coef. INDEX2 (última regresión)	0.0466 (p=0.000)	0.0470 (p=0.000)	0.0381 (p=0.000)	0.0376 (p=0.000)	0.0508 (p = 0.028)
Sesgo de residuos	-0.1854	-0.2361	-0.1650	-0.1264	-0.1392
Curtosis de residuos	1.3026	1.2706	1.2882	1.4789	1.3311

Estadístico de Durbin-Watson	1.7036	1.6980	1.6807	1.6881	1.6759
Evaluación	Bueno	Bueno	Excelente	Bueno	Regular

Figura A6. Ameriprise

INDEX2 i PRICE_Growth (AMP)



residuals i PRICE (AMP)



residuals i price growth (AMP)

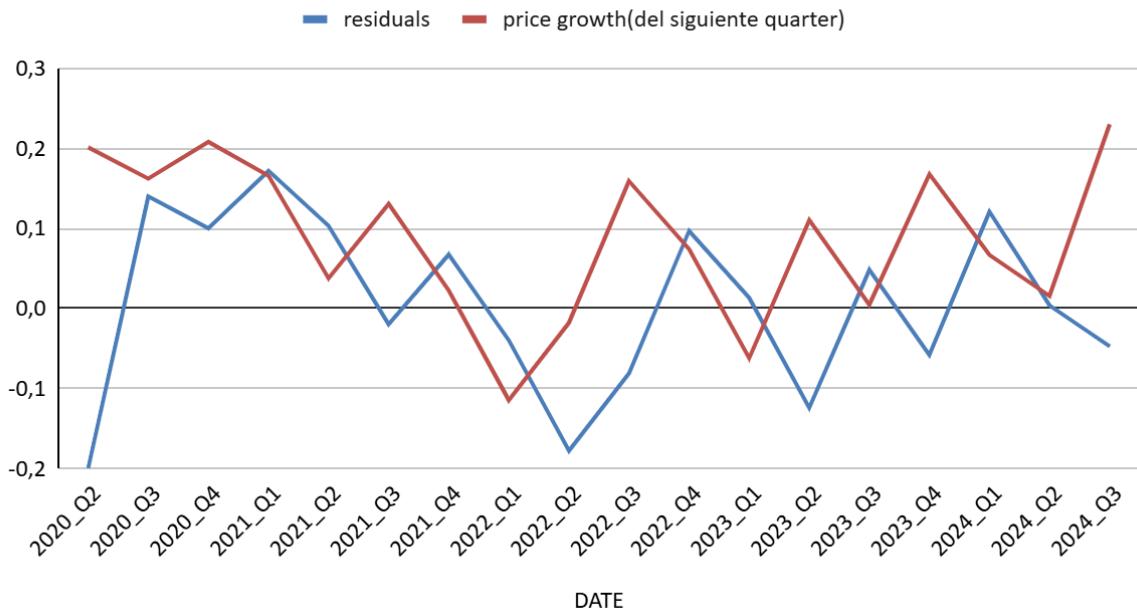


Figura A7. American Express

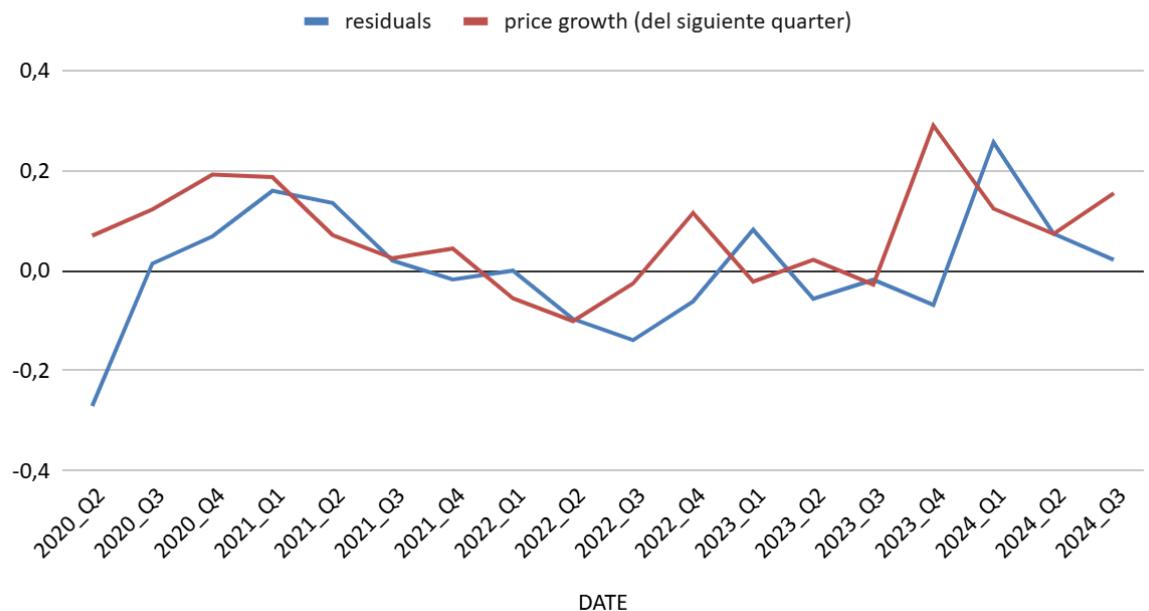
INDEX2 i PRICE_Growth (AXP)



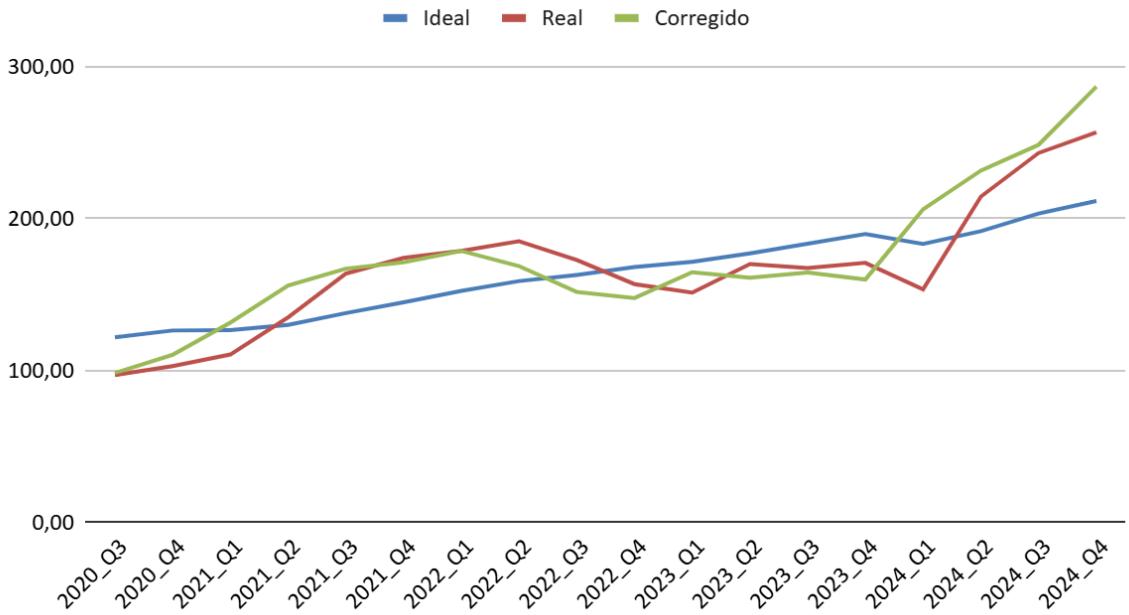
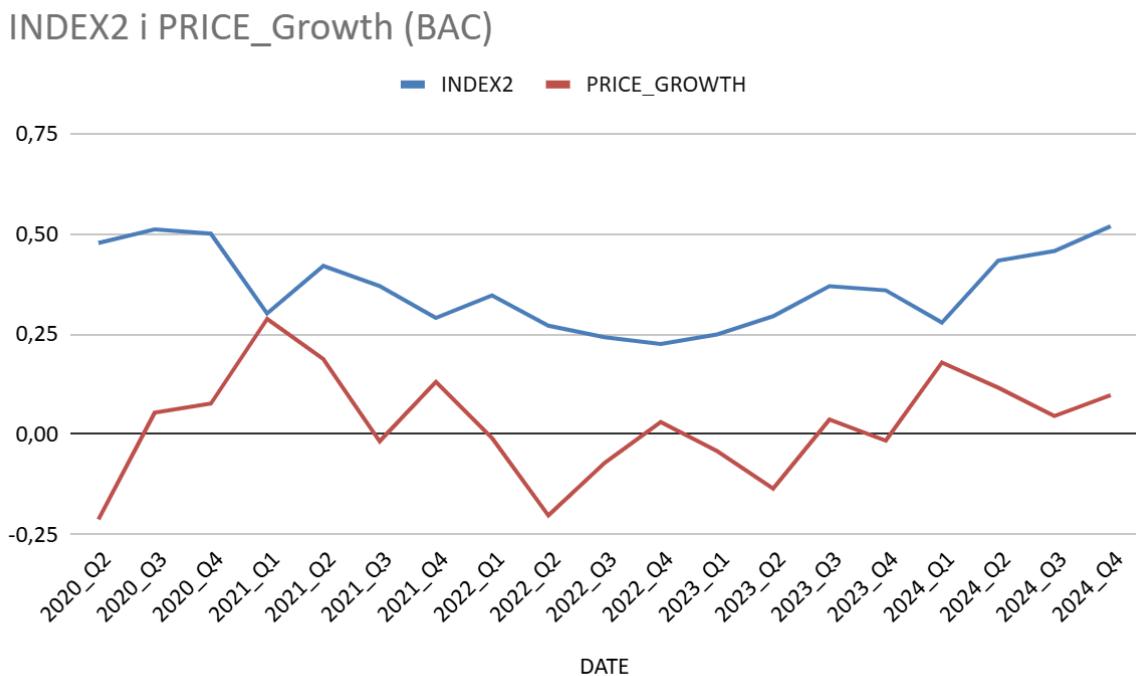
residuals i PRICE (AXP)



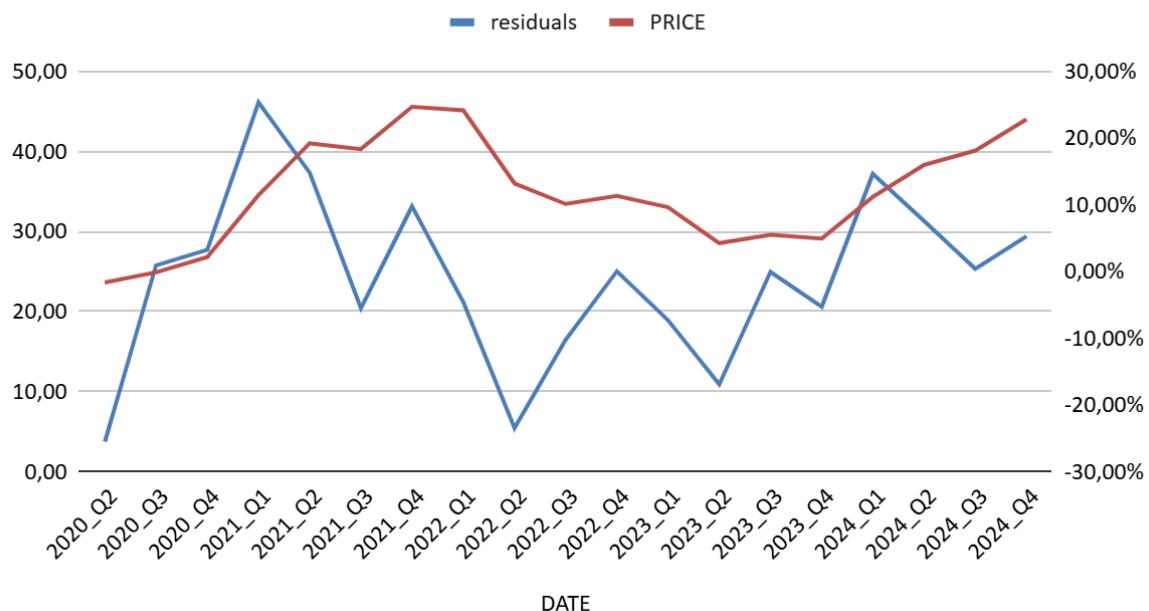
residuals i price growth (AXP)



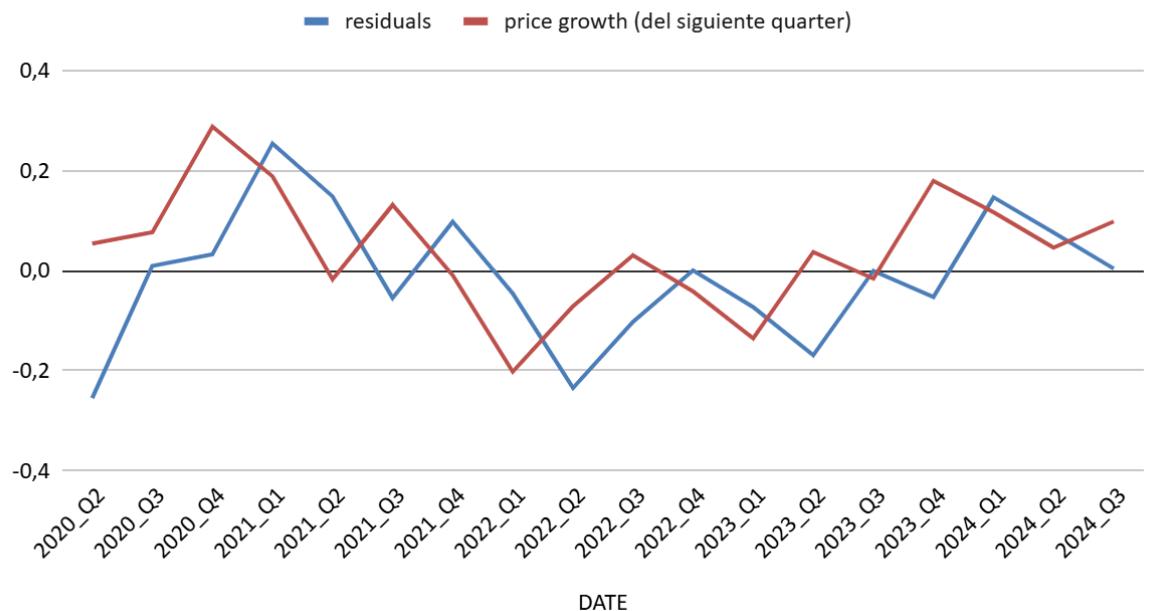
axp

**Figura A8.** Bank of America

residuals i PRICE (BAC)



