Modelos de Machine Learning aplicados a la proyección del precio WTI del petróleo.

Universidad de los Andes
Maestría en Economía Aplicada
Andres Felipe Martinez, Angela Paola Morales Guio y Oscar Cortes

Repositorio: https://github.com/paolamguio/Proyecto_Final

Resumen

Considerando la importancia del petróleo en la economía mundial, así como la volatilidad de su precio de referencia, este trabajo analiza la predicción el precio internacional del petróleo WTI (West Texas Intermediate), con el fin de orientar la toma de decisiones de gobiernos e inversionistas. Se propone un modelo que incluye variables de google trends para el pronóstico un paso adelante, para el desarrollo del modelo se usan datos de producción y precios del petróleo, datos de producción y precios del gas natural. Se utilizan modelos univariados ARIMA, GARCH y T-GARCH para estimar la media y la varianza de las series; y se utilizan modelos multivariados Elastic Net, destilación por componentes principales y NNET autorregresivo para la regularización. Se prueban diferentes configuraciones de los modelos univariados y multivariados para la predicción del precio WTI, utilizando el MSE para la comparación y selección de la combinación más precisa. Se analiza también el efecto de incluir variables de google trend para el mejoramiento de las predicciones. Los resultados empíricos muestran un buen comportamiento del modelo seleccionado con bajos RMSE para la predicción.

Palabras claves: predicción precio petróleo; volatilidad; modelos univariados; modelos multivariados; RMSE.

I. INTRODUCCIÓN

El incremento en el precio del petróleo ha acentuado el problema inflacionario que afecta a muchas de las economías del mundo, aumento que sólo beneficia a unos pocos países cuyas economías se concentran en la producción y exportación y, perjudica a muchos otros importadores de este recurso natural, de igual forma, la variación de los precios del petróleo ha convulsionado la economía a nivel internacional en años pasados. Es por esto, que el objetivo de este trabajo es encontrar el mejor modelo para pronosticar una eventual subida o caída del precio del petróleo, que permitiera tomar medidas para mitigar o aprovechar dicha situación, debido a la alta dependencia de la economía del país para la generación de regalías y así la financiación de programas sociales. Por otra parte, el país aprobó la política de transición energética (DNP, 2022), donde se identifica el gran potencial que tiene el país en energías renovables como la eólica, solar, y geotérmica, así como en hidrógeno, por lo cual, la predicción de los precios del petróleo apoyaría las decisiones en el cambio que se ha planteado el Gobierno Nacional en los próximos 8 años.

Algunos estudios han utilizado métodos de Big Data aplicados a series de tiempo para predecir el precio del petróleo, Yu et al (2019) y Choi y Shin (2022) utilizan como predictores búsquedas relacionadas con petróleo en Google Trends, encontrando que los modelos que incluyen estas variables producen mejores pronósticos para el petróleo. En Colombia, investigadores de la Universidad de Antioquia y la Universidad de Eafit propusieron un modelo basado en redes neuronales para el pronóstico del precio internacional del petróleo utilizando estructuras de redes con datos de series de precios diarios de petróleo WTI (West Texas Intermediate), el índice S&P500 y índice del dólar estadounidense DXY.

En este trabajo se presenta un modelo de predicción del precio del petróleo WTI a partir de series entre los

años 1989 a 2020 de los precios y producción del petróleo WTI y el precio y producción del gas natural. Adicionalmente, como novedad, el modelo cuenta con series variables de google trends entre los años 2004 y 2020 para su calibración, lo que repercute en una mayor precisión en los resultados de predicción.

Este documento se compone de cinco secciones incluida esta introducción. La segunda sección enumera y detalla las variables utilizadas para el entrenamiento del modelo y como se limpiaron para este fin. La tercera sección describe las metodologías utilizadas para la elaboración del modelo, mientras que en la cuarta sección se presentan los resultados arrojados por el modelo. Finalmente, en la última sección se resaltan las principales conclusiones y recomendaciones con base al trabajo realizado.

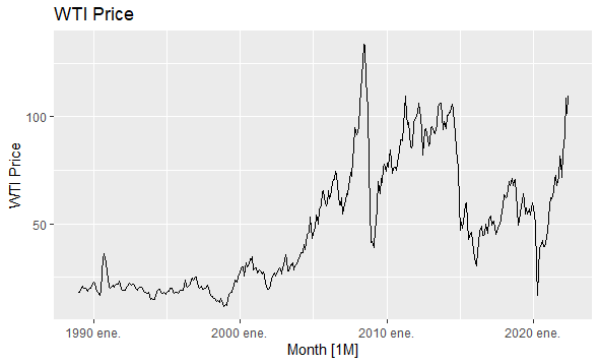
II. DATOS

En línea con el objetivo de este documento, para proyectar precios del petróleo por barril es necesario contar con series de tiempo de las variables que se consideran relevantes e influyentes en la definición del WTI, por lo que se obtuvieron los datos mensuales del precio del petróleo promedio como la variable explicativa principal, junto con otras variables de importancia como lo son: la producción de petróleo en miles de barriles debido a la relación entre la oferta y demanda del crudo como factor fundamental en la formación de precios de este, el precio de importación del Gas Natural y la producción de millones de pies cúbicos de Gas Natural, ya que existe una correlación positiva entre estos dos recursos naturales en particular, porque el gas natural es un derivado del proceso de extracción del petróleo (Lioudis, 2022) [1]. Las anteriores variables se obtuvieron en periodos mensuales de enero 1990 a mayo 2022[2] , lo que permite contar con un total de 401 observaciones por variable.

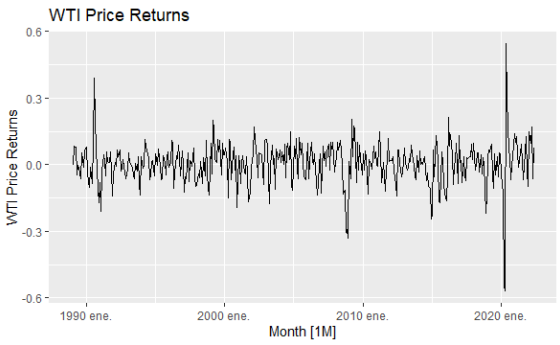
En la tabla 1 del Anexo, se observa que el precio del barril del petróleo (WTI) es una variable con alta volatilidad que ha llegado a presentar un mínimo de 11,35 USD/Barril hasta 133,88 D/Barril en el periodo en análisis, lo que refleja una serie de tiempo no estacionaria, que lleva a suponer que hay diversos factores que influyen en el precio del crudo y la influencia de cada uno de estos no es consistente o linean a lo largo del tiempo, Cen et al, 2018 (ver gráfica 1.a). La volatilidad también se ve reflejada en la producción de petróleo mensual, así como en el precio y producción de Gas Natural.
Debido a la volatilidad de la serie WTI, es necesario transformar la serie a estacionaria por medio de los retornos logarítmicos de esta (ver gráfica 1.b), sin embargo, al observarse outliers en esta última serie, se ajusta la misma por outliers considerando un análisis cuantílico e imputando los outliers identificados (ver gráfica 1.c).

Gráfica 1. Precios Históricos WTI

1.a. Históricos precios WTI (serie original)

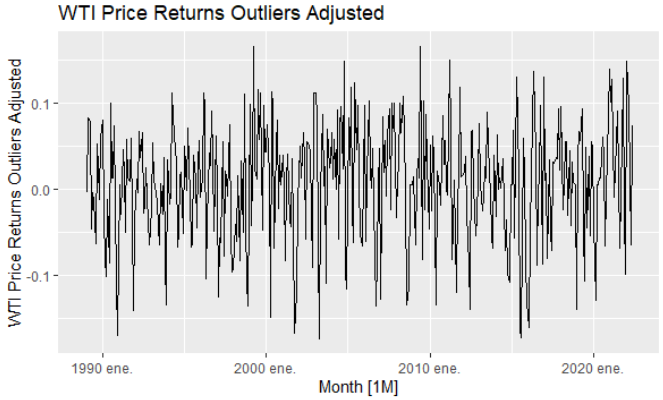


1.b. Históricos precios WTI (retornos logarítmicos)