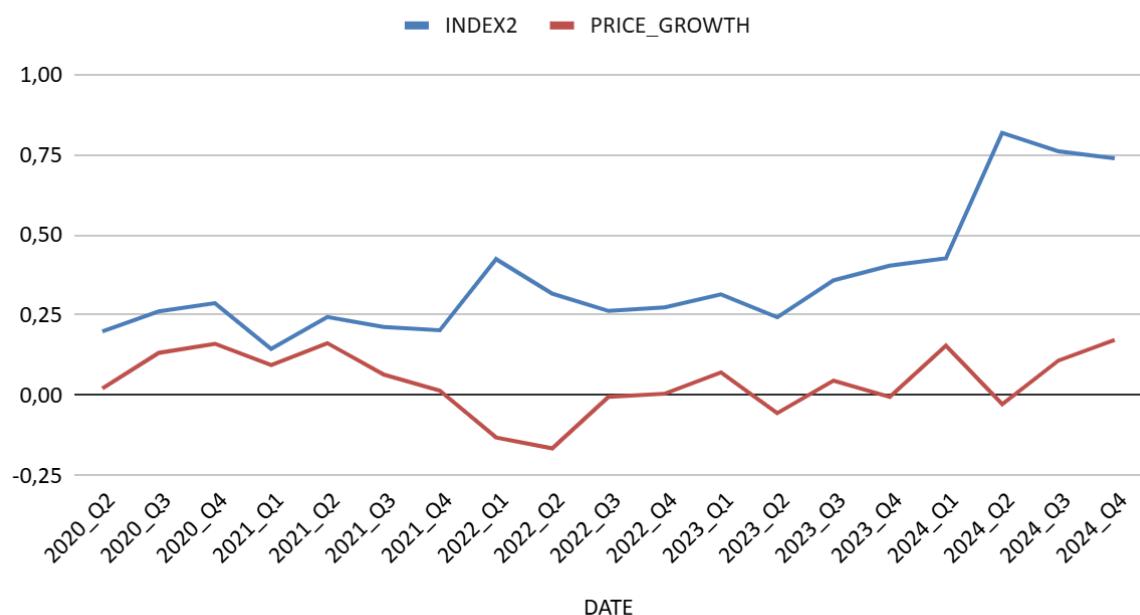
residuals i price growth (BAC)



bac

**Figura A9.** Blackrock INC

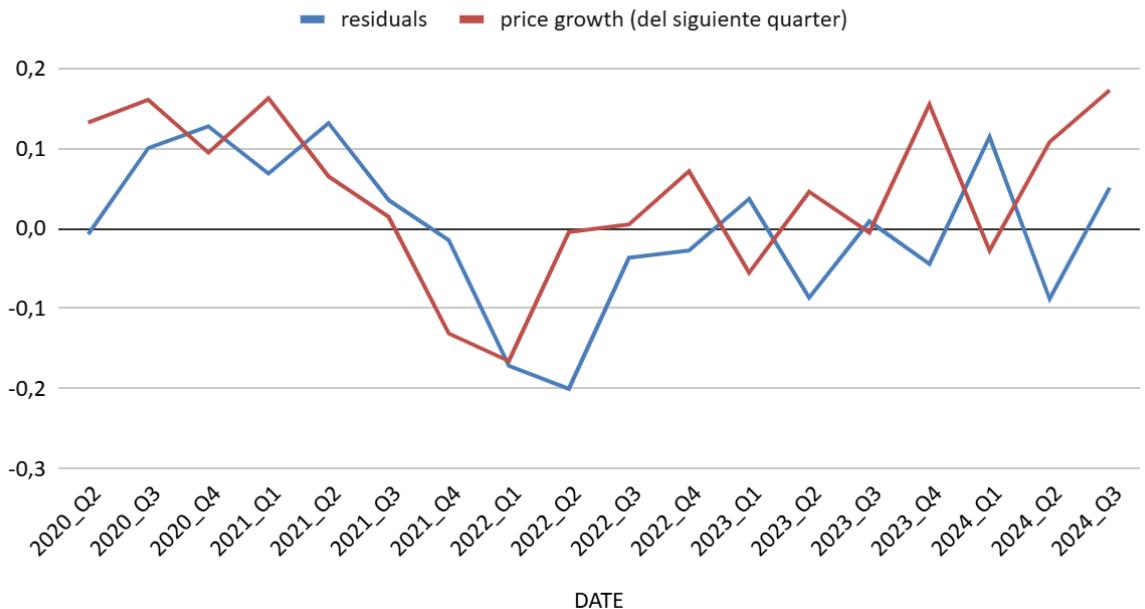
INDEX2 i PRICE_Growth (BLK)



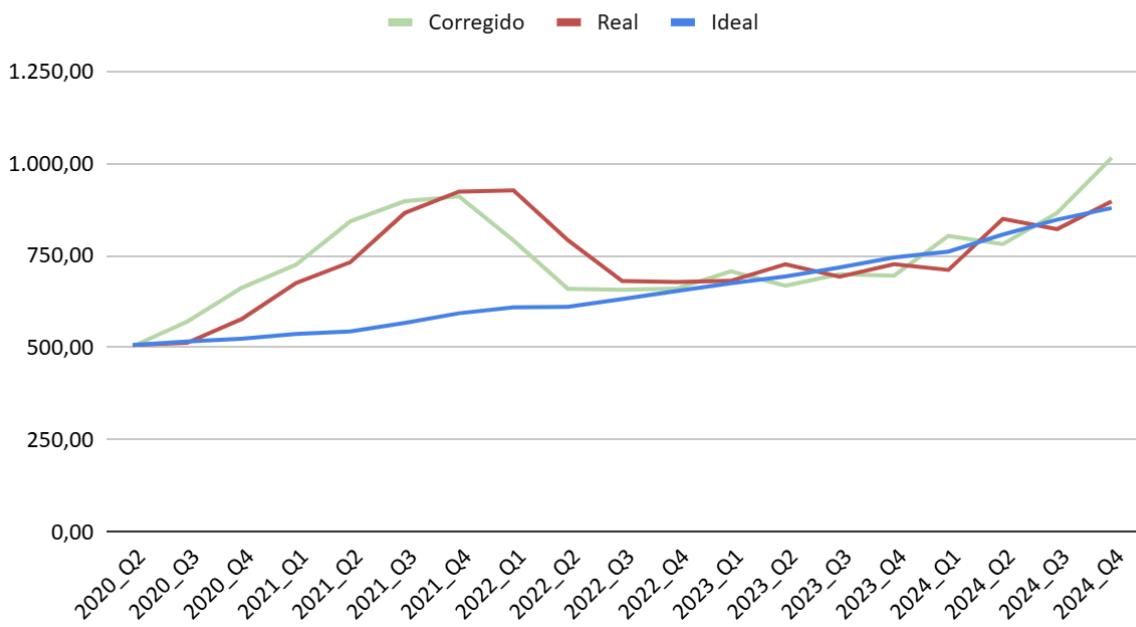
residuals i PRICE (BLK)



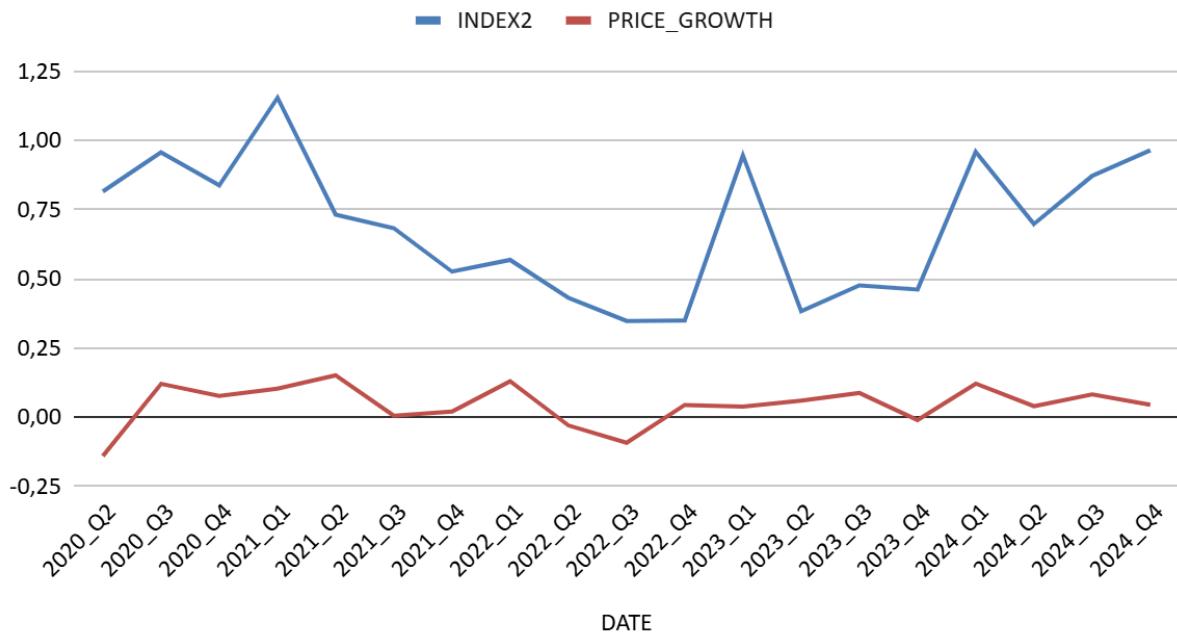
residuals i price growth (BLK)



blk

**Figura A10.** Berkshire Hathaway

INDEX2 i PRICE_Growth (BRK)



residuals i PRICE (BRK)



residuals i price growth (BRK)

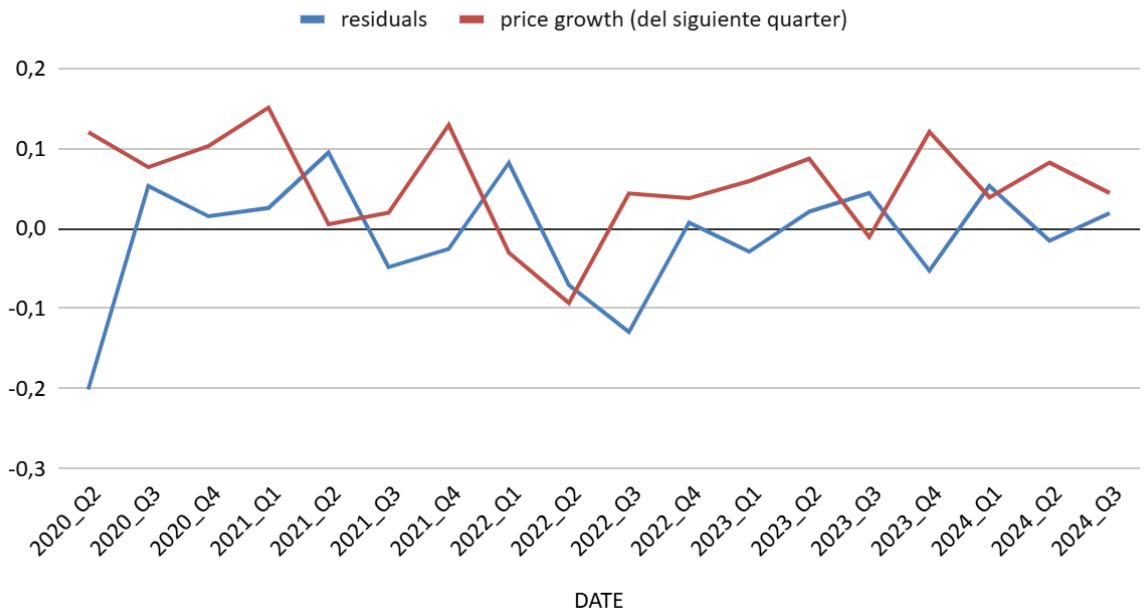
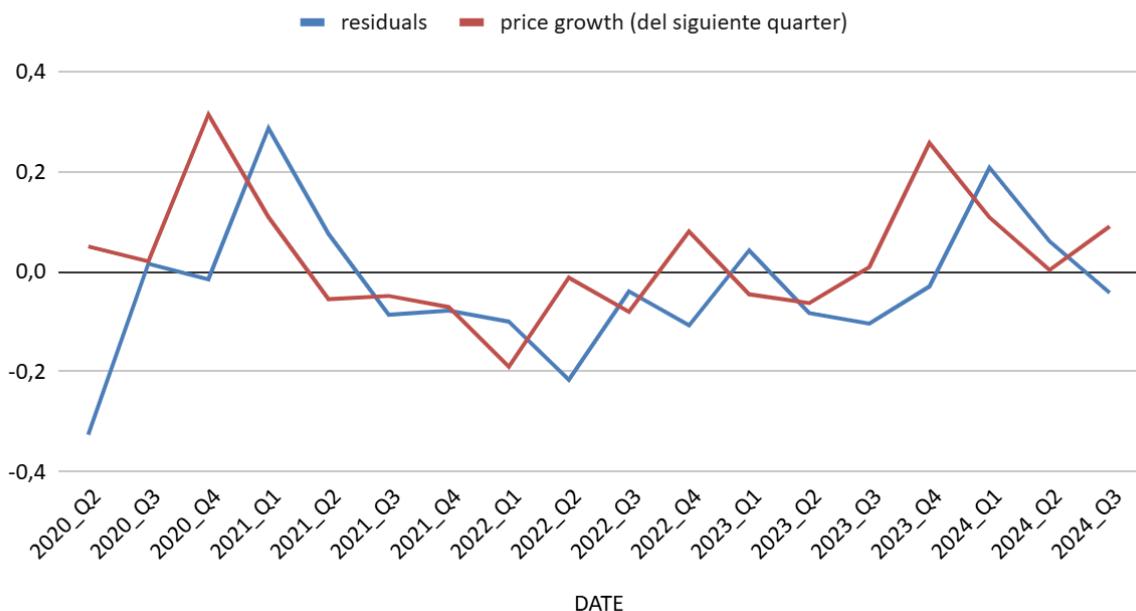