[1] La relación entre el Petróleo y el Gas Natural es una relación positiva, pero limitada, el coeficiente de correlación positivo entre estas variables ha ido decreciendo con el paso del tiempo por las virtudes del gas natural. Este último ayuda a generar energía más limpia y con menos generación de dióxido de carbono respecto al petróleo.
[2] Los datos se obtuvieron de US Energy Information Administration y Google Trends

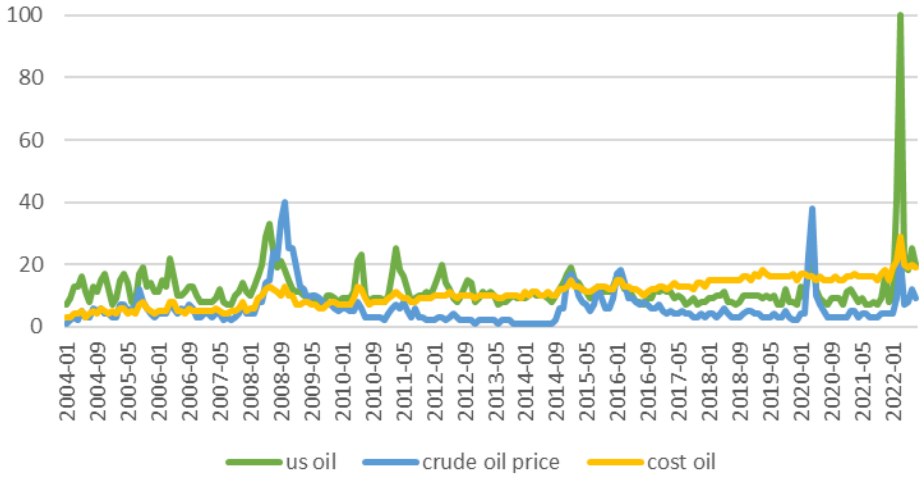
1.c Históricos precios WTI (imputación ourliers)



Fuente: Elaboración propia

Adicionalmente, se tuvo en cuenta datos de Google Trends[3] para el periodo de enero de 2004 a mayo de 2022 (generando 221 observaciones por variable de interés, ver Tabla 2 del Anexo) que representan las siguientes búsquedas en Estados Unidos relacionadas con petróleo: crude oil price, precio del petróleo, price of crude oil, oil price today, today crude oil price, oil consumption, world oil consumption, oil production, us oil, oil, cost oil, entre otras; evidenciando que los términos que lograron un indicador 100 en su búsqueda en algún mes del periodo en análisis, fueron crude oil price, us oil, oil, cost oil y biodiesel, reflejadas en la gráfica 2, variables explicativas de los modelos relacionados en el numeral III del presente documento.

Gráfica 2. Indicador Google Trends



Fuente: Elaboración propia

En las gráficas 1.b y 2, se observa que los outliers coinciden con el incremento en búsquedas en google trends de los términos relacionados con el petróleo, periodos que a su vez van en armonía con la crisis económica de 2008, la emergencia sanitaria mundial por la COVID-19 a inicios del año 2020 y con la Guerra entre Rusia y Ucránia en 2021. Lo anterior hace relevante la inclusión de estos variables a los modelos de predicción pues estas búsquedas pueden estar relacionadas con periodos en los cuales se presentaron impactos en el precio del petróleo.

[3] Herramienta de Google Labs que muestra los términos de búsqueda más populares, en donde los términos se indexan en función del valor de bus búsquedas, siendo **100** el valor máximo de interés de búsqueda, 50 el valor medio de interés de búsqueda y 0 el valor nulo de búsqueda del término.

III. MODELO

Considerando que los datos corresponden a series de tiempo, se realizaron particiones a la base de datos correspondiente al periodo de entrenamiento y de prueba, considerando un periodo de entrenamiento de agosto 1989 a diciembre 2020 para los modelos que no incluyen variables de google trends y julio 2004 a diciembre 2020 para los modelos que incluyen variables de google trends.

Por otro lado, para realizar validación cruzada con datos de series de tiempo se considera un proceso de rolling window para pronosticar un paso adelante la serie de los retornos del precio del petróleo, lo que quiere decir que se realizarán los pronósticos un paso adelante incluyendo el dato adicional para realizar el siguiente paso, por lo tanto, se tiene en cuenta una ventana móvil inicial de agosto 1989 a octubre 2014 para modelos sin google trends y julio 2004 a agosto 2017 para modelos con google trends, incrementando un mes adicional la ventana hasta llegar a diciembre 2020, de esta manera se obtienen pronósticos un mes adelante para los periodos noviembre 2014 a diciembre 2020 y septiembre 2017 a diciembre 2020, calculando con esta información los siguientes estándares relativos, MSE, RMSE, MAE y U-Theil, para seleccionar el mejor modelo.

Teniendo en cuenta la serie de los retornos logarítmicos del precio del petróleo ajustada por outliers se consideraron dos tipos de modelos, inicialmente se tienen modelos univariados en los cuales se modela la media y la varianza de la serie como ARIMA, GARCH y T-GARCH y por otro lado, se tienen modelos multivariados que consideran regularización con Elastic Net, destilación por componentes principales y NNET autorregresivo. Para los modelos multivariados se estiman dos modelos por cada metodología, un modelo 1 que no incluye las variables de google trends y un modelo 2 que incluye las variables de google trends, adicionalmente, para estos modelos se incluyen los 6 primeros rezagos de cada variable, esto considerando importante incluir periodos anteriores de todas las variables ya que estas series de tiempo presentan autocorrelación. La Tabla 3 presenta el accuracy de los modelos estimados

Tabla 3. Precisión (Accuracy) de los modelos estimados

Model	ME	RMSE	MAE	Theil's U
ARIMA Model	-0,0108	0,0863	0,0666	1,8499
GARCH Model	0,0066	0,0704	0,0560	1,0533
TGARCH Model	0,0066	0,0707	0,0561	1,0604
Elastic Net Model 1	0,0033	0,0686	0,0553	0,9013
Elastic Net Model 2 Google trends	0,0082	**0,0541**	0,0428	0,9414
Principal Component Model 1	0,0025	0,0720	0,0577	0,9945
Principal Component Model 2 Google trends	0,0049	**0,0596**	0,0468	1,0080
NNETAR Model	-0,0039	0,0739	0,0579	1,0451
NNETAR Model 1	0,0325	0,0783	0,0669	1,1787
NNETAR Model 2 Google trends	0,0458	0,0761	0,0644	1,0602