Figura A11. *Citigroup*

residuals i price growth (C)



C

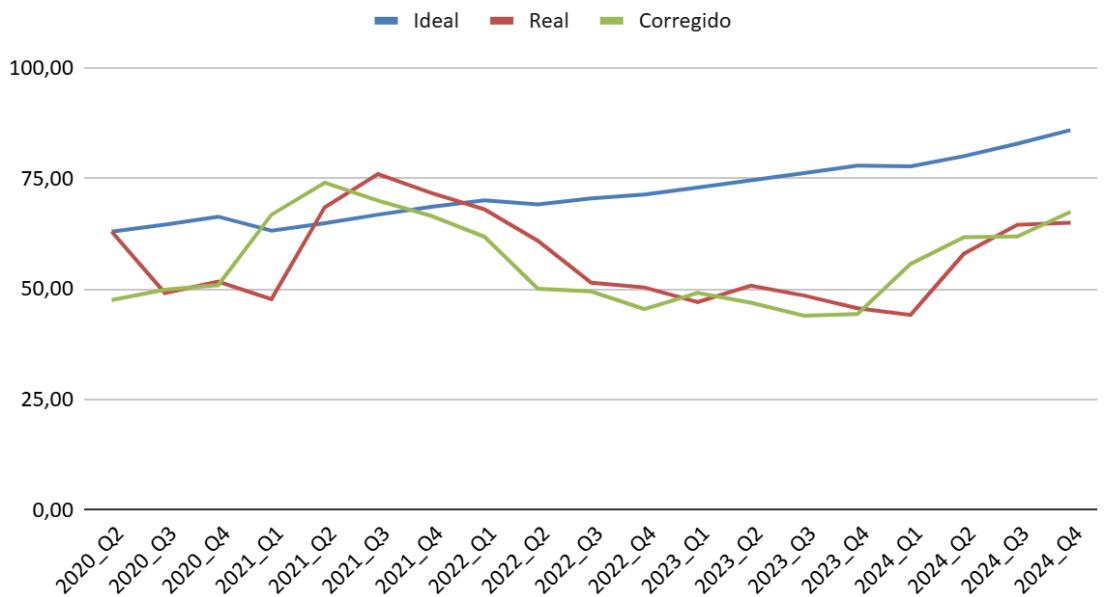
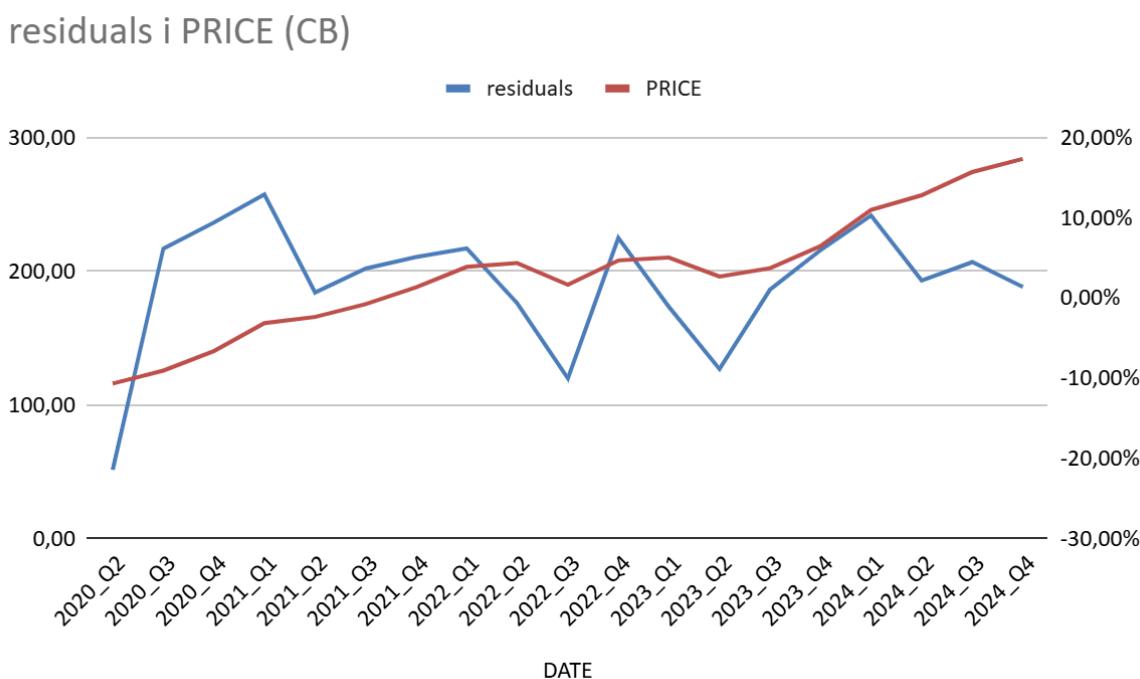
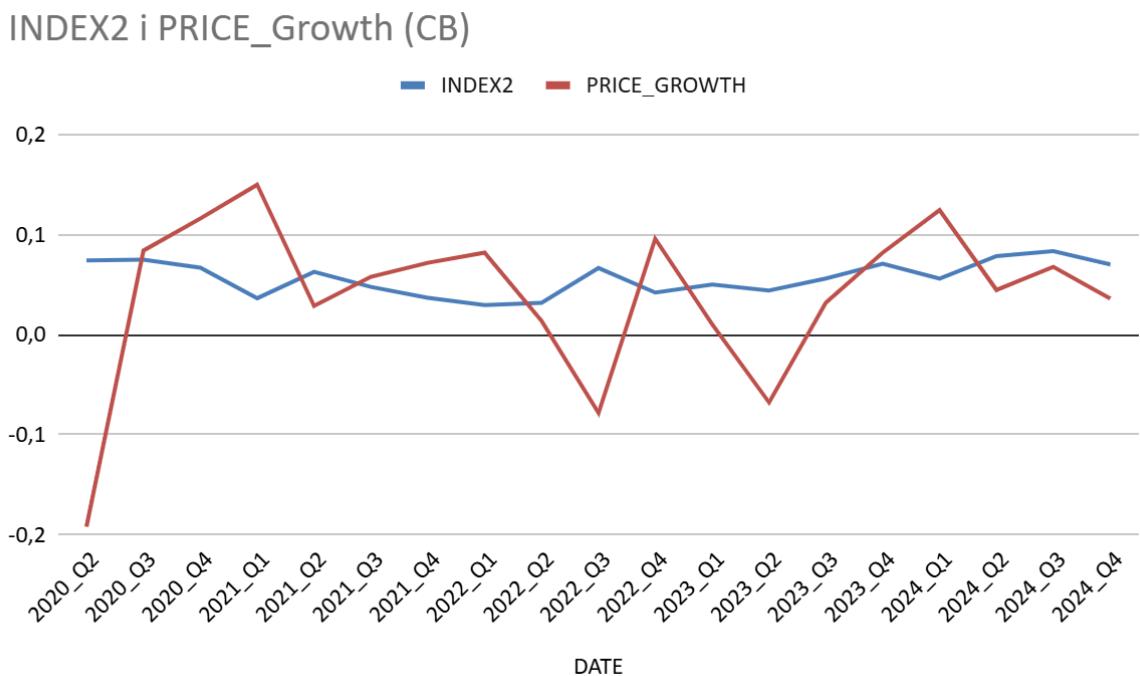


Figura A12. Chubb LTD

residuals i price growth (CB)

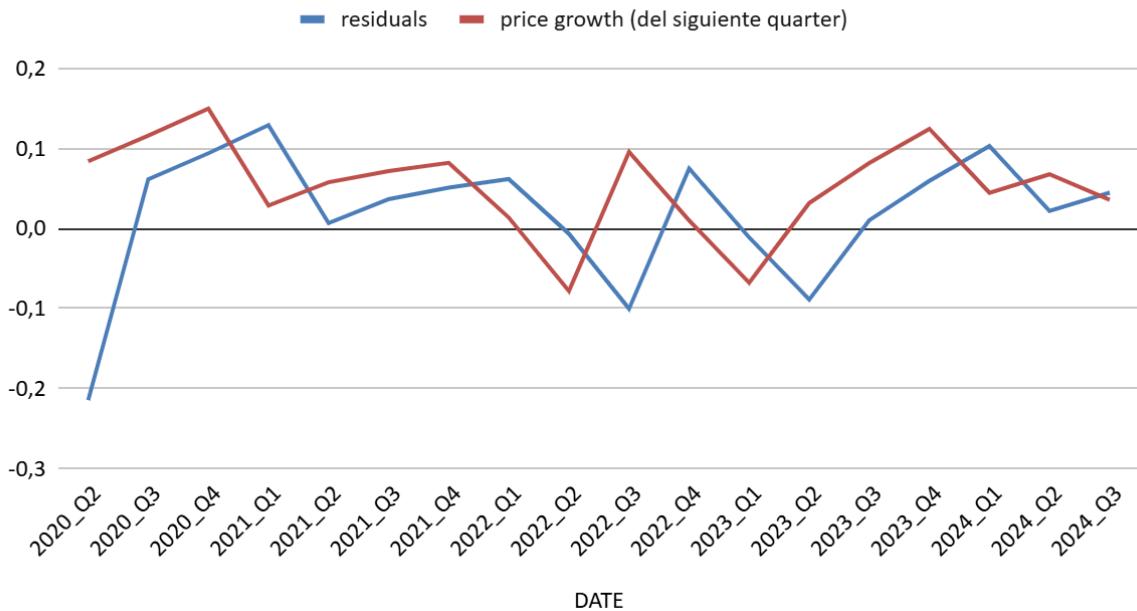
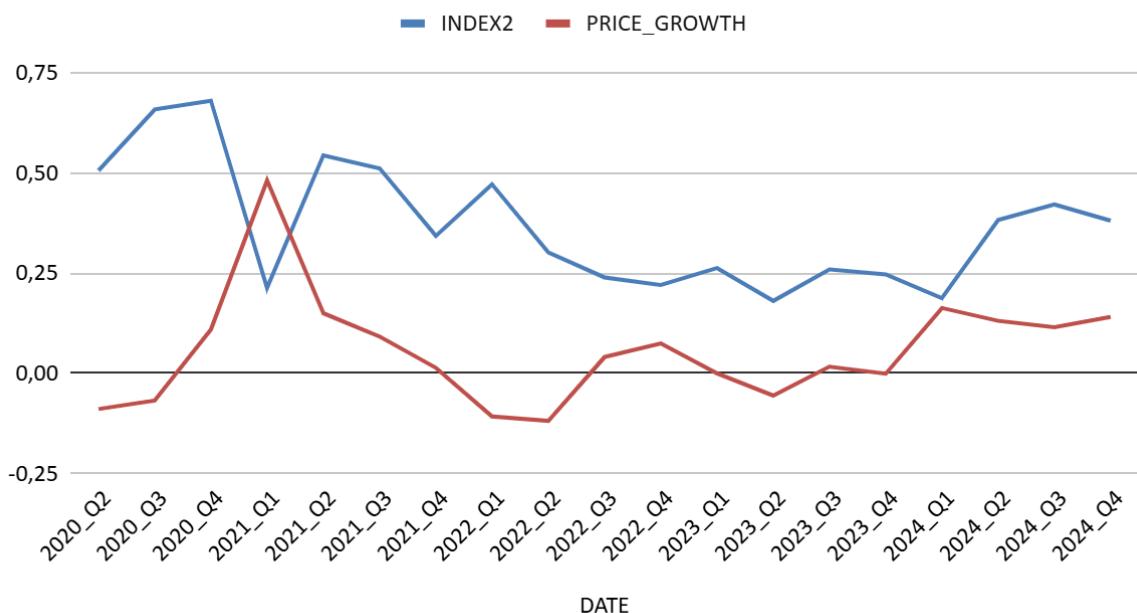


Figura A13. Goldman Sachs

INDEX2 i PRICE_Growth (GS)



residuals i PRICE (GS)



residuals i price growth (GS)



Figura A14. JP Morgan

residuals i price growth (JPM)

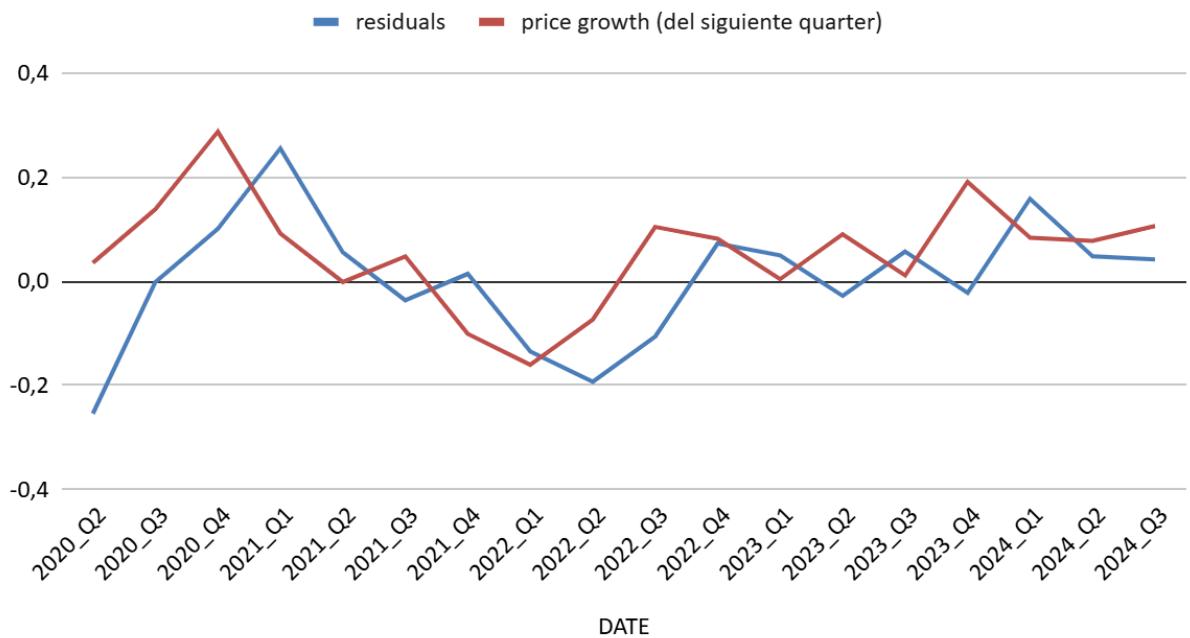
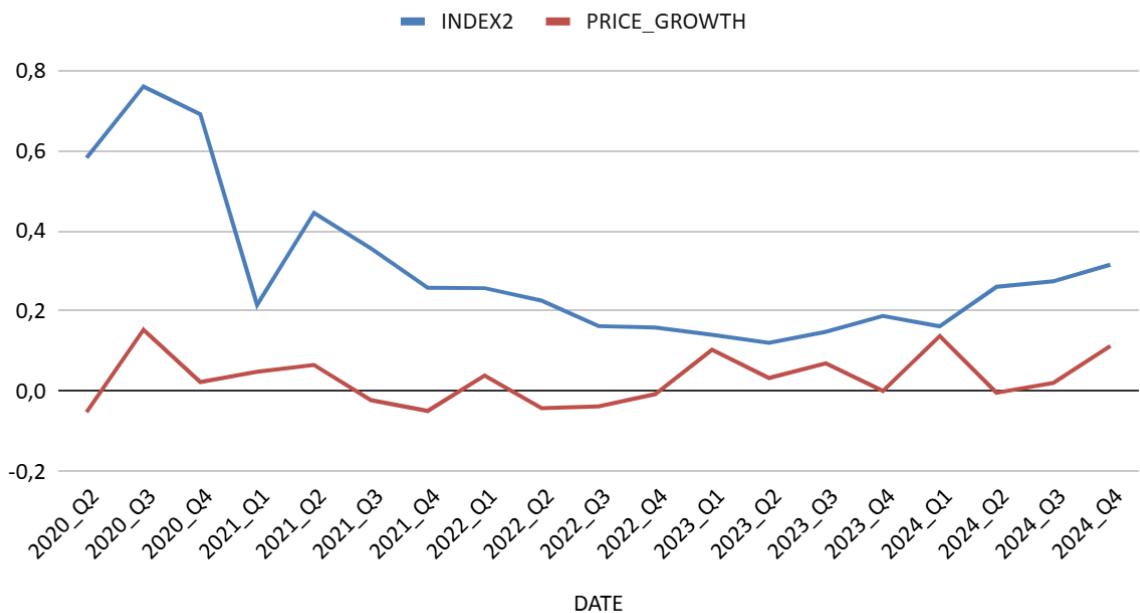
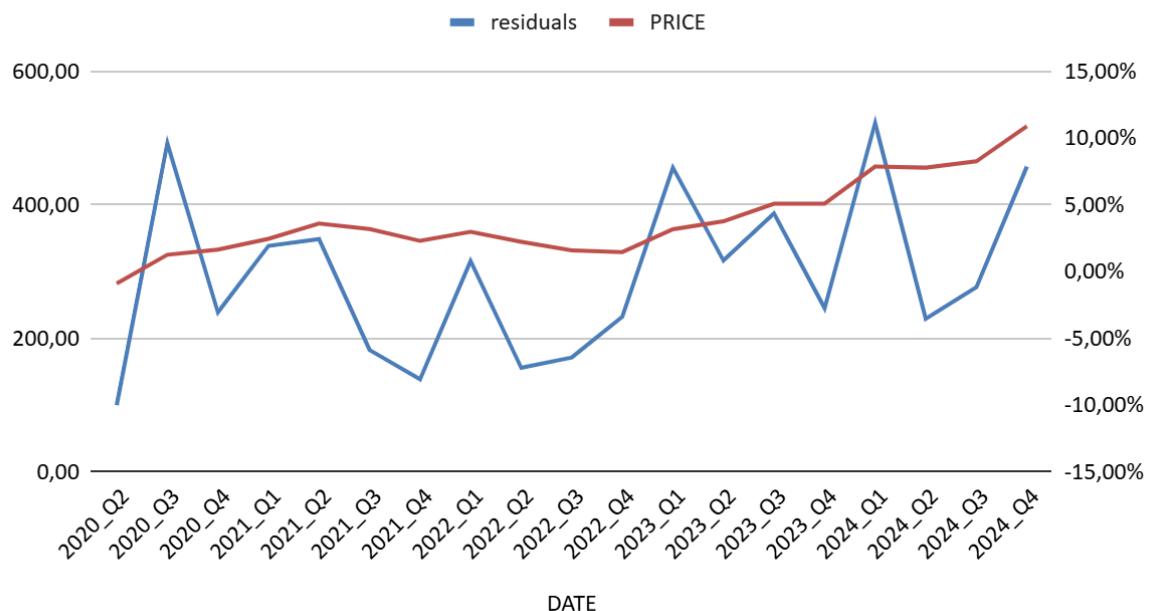


Figura A15. Mastercard

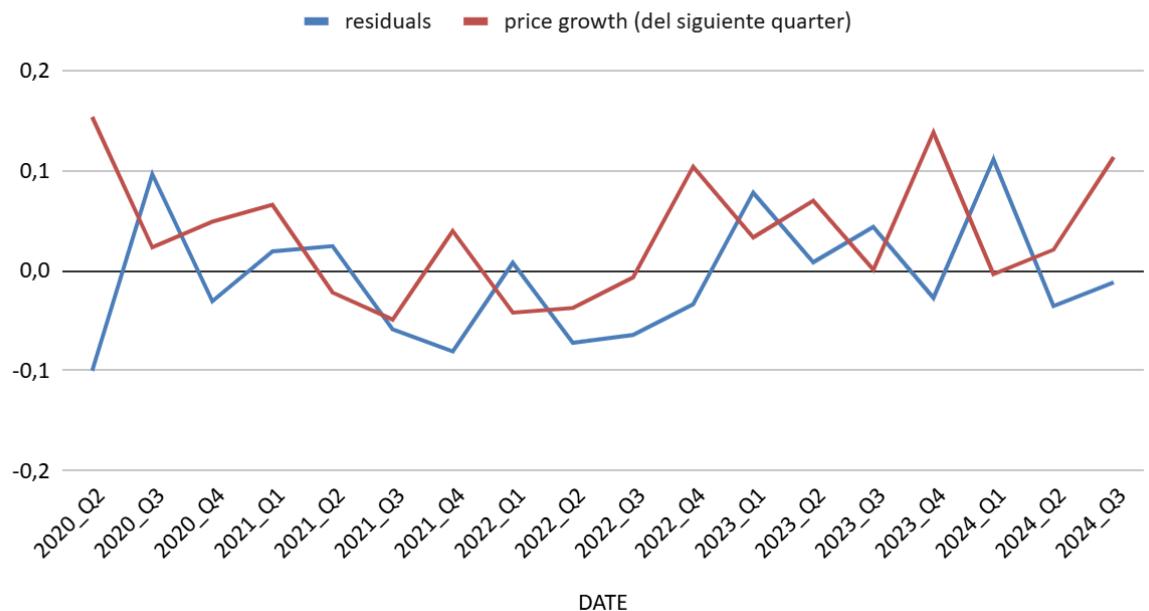
INDEX2 i PRICE_Growth (MA)



residuals i PRICE (MA)



residuals i price growth (MA)



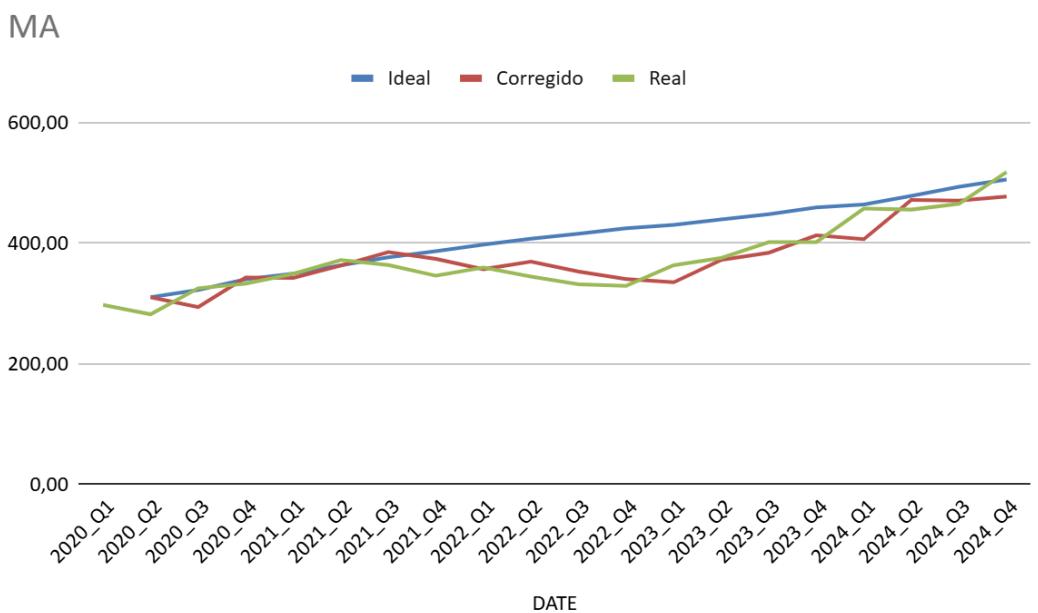
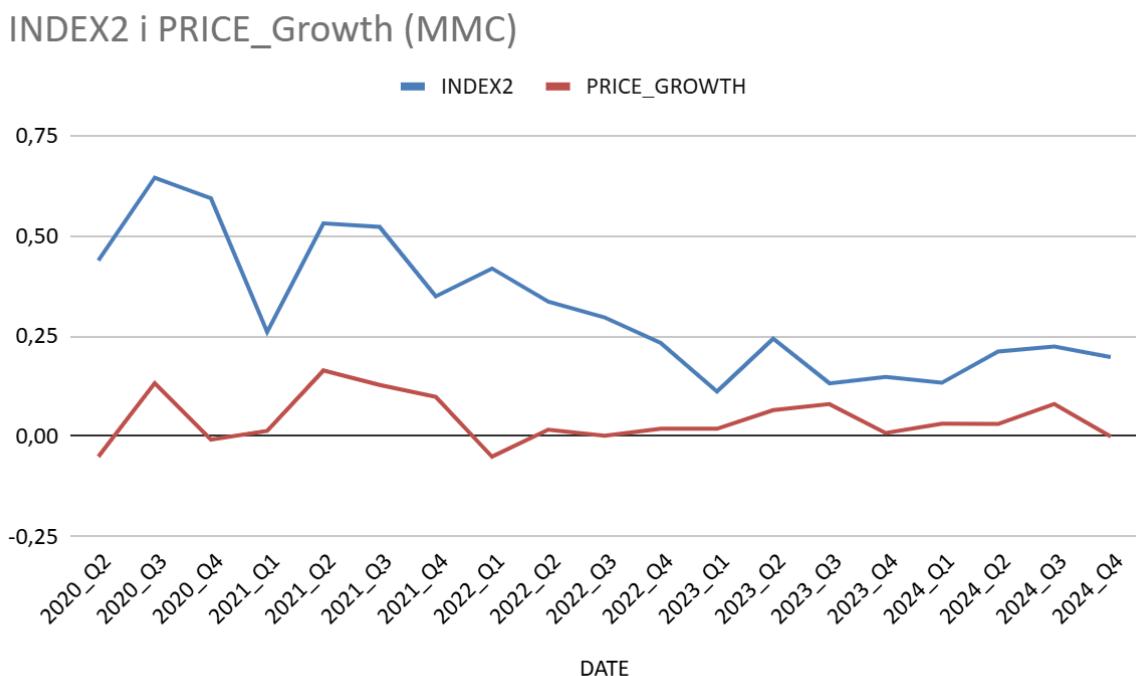


Figura A16. Marsh & McLennan Companies Inc



residuals i PRICE (MMC)



residuals i price growth (MMC)