Fuente: Elaboración propia

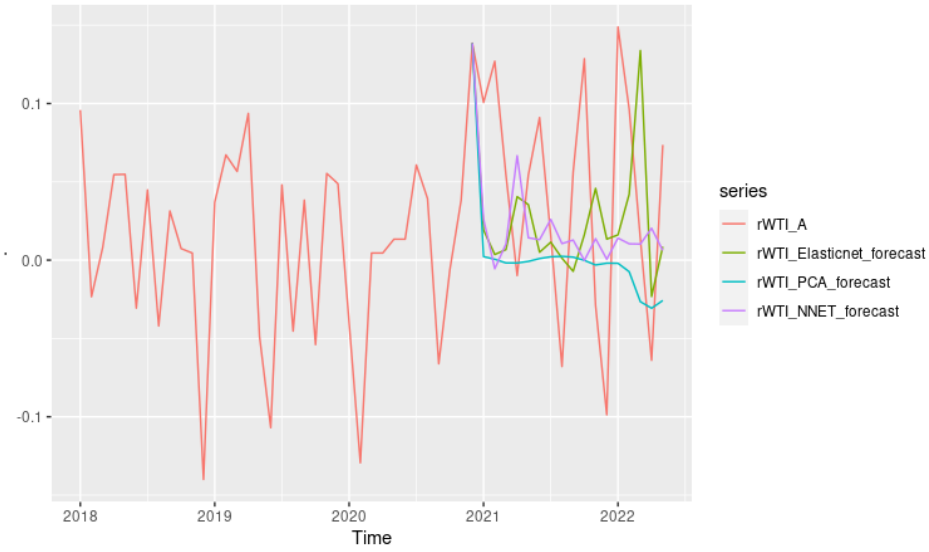
Esta tabla presenta los estandares relativos de los modelos entrenados, se puede evidenciar que el modelo estimado por Elastic Net que incluye los predictores de Google trend presenta el menor RMSE, por lo tanto, teniendo en cuenta este criterio, este es el modelo seleccionado, adicionalmente este modelo presenta un estandar relativo de U-Theil el cual muestra que el modelo explica un 94% adicional comparado con un modelo de caminata aleatoria. Las variables que se utilizaron para entrenar este modelo fueron principalmente la Producción de Petróleo, el precio del gas natural, la producción de gas natural, las variables de Google trend y todos los rezagos hasta 6 meses de estas variables. Este modelo de Elastic Net reduce los coeficientes de las variables considerando la importancia de cada variable para predecir los retornos del

precio del petróleo.

IV. RESULTADOS

Considerando los modelos estimados, se realiza la predicción de los retornos del precio del petróleo para el periodo enero 2021 a mayo 2022

Gráfica 3. Comparación de predicciones de los modelos seleccionados



Fuente: Elaboración propia

La gráfica 3 muestra una comparación del modelo seleccionado con la serie original y con otros modelos que también presentaron menores RMSE, se evidencia que la predicción del modelo de reducción Elastic Net se acerca más a la serie original de los retornos, por lo cual se considera el mejor modelo para predecir el precio del petróleo.

Por otro lado, la tabla anterior muestra que los modelos multivariados son mejores que los modelos univariados de media y varianza, y también presenta que incluir como predictores las variables obtenidas de google trends mejora la predicción de los modelos pues para los modelos de Elastic Net, Componentes Principales y NNET presentan menores RMSE comparados con los modelos que no incluyen estas variables para predecir el precio del petróleo.

V. CONCLUSIONES Y RECOMENDACIONES

Para la calibración del modelo se requiere que las series de tiempo del precio del petróleo WTI sean estacionarios, por lo tanto, se transformaron las series mediante la resta de los logaritmos de los datos de periodo consecutivos, hallando los crecimientos porcentuales de las series. Posteriormente de manera cuantitativa se ubicaron los outliers en las series estacionarias y se reemplazaron por la media del cuartil donde se encontraban.

Considerando el impacto no lineal que tienen diferentes factores sobre el precio del petróleo WTI, se decidió incluir en los modelos multivariados variables de google trends que permiten pronosticar incrementos y caídas fuertes del precio del petróleo correspondiente a eventos externos coyunturales y no estructurales que afectan el predio del petróleo.

Acorde con la comparación de RMSE entre los modelos, la evidencia empírica identifica al modelo de reducción Elastic Net como el modelo que más se acerca a la serie original de los retornos del WTI, siendo este el mejor modelo de predicción que además, incluye las variables obtenidas de Google trends, reflejando que el modelo multivariado genera mejores predicciones que los modelos univariados, sin embargo, dicho modelo en investigaciones futuras podría complementarse con variables macroeconómicas de Estados Unidos y de los principales países productores de petróleo con el fin de obtener modelos de predicción más precisos.

Finalmente, con las predicciones del modelo se pueden tomar las medidas respectivas para mitigar o potenciar los efectos de fuertes subidas o caídas del precio del petróleo de manera que las empresas puedan cubrirse a través de futuros en posiciones cortas o largas. Igualmente, en el camino de transición energética se puede sustituir el consumo de energía de fuentes fósiles en épocas de alza con energías renovables como energías eólicas o solares.