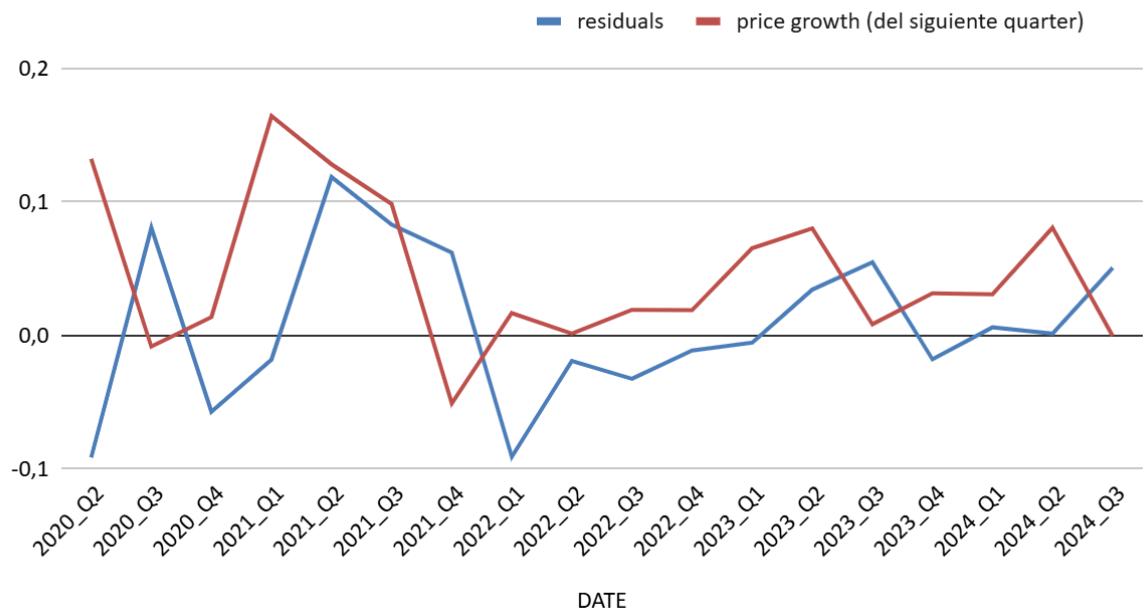
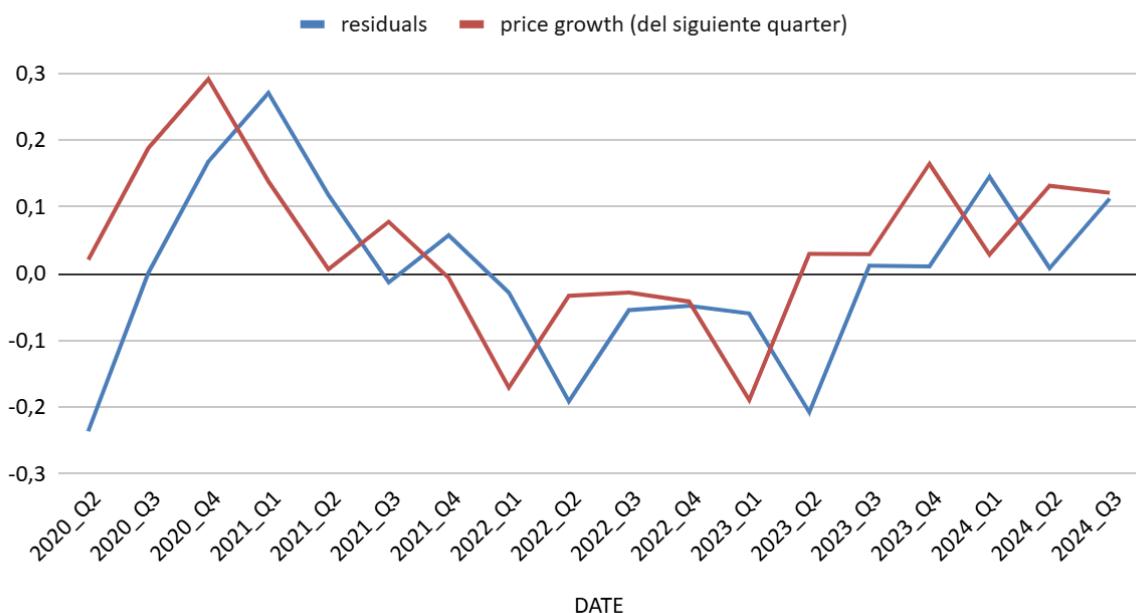
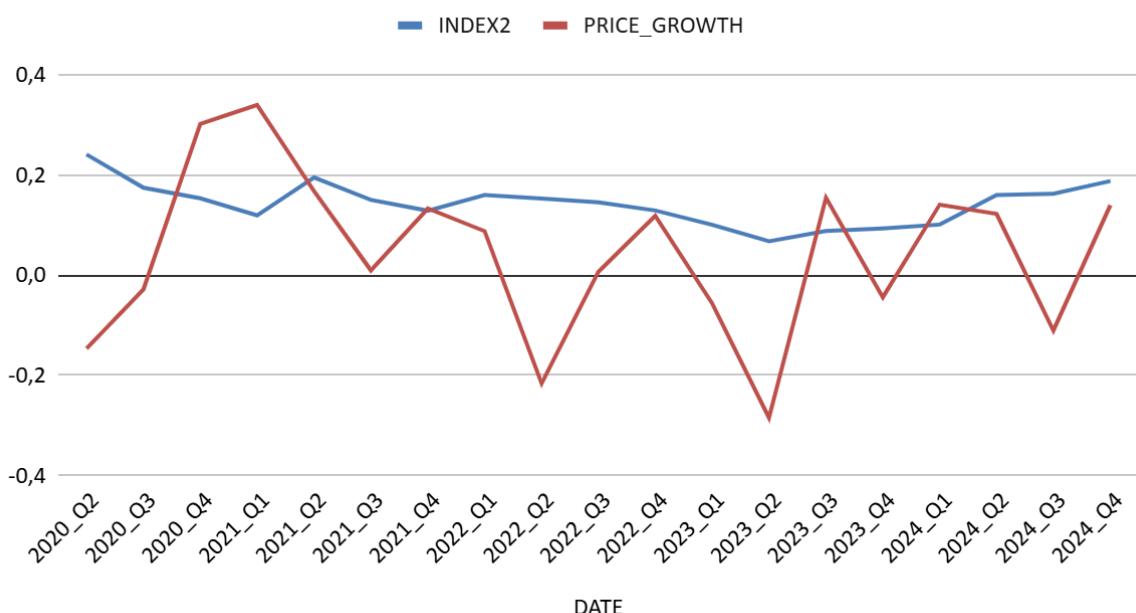


Figura A17. PNC Financial Services

residuals i price growth (PNC)

**Figura A18.** Charles Schwab

INDEX2 i PRICE_Growth (SCHW)



residuals i PRICE (SCHW)



residuals i price growth (SCHW)

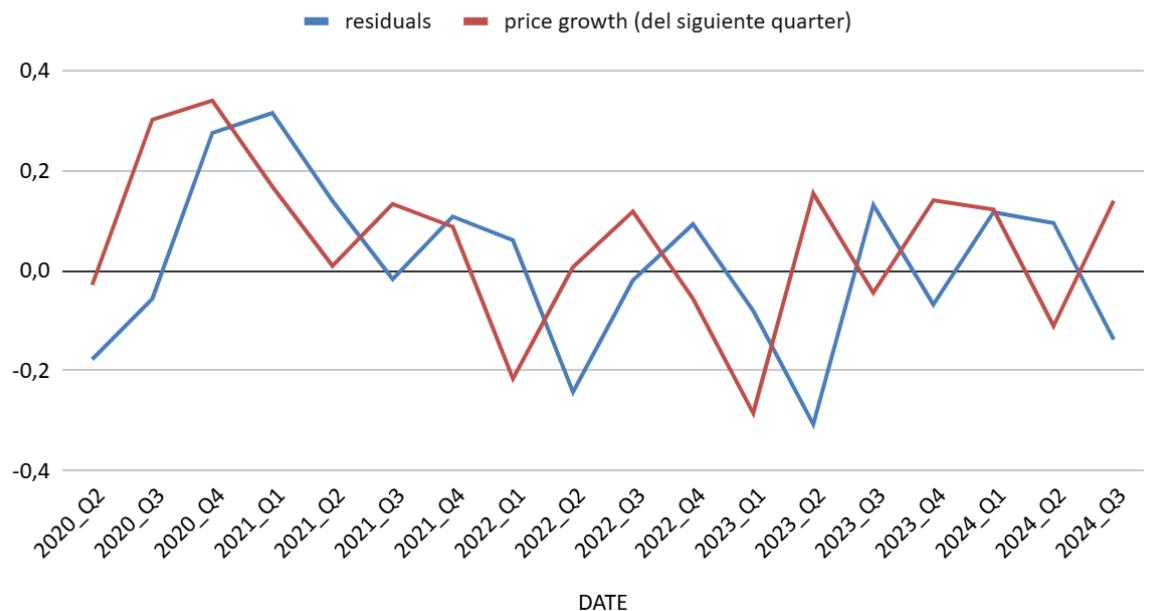
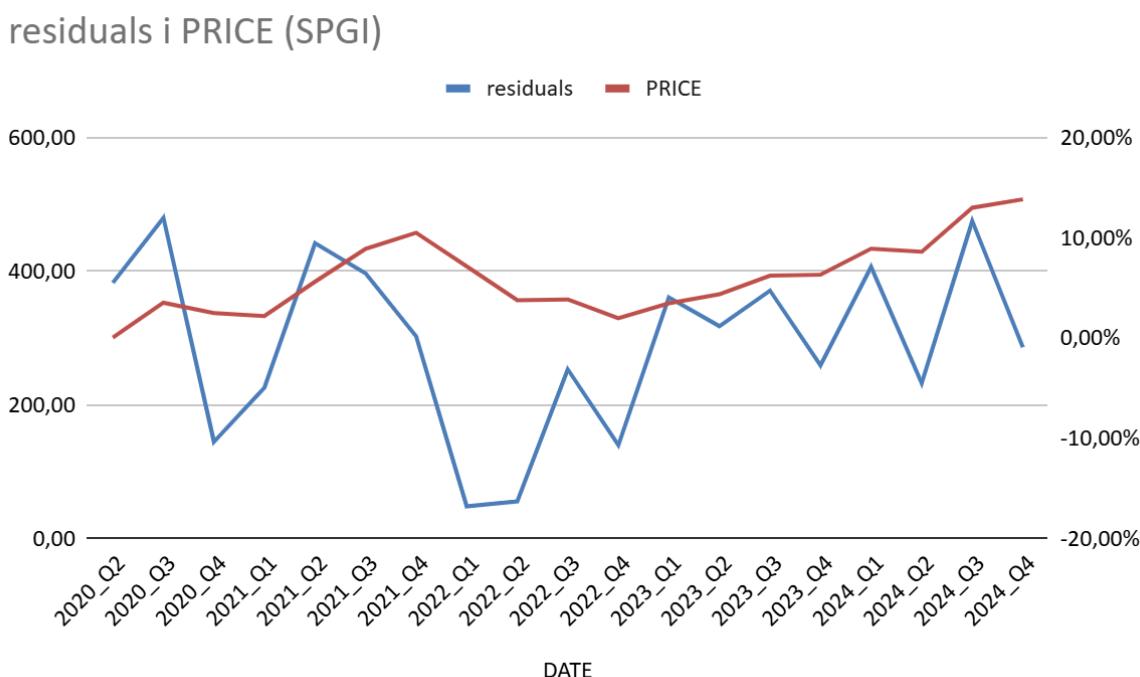
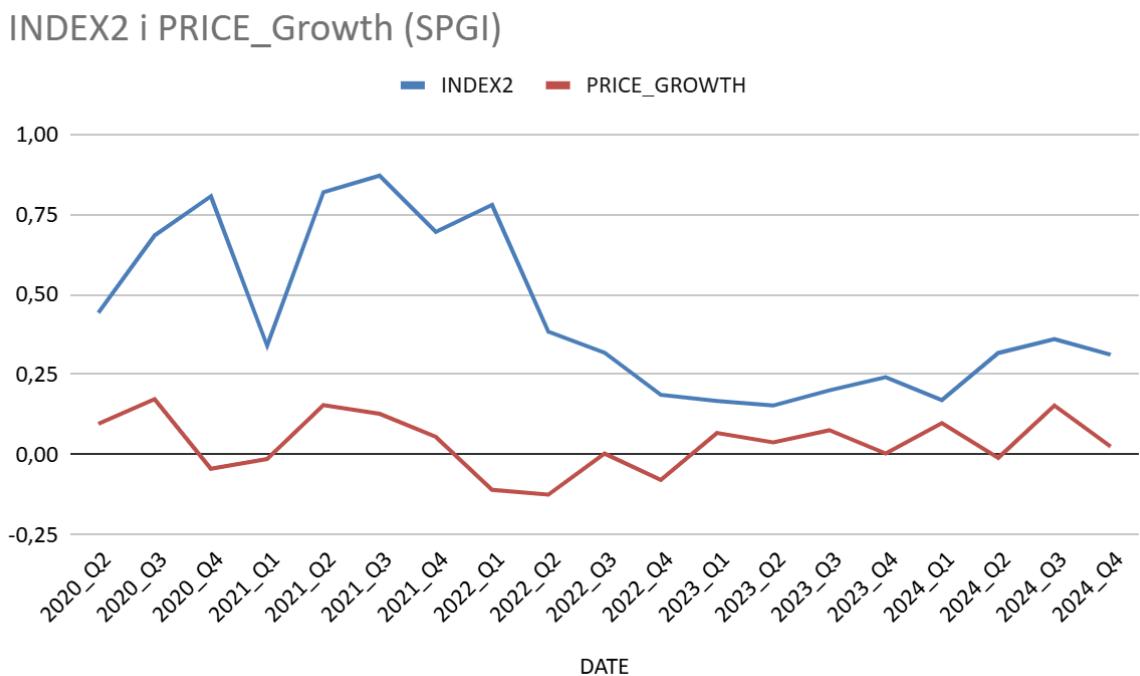


Figura A19. SPGI global

residuals i price growth (SPGI)

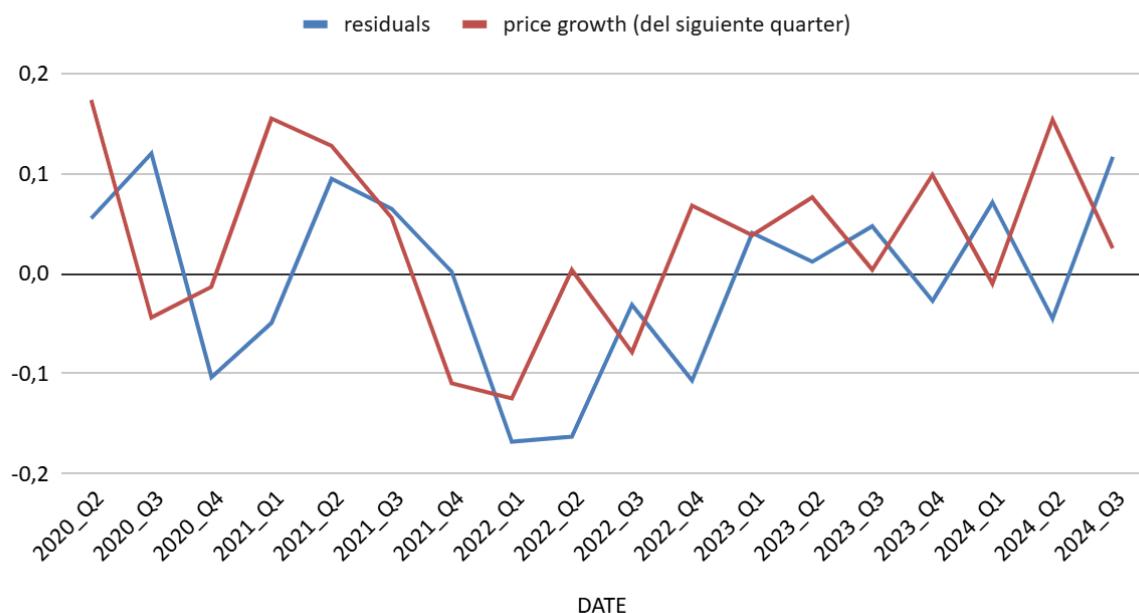
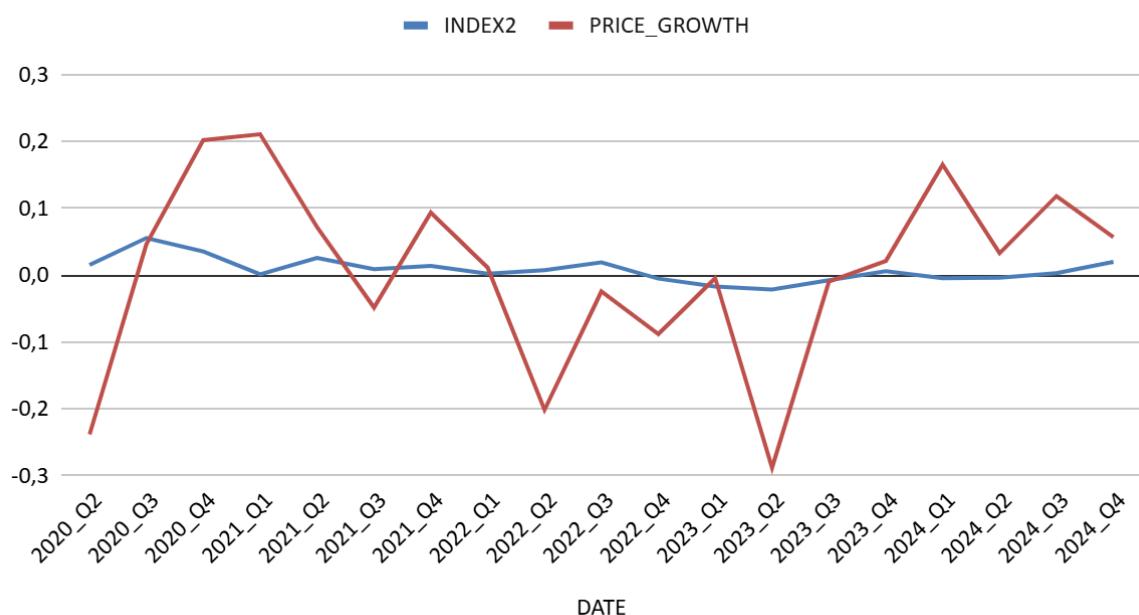


Figura A20. Trust Financial Corp

INDEX2 i PRICE_Growth (TFC)



residuals i PRICE (TFC)



residuals i price growth (TFC)

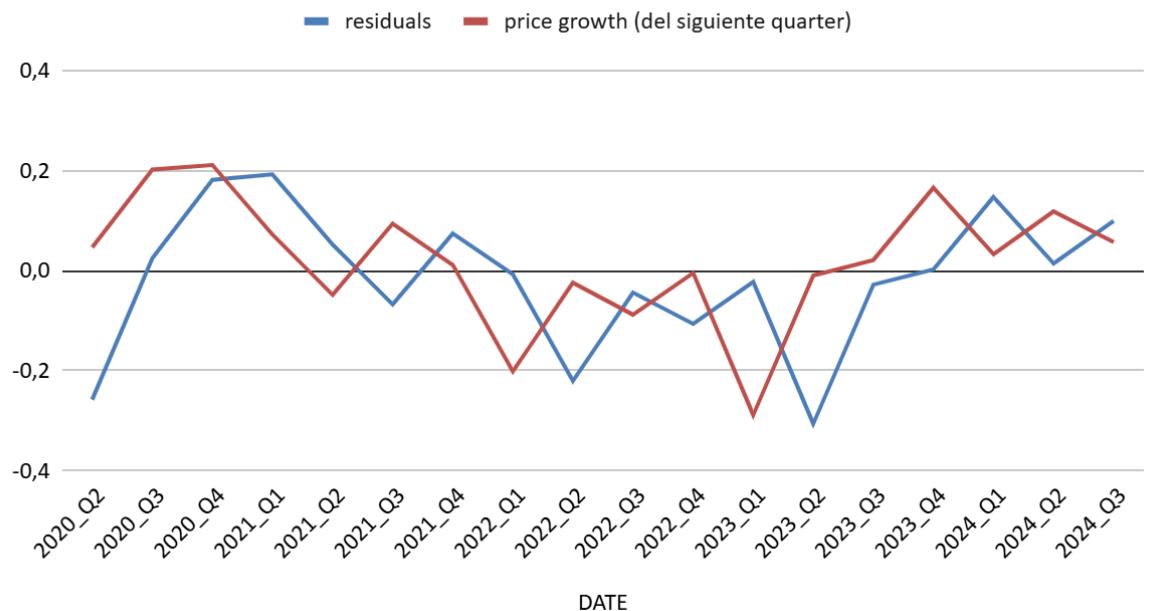


Figura A21. US Bancorp

residuals i PRICE (USB)



residuals i price growth (USB)

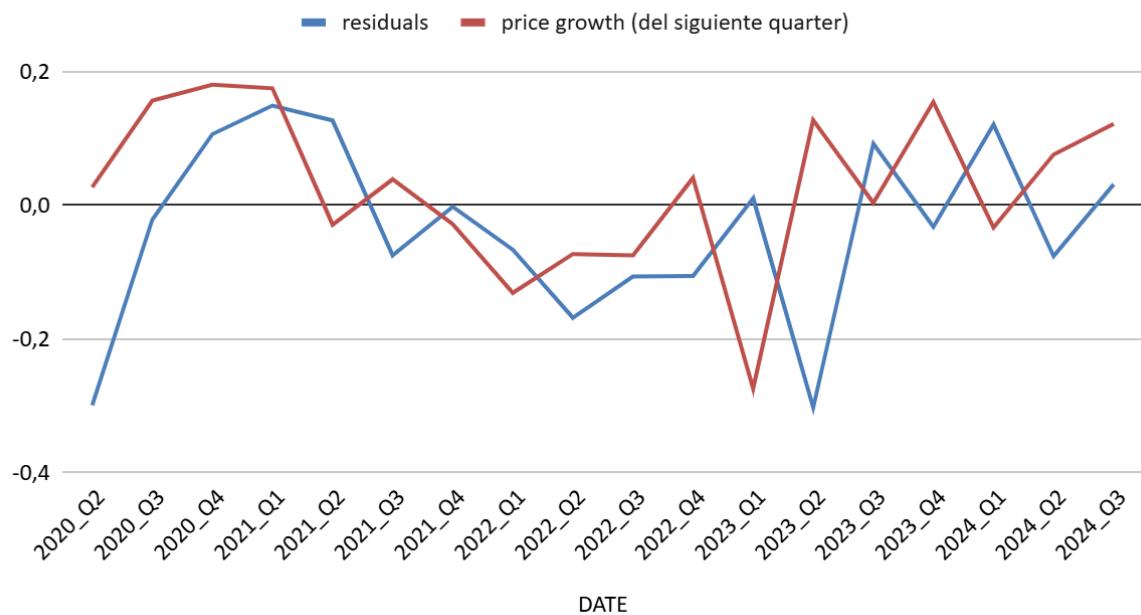
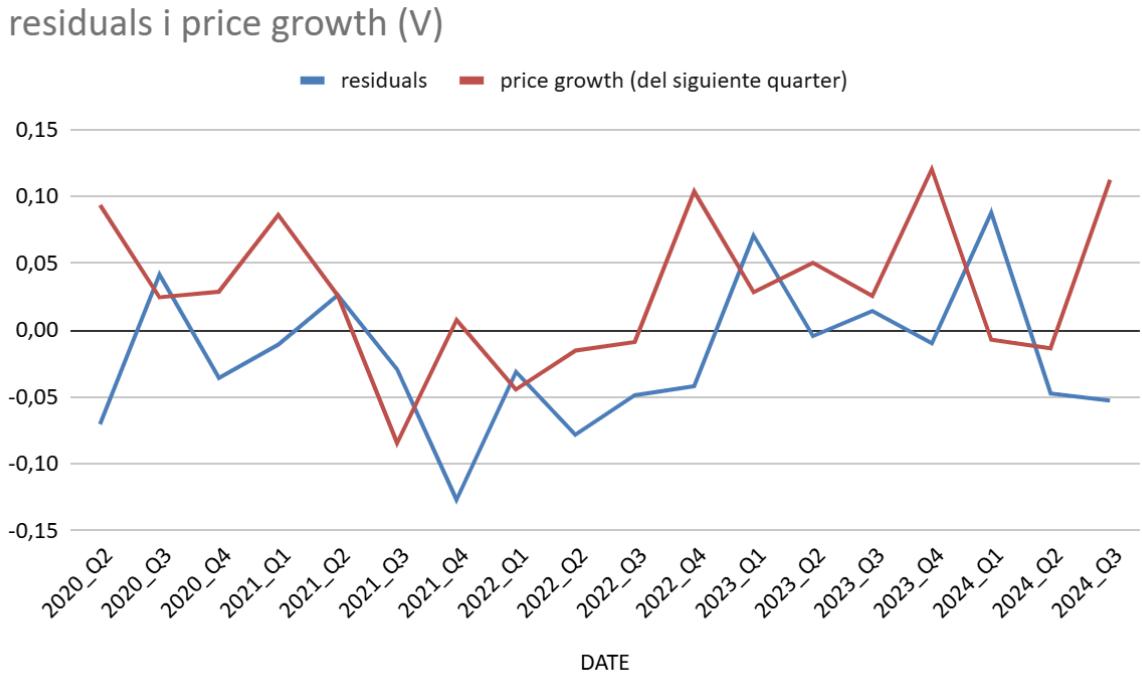
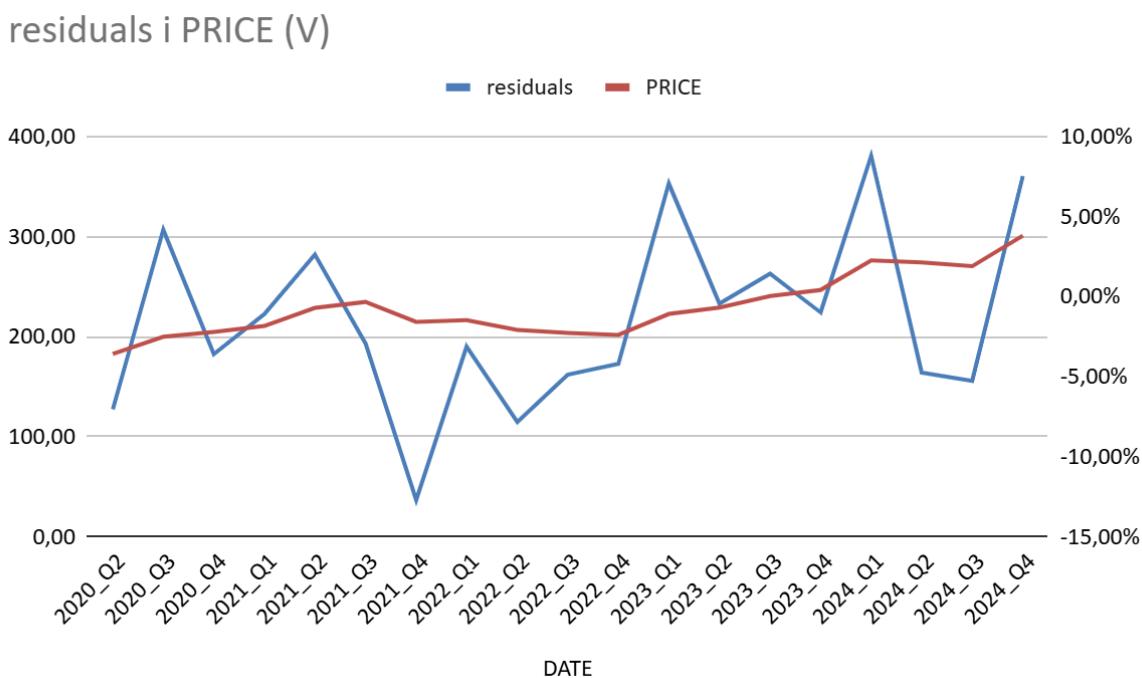


Figura 22. Visa

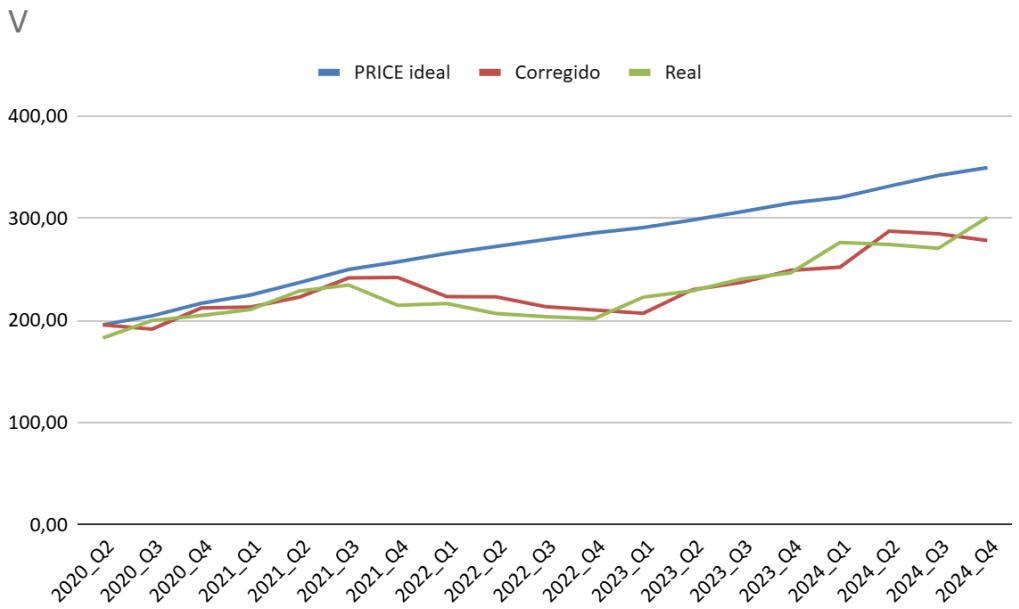
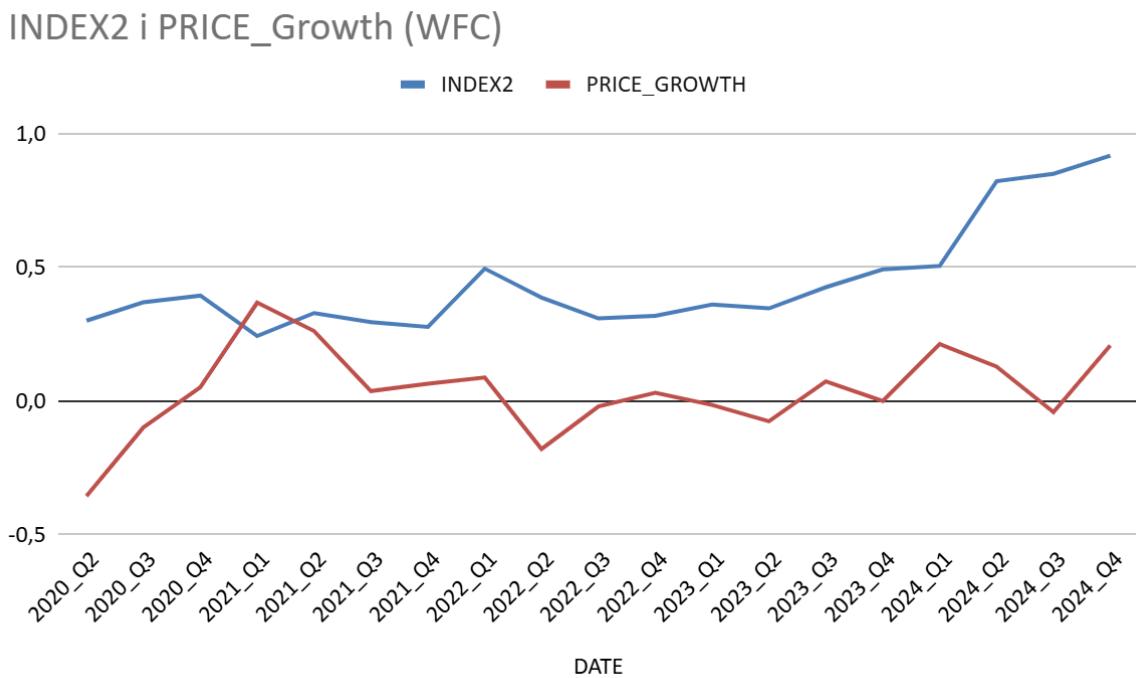