Referencias

Cen, Zhongpei & Wang, Jun. (2018). Crude oil price prediction model with long short term memory deep learning based on prior knowledge data transfer. Energy. 169. 10.1016/j.energy.2018.12.016.

Choi, J.-E., & Shin, D. W. (2022, January 31). How to improve oil consumption forecast using google trends from online big data?: the structured regularization methods for large vector autoregressive model. Communications for Statistical Applications and Methods. Communications for Statistical Applications and Methods. https://doi.org/10.29220/csam.2022.29.1.041

Lioudis, N (2022, 13 de junio). How Crude Oil Affects Natural Gas Prices. *Ivestopedia.* https://www.investopedia.com/articles/economics/08/crude-and-gas-prices.asp

Yu, L. (2019). Online big data-driven oil consumption forecasting with Google trends. International journal of forecasting, 35(1)

Villada, F., Arroyave, D., & Villada, M. (2014). Pronóstico del Precio del Petróleo mediante Redes Neuronales Artificiales. Revista Información Tecnológica, Vol. 25.

Anexo

Tabla 1. Estadísticas descriptivas variables series de tiempo

Variable	N	min	max	median	mean	std.dev
WTI (USD/Barril)	401	11,35	133,88	40,94	47,89	29,05
Producción de petróleo (miles de barriles)	401	119.208	400.219	198.691	219.519	63.178
Precio de importación del Gas Natural Gas (USD/millones de pies cúbicos)	401	1,34	11,99	2,80	3,61	2,12
Producción Gas Natural (millones de pies cúbicos)	401	1.649.263	3.679.110	2.056.810	2.292.110	504.807

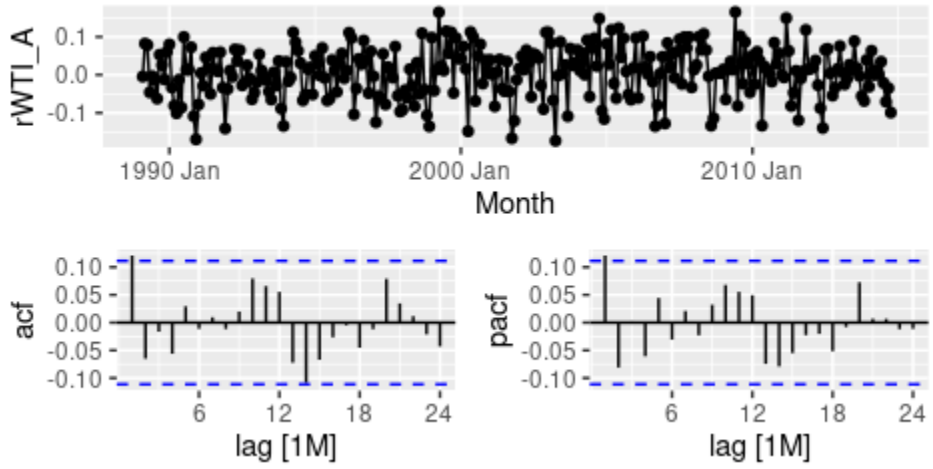
Fuente: Elaboración propia

Tabla 2. Estadísticas descriptivas variables google trends

Variable	N	min	max	median	mean	std.dev
crude oil price	221	3	100	10	15,04	14,61
precio del petroleo	221	0	5	1	1,13	0,72
price of crude oil	221	0	17	2	3,21	2,61
oil price today	221	0	58	4	6,79	8,35
today crude oil price	221	0	23	1	1,99	2,73
oil consumption	221	1	10	3	3,17	1,34
world oil consumption	221	0	2	1	1	0,10
oil production	221	3	26	5	5,93	2,68
us oil	221	7	100	10	11,88	7,56
oil	221	30	100	52	57,13	17,67
energy consumption	221	2	6	3	3,53	1,12
cost oil	221	9	100	35	36,22	14,85
cost of oil	221	4	36	13	12,50	3,97
fossil fuels	221	2	37	13	12,70	6,08
biodiesel	221	3	100	9	20,76	22,56
oil shock	221	1	2	1	1,11	0,32
OPEC	221	2	51	10	12,28	8,09
production gas	221	2	14	4	4,18	1,32
gulf mexico	221	1	9	1	1,08	0,65
Middle east	221	1	6	2	2,11	0,77
gas price	221	1	10	1	1,31	0,89