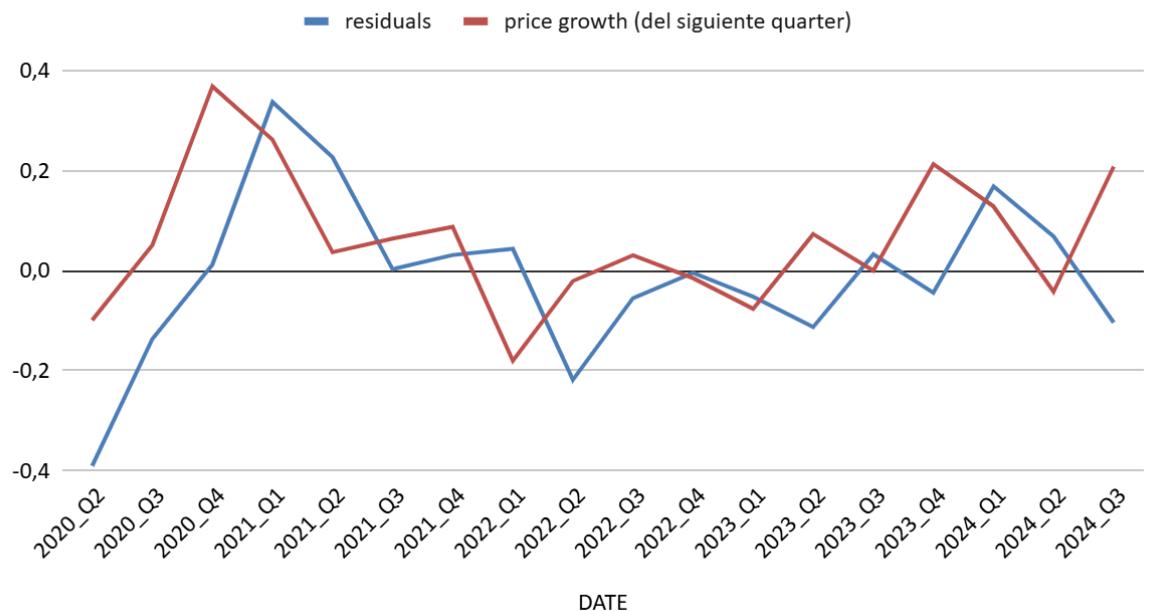
Figura A23. Wells Fargo Company



residuals i PRICE (WFC)



residuals i price growth (WFC)



Codigo Completo:

```

import pandas as pd

import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LassoCV
from sklearn.model_selection import train_test_split
from statsmodels.stats.outliers_influence import
variance_inflation_factor
import statsmodels.api as sm
from statsmodels.stats.stattools import durbin_watson
from linearmodels.panel import PanelOLS
from skopt import gp_minimize
from skopt.space import Real
from skopt.utils import use_named_args
from sklearn.metrics import mean_squared_error, mean_absolute_error,
mean_absolute_percentage_error
from scipy.stats import skew, kurtosis, shapiro, anderson
from statsmodels.stats.diagnostic import het_breuschpagan, het_white,
acorr_ljungbox
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LinearRegression

# Cargar los archivos Excel
fundamentales = pd.read_excel('fundamentales tfg.xlsx')
var_control = pd.read_excel('var control tfg.xlsx')

# Convertir TICKER y DATE a categórico
fundamentales['TICKER'] = pd.Categorical(fundamentales['TICKER'])
fundamentales['DATE'] = pd.Categorical(fundamentales['DATE'])
var_control['DATE'] = pd.Categorical(var_control['DATE'])

# Fusionar los DataFrames en DATE
data = pd.merge(fundamentales, var_control, on='DATE', how='left')

# Eliminar espacios adicionales en los nombres de las columnas
data.columns = data.columns.str.strip()

# Fix DATE data type after merge
data['DATE'] = pd.Categorical(data['DATE'])

```

```

# Sort the data by TICKER and DATE to ensure correct order for return
calculation
data = data.sort_values(['TICKER', 'DATE'])

# Calcular el crecimiento de PRICE entre trimestres para cada empresa y
aplicar shift(-1)
data['PRICE_Growth'] = data.groupby('TICKER',
observed=True) ['PRICE'].pct_change()

# Calcular la correlación entre todas las variables y PRICE_Growth
antes de estandarizarlas
correlation_before_scaling =
data.corr(numeric_only=True) ['PRICE_Growth'].dropna()

# Mostrar la correlación de cada variable con PRICE_Growth
print("Correlación con PRICE_Growth antes de estandarizar:")
print(correlation_before_scaling)

# Exportar la correlación a un archivo Excel
correlation_before_scaling.to_frame(name='Correlation').to_excel('corre
lation_with_price_growth_before_scaling.xlsx')

print("Columnas disponibles en el DataFrame:")
print(data.columns)

# Lista de variables a estandarizar
variables_a_estandarizar = ['CASH', 'DEBT', 'EPS', 'FCFMG',
'NETPROFITMG',
'OPERATINGMG', 'QUICK', 'ROA', 'ROE',
'EMPLOYEES', 'REVENUES']

# Aplicar Min-Max Scaling para cada trimestre (DATE)
scaler = MinMaxScaler()
data[variables_a_estandarizar] = data.groupby('DATE',
observed=True) [variables_a_estandarizar].transform(
    lambda x: scaler.fit_transform(x.values.reshape(-1, 1)).flatten()
)

# Excluir Q1_2025 después de calcular PRICE_Growth y estandarizar
data = data[data['DATE'] != 'Q1_2025']

```

```

data[variables_a_estandarizar] = data.groupby('TICKER',
observed=True)[variables_a_estandarizar].shift(1)

# Definir los pesos iniciales para cada variable
weights = {
    'CASH': 0.2,
    'DEBT': -1,
    'EPS': 0.2,
    'FCFMG': 0.2,
    'NETPROFITMG': 0.25,
    'OPERATINGMG': 0.25,
    'QUICK': 0.2,
    'ROA': 0.15,
    'ROE': 0.15,
    'EMPLOYEES': 0.2,
    'REVENUES': 0.2
}

# Verificar que los pesos sumen 1
if not np.isclose(sum(weights.values()), 1):
    raise ValueError("Los pesos no suman 1. Por favor, ajusta los
valores.")

# Filtrar los pesos iniciales para que coincidan con las columnas en
variables_a_estandarizar
filtered_weights_initial = {key: weights[key] for key in
variables_a_estandarizar if key in weights}

# Calcular el índice inicial para cada empresa y trimestre
data['INDEX'] =
data[filtered_weights_initial.keys()].dot(pd.Series(filtered_weights_in
itial))

# Eliminar espacios adicionales en los nombres de las columnas
data.columns = data.columns.str.strip()

# Definir los grupos de variables
MACRO_VARS = ['Interest rates', 'Real GDP Growth', 'Exchange Rates',
              'Credit spreads', 'Unemployment rate']

INDICATOR_VARS = ['VIX', 'XLF', 'SPY']

```

```

FEATURE_VARS = ['VOLUME', 'SHARES']

# Filtrar las columnas de cada grupo que están presentes en el
DataFrame
MACRO_VARS = [var for var in MACRO_VARS if var in data.columns]
INDICATOR_VARS = [var for var in INDICATOR_VARS if var in data.columns]
FEATURE_VARS = [var for var in FEATURE_VARS if var in data.columns]

# Función para calcular log-differences
def log_difference(series):
    # Manejar valores negativos sumando un desplazamiento positivo
    min_value = series.min()
    if (min_value <= 0):
        series += abs(min_value) + 1e-6 # Desplazar para que todos los
valores sean positivos
    return np.log(series).diff()

# Estandarizar MACRO_VARS usando log-differences, excepto 'Real GDP
Growth' y 'Interest rates'
if MACRO_VARS:
    for var in MACRO_VARS:
        if var == 'Interest rates':
            # Calcular el cambio porcentual para Interest rates
            data[var] = data.groupby('TICKER',
observed=True)[var].pct_change()
        elif var != 'Real GDP Growth': # Excluir 'Real GDP Growth' del
cálculo de log-differences
            # Calcular log-differences para las demás variables
            data[var] = data.groupby('TICKER',
observed=True)[var].transform(log_difference)
else:
    print("No se encontraron columnas para MACRO_VARS en el
DataFrame.")

# Estandarizar INDICATOR_VARS
if INDICATOR_VARS:
    for var in INDICATOR_VARS:
        data[var] = log_difference(data[var])
else:

```

```

print("No se encontraron columnas para INDICATOR_VARS en el
DataFrame.")

# Estandarizar FEATURE_VARS (incluyendo SHARES) usando log-differences
dentro de cada empresa (por TICKER)
if FEATURE_VARS:
    for var in FEATURE_VARS:
        if var == 'SHARES':
            # Calcular log-differences para SHARES
            data['SHARES'] = data.groupby('TICKER',
observed=True) ['SHARES'].transform(log_difference)

            # Aplicar Min-Max Scaling por periodo (DATE) después de
log-differences
            scaler_shares = MinMaxScaler()
            data['SHARES'] = data.groupby('DATE',
observed=True) ['SHARES'].transform(
                lambda x:
scaler_shares.fit_transform(x.values.reshape(-1, 1)).flatten()
            )
        else:
            # Estandarizar las demás variables de FEATURE_VARS usando
log-differences
            data[var] = data.groupby('TICKER',
observed=True) [var].transform(log_difference)
    else:
        print("No se encontraron columnas para FEATURE_VARS en el
DataFrame.")

# Asegurarse de que DATE no tenga espacios adicionales
data['DATE'] = data['DATE'].astype(str).str.strip()

# Definir las variables independientes (X) y la variable dependiente
(y)
X_vars = ['INDEX'] + MACRO_VARS + INDICATOR_VARS + FEATURE_VARS
y_var = 'PRICE_Growth'

# Verificar que las variables estén en el DataFrame
if not X_vars:
    raise ValueError("No se encontraron variables independientes en el
DataFrame.")

```

```

if y_var not in data.columns:
    raise ValueError(f"La variable dependiente '{y_var}' no se
encuentra en el DataFrame.")

# Eliminar filas con NaN en las variables seleccionadas
data = data.dropna(subset=X_vars + [y_var])

# Definir X (independientes) y y (dependiente)
X = data[X_vars]
y = data[y_var]

# Calcular la correlación entre las variables independientes y
PRICE_Growth
correlation_data = data[X_vars + [y_var]].corr().loc[y_var].drop(y_var)

# Mostrar la correlación de cada variable con PRICE_Growth
print("Correlación con PRICE_Growth:")
print(correlation_data)

# Exportar la correlación a un archivo Excel
correlation_data.to_frame(name='Correlation').to_excel('correlation_with_price_growth.xlsx')

# Dividir los datos en conjuntos de entrenamiento y prueba
X = data[X_vars]
y = data[y_var]

# Calcular el Variance Inflation Factor (VIF) para las variables
independientes
def calculate_vif(X):
    vif_data = pd.DataFrame()
    vif_data["Feature"] = X.columns
    vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in
range(X.shape[1])]
    return vif_data

# Iteratively eliminate variables with VIF > 10
while True:
    vif_data = calculate_vif(X)
    print("Variance Inflation Factor (VIF):")
    print(vif_data)

```

```

# Export the VIF data to an Excel file
vif_data.to_excel('vif_data.xlsx', index=False)

# Identify variables with VIF > 10
high_vif_features = vif_data[vif_data["VIF"] >
10]["Feature"].tolist()
if not high_vif_features:
    break # Exit the loop if no variables have VIF > 10

print(f"Removing variables with VIF > 10: {high_vif_features}")
X = X.drop(columns=high_vif_features)

# Asegurarse de que TICKER y DATE sean índices para el modelo de panel
data = data.set_index(['TICKER', 'DATE'])

# Verificar que las variables estén en el DataFrame
if not X_vars:
    raise ValueError("No se encontraron variables independientes en el
DataFrame.")
if y_var not in data.columns:
    raise ValueError(f"La variable dependiente '{y_var}' no se
encuentra en el DataFrame.")

# Eliminar filas con NaN en las variables seleccionadas
data = data.dropna(subset=X_vars + [y_var])

# Calcular el Variance Inflation Factor (VIF) para las variables
independientes
def calculate_vif(X):
    vif_data = pd.DataFrame()
    vif_data["Feature"] = X.columns
    vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in
range(X.shape[1])]
    return vif_data

# Dividir los datos en conjuntos de entrenamiento y prueba después de
eliminar variables con alto VIF
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

```

```

# Aplicar Lasso con las variables restantes
lasso = LassoCV(cv=5, random_state=42).fit(X_train, y_train)

# Mostrar los coeficientes de las variables seleccionadas
selected_features = pd.DataFrame({'Feature': X.columns, 'Coefficient': lasso.coef_})
print("Coeficientes de Lasso:")
print(selected_features)

# Filtrar las variables con coeficientes diferentes de 0
selected_features = selected_features[selected_features['Coefficient'] != 0]
selected_vars = selected_features['Feature'].tolist()
print(f"Variables seleccionadas después de Lasso: {selected_vars}")

# Redefinir X con las variables seleccionadas
X = X[selected_vars]

# Agregar una constante para el término independiente
X = sm.add_constant(X)

# Ajustar el modelo OLS
model = sm.OLS(y, X).fit(cov_type='HC3') # Usar errores estándar
robustos para heterocedasticidad

# Mostrar los resultados de la regresión
print(model.summary())

# Calcular el Mean Squared Error (MSE)
mse = np.mean((model.resid) ** 2)
print(f"Mean Squared Error (MSE): {mse:.4f}")

# Obtener los coeficientes, R2 y R2 ajustado
coefficients = model.params
r_squared = model.rsquared
adjusted_r_squared = model.rsquared_adj

print("\nCoefficients:")
print(coefficients)
print(f"\nR2: {r_squared:.4f}")
print(f"Adjusted R2: {adjusted_r_squared:.4f}")

```

```

# Exportar los coeficientes de Lasso a un archivo Excel
selected_features.to_excel('lasso_coefficients.xlsx', index=False)

# Exportar los resultados de la regresión OLS a un archivo Excel
coefficients.to_excel('ols_coefficients.xlsx', index=True)

# Exportar el índice inicial a un archivo Excel
data.reset_index(inplace=True)
data[['TICKER', 'DATE', 'INDEX']].to_excel('initial_index.xlsx',
index=False)

# Parámetro alpha para ponderar MSE y correlación
alpha = 0.25 # Ajusta este valor según la importancia relativa de mse
y correlation

data = data.dropna(subset=variables_a_estandarizar + [y_var]) # Eliminar filas con NaN
y = data[y_var] # Actualizar y con los datos filtrados

def objective_function(**new_weights):
    # Verificar que los pesos estén dentro de los rangos (para depuración)
    for key, value in new_weights.items():
        if key == 'DEBT':
            if not (-0.99 <= value <= -0.01):
                print(f"Advertencia: {key}={value} fuera del rango [-0.99, -0.01], ajustando...")
                new_weights[key] = max(-0.99, min(-0.01, value))
        else:
            if not (0.01 <= value <= 0.99):
                print(f"Advertencia: {key}={value} fuera del rango [0.01, 0.99], ajustando...")
                new_weights[key] = max(0.01, min(0.99, value))

    # Calcular la suma total de los pesos
    total_weight = sum(new_weights.values())
    if total_weight == 0: # Evitar división por cero
        print("Advertencia: Suma de pesos es 0, devolviendo valor muy negativo")
        return -1e10 # Valor muy negativo en lugar de -inf

```

```

# Normalizar los pesos para que sumen 1
normalized_weights = {key: value / total_weight for key, value in
new_weights.items()}

# Calcular el índice con los nuevos pesos
data['INDEX2'] =
data[variables_a_estandarizar].dot(pd.Series(normalized_weights))

# Actualizar X con el nuevo INDEX2
X_vars_adjusted = [var for var in selected_vars if var != 'INDEX']
+ ['INDEX2']
X_new = sm.add_constant(data[X_vars_adjusted])

# Ajustar el modelo OLS con el nuevo X_new
try:
    model = sm.OLS(y, X_new).fit(cov_type='HC3')
    mse = mean_squared_error(y, model.predict(X_new))
except Exception as e:
    print(f"Error en OLS: {e}, devolviendo valor muy negativo")
    return -1e10 # Valor muy negativo en lugar de -inf

# Calcular la correlación entre 'INDEX2' y 'PRICE_Growth'
correlation = data[['INDEX2', 'PRICE_Growth']].corr().iloc[0, 1]
if pd.isna(correlation) or np.isinf(correlation):
    print(f"Correlación inválida: {correlation}, devolviendo valor
muy negativo")
    return -1e10 # Valor muy negativo en lugar de -inf

# Mostrar los valores calculados en esta iteración
print(f"MSE: {mse:.7f}")
print(f"Correlación entre INDEX2 y PRICE_Growth:
{correlation:.4f}")

# Estandarizar el MSE a una escala de [0, 1]
mse_score = 1 / (1 + mse)
if np.isinf(mse_score) or np.isnan(mse_score):
    print(f"MSE_score inválido: {mse_score}, devolviendo valor muy
negativo")
    return -1e10*9999 # Valor muy negativo en lugar de -inf

```

```

# Penalizar pesos negativos en fundamentales que deberían ser
positivos
positive_vars = ['ROE', 'EPS', 'NETPROFITMG', 'OPERATINGMG',
'QUICK', 'CASH', 'REVENUES']
penalty = sum(max(0, -normalized_weights.get(key, 0)) for key in
positive_vars if key in normalized_weights) * 0.5
print(f"Penalización por pesos negativos: {penalty:.4f}")

# Regularización L2 para los pesos
l2_penalty = 0.01 * sum(w ** 2 for w in
normalized_weights.values())
print(f"Penalización L2: {l2_penalty:.4f}")

# Calcular la métrica a optimizar
metric = alpha * mse_score + (1 - alpha) * correlation - penalty -
l2_penalty
return metric

# Crear los límites para cada peso
pbounds = {
    key: (-0.99, -0.01) if key == 'DEBT' else (0.01, 0.99)
    for key in weights.keys()
}

# Inicializar la optimización bayesiana
from bayes_opt import BayesianOptimization

optimizer = BayesianOptimization(
    f=objective_function,
    pbounds=pbounds,
    verbose=2,
    random_state=42,
)

# Definir los pesos iniciales como una copia de weights*
initial_weights = weights.copy()

# Establecer los pesos iniciales como punto de partida
optimizer.probe(params=initial_weights, lazy=True)

# Ejecutar la optimización

```

```

optimizer.maximize(init_points=40, n_iter=400)

# Obtener los mejores pesos
best_weights = optimizer.max['params']

# Normalizar los mejores pesos para que sumen 1 y respeten los límites
# de -1 a 1
total_weight = sum(best_weights.values())
if total_weight != 0: # Evitar división por cero
    best_weights = {key: max(-1, min(1, value / total_weight)) for key,
value in best_weights.items()}
else:
    print("Suma de pesos optimizados es 0, no se puede normalizar")
    best_weights = weights # Usar los pesos iniciales como respaldo

print("Mejores pesos encontrados:", best_weights)

# Filtrar los pesos optimizados para que coincidan con las columnas en
variables_a_estandarizar
filtered_weights = {key: best_weights[key] for key in
variables_a_estandarizar if key in best_weights}

# Recalcular el índice con los pesos filtrados
data['INDEX2'] =
data[filtered_weights.keys()].dot(pd.Series(filtered_weights))

# Exportar el índice optimizado a un archivo Excel
data.reset_index(inplace=True)
data[['TICKER', 'DATE', 'INDEX2',
y_var]].to_excel('optimized_index.xlsx', index=False)

# Recalcular la correlación y el MSE con el nuevo INDEX2
X_vars_adjusted = [var for var in selected_vars if var != 'INDEX'] +
['INDEX2']
X_new = sm.add_constant(data[X_vars_adjusted])

model = sm.OLS(y, X_new).fit(cov_type='HC3')
optimized_correlation = data[['INDEX2', 'PRICE_Growth']].corr().iloc[0,
1]
final_mse = mean_squared_error(y, model.predict(X_new))

```

```

print(model.summary())
print(f"\nCorrelación optimizada entre INDEX2 y PRICE_Growth:
{optimized_correlation:.6f}")
print(f"Mean Squared Error (MSE) después de la optimización:
{final_mse:.8f}")

# Obtener los coeficientes, R2 y R2 ajustado
coefficients = model.params
r_squared = model.rsquared
adjusted_r_squared = model.rsquared_adj

print("\nCoefficients:")
print(coefficients)
print(f"\nR2: {r_squared:.4f}")
print(f"Adjusted R2: {adjusted_r_squared:.4f}")

# Validación Cruzada
model_sk = LinearRegression().fit(X_new, y)
cv_scores = cross_val_score(model_sk, X_new, y, cv=5,
scoring='neg_mean_squared_error')
cv_mse_mean = -cv_scores.mean()
cv_mse_std = cv_scores.std()

print(f"\nCross-Validation Metrics:")
print(f"Cross-Validated MSE (Mean): {cv_mse_mean:.7f}")
print(f"Cross-Validated MSE (Std Dev): {cv_mse_std:.7f}")

# Predicción Fuera de Muestra
X_train, X_test, y_train, y_test = train_test_split(X_new, y,
test_size=0.2, random_state=42)
model_train = sm.OLS(y_train,
sm.add_constant(X_train)).fit(cov_type='HC3')
y_pred_test = model_train.predict(sm.add_constant(X_test))
mse_test = mean_squared_error(y_test, y_pred_test)
rmse_test = np.sqrt(mse_test)
mae_test = mean_absolute_error(y_test, y_pred_test)

print(f"\nTest Set Metrics (Out-of-Sample):")
print(f"Test MSE: {mse_test:.7f}")
print(f"Test RMSE: {rmse_test:.7f}")
print(f"Test MAE: {mae_test:.7f}")

```

```

# Calcular el VIF de las variables de la regresión 1 con el índice
# optimizado
vif_data_optimized = calculate_vif(X_new.drop(columns=['const'])) # Excluir la constante para VIF
print("\nVariance Inflation Factor (VIF) después de optimizar INDEX2:")
print(vif_data_optimized)
vif_data_optimized.to_excel('vif_data_optimized.xlsx', index=False)

# Ensure y and INDEX2 are properly aligned and extracted from data
y = data['PRICE_Growth'] # Dependent variable
X = data[['INDEX2']] # Independent variable as a DataFrame
X = sm.add_constant(X) # Add a constant term for the intercept

# Fit the OLS model
model = sm.OLS(y, X).fit(cov_type='HC3')

# Calculate MSE
mse_index = mean_squared_error(y, model.predict(X))

# Print results
print("\nResultados del modelo OLS con INDEX2 como variable
independiente (PRICE_Growth):")
print("-----")
print(model.summary())
print(f"Mean Squared Error (MSE) between INDEX2 and PRICE_Growth:
{mse_index:.4f}")
data['residuals'] = model.resid # Guardar los residuos en el DataFrame

# 1. Basic Residual Statistics
residuals_mean = data['residuals'].mean()
residuals_std = data['residuals'].std()
residuals_skew = skew(data['residuals'].dropna())
residuals_kurtosis = kurtosis(data['residuals'].dropna())
dw_stat = durbin_watson(model.resid)

print(f"\nBasic Residual Statistics:")
print(f"Mean of Residuals: {residuals_mean:.7f} (Ideal: close to 0)")
print(f"Standard Deviation of Residuals: {residuals_std:.7f}")
print(f"Skewness of Residuals: {residuals_skew:.4f} (Ideal: close to
0)")

```

```

print(f"Kurtosis of Residuals: {residuals_kurtosis:.4f} (Ideal: close
to 3 for normal distribution)")
print(f"Durbin-Watson Statistic: {dw_stat:.4f} (Ideal: close to 2,
indicates no autocorrelation)")

# 2. Advanced Residual Diagnostics
# Residual Standard Error (RSE)
rse = np.sqrt(np.sum(model.resid ** 2) / (model.df_resid))

# Normality Tests
shapiro_test, shapiro_p = shapiro(model.resid)
anderson_result = anderson(model.resid)
anderson_stat = anderson_result.statistic
anderson_critical_value_5 = anderson_result.critical_values[2] # 5%
significance level

print(f"\nAdvanced Residual Diagnostics:")
print(f"Residual Standard Error (RSE): {rse:.7f} (Measures typical
prediction error)")
print(f"Shapiro-Wilk Test (p-value): {shapiro_p:.7f} (p > 0.05
indicates normality)")
print(f"Anderson-Darling Statistic: {anderson_stat:.7f} (Normality if <
{anderson_critical_value_5} at 5%)")

# 4. Residual Influence and Outliers
influence = model.get_influence()
cooks_d = influence.cooks_distance[0]
leverage = influence.hat_matrix_diag

print(f"\nResidual Influence and Outliers:")
print(f"Max Cook's Distance: {np.max(cooks_d):.7f} (Values > 1 suggest
influential points)")
print(f"Max Leverage Value: {np.max(leverage):.7f} (High values
indicate potential outliers)")

# 5. Additional Error Metrics for Second Regression
y_pred = model.predict(X)
rmse_index = np.sqrt(mse_index)
mae_index = mean_absolute_error(y, y_pred)

print(f"\nAdditional Error Metrics for Second Regression:")

```

```

print(f"Root Mean Squared Error (RMSE): {rmse_index:.7f}")
print(f"Mean Absolute Error (MAE): {mae_index:.7f}")

# 6. Export Residuals and Diagnostics
residuals_df = pd.DataFrame({
    'Residuals': model.resid,
    'Fitted_Values': model.fittedvalues,
    'Cook_Distance': cooks_d,
    'Leverage': leverage
})
residuals_df.to_excel('residuals_diagnosticos.xlsx', index=False)

print(f"\nResiduals and diagnostics exported to
'residuals_diagnosticos.xlsx'")
print(f"Residual summary exported to 'residuals_summary.xlsx'")
print(max('INDEX2'))

# Exportar los mejores pesos a un archivo Excel
pd.DataFrame.from_dict(best_weights, orient='index',
columns=['Weight']).to_excel('optimized_weights_bayesian.xlsx')

# Exportar el DataFrame con el indice optimizado
data.to_excel('data_FINAL.xlsx', index=False)

# Asegurarse de que TICKER y DATE estén como columnas antes de exportar
data.reset_index(inplace=True)

# Opcional: Guardar los residuos con TICKER y DATE en un archivo Excel
data[['TICKER', 'DATE',
'residuals']].to_excel('residuals_by_ticker_date.xlsx', index=False)

```