Fuente: Elaboración propia

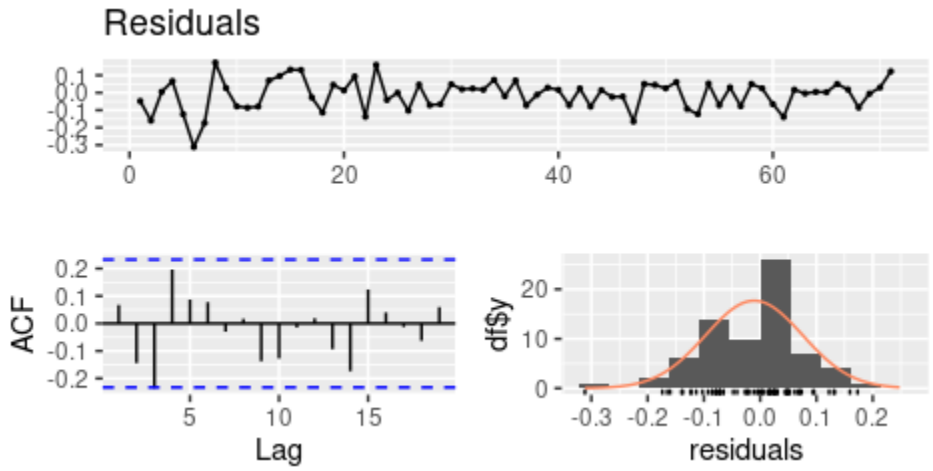
Gráfica 4. Autocorrelación simple y parcial de la serie de retornos logarítmicos ajustada por outliers



Fuente: Elaboración propia

Esta gráfica muestra el comportamiento de la autocorrelación simple y parcial de la serie de los retornos logarítmicos ajustada por outliers, se evidencia que despues de ajustar la serie persisten algunos rezagos con autocorrelación, por lo tanto se modela la media de la serie con un modelo ARIMA.

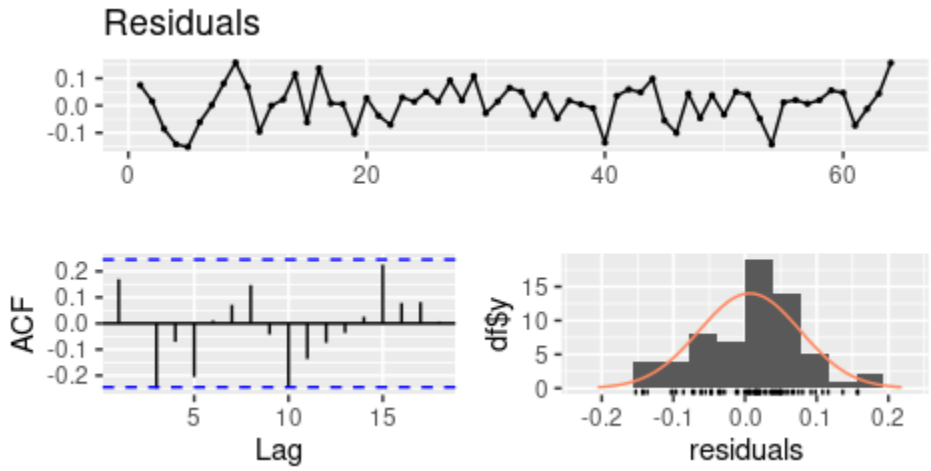
Gráfica 5. Errores modelo ARIMA



Fuente: Elaboración propia

Esta gráfica presenta el comportamiento de los errores de pronóstico del modelo ARIMA, encontrando que estos se comportan como un proceso ruido blanco, pues no presentan autocorrelación.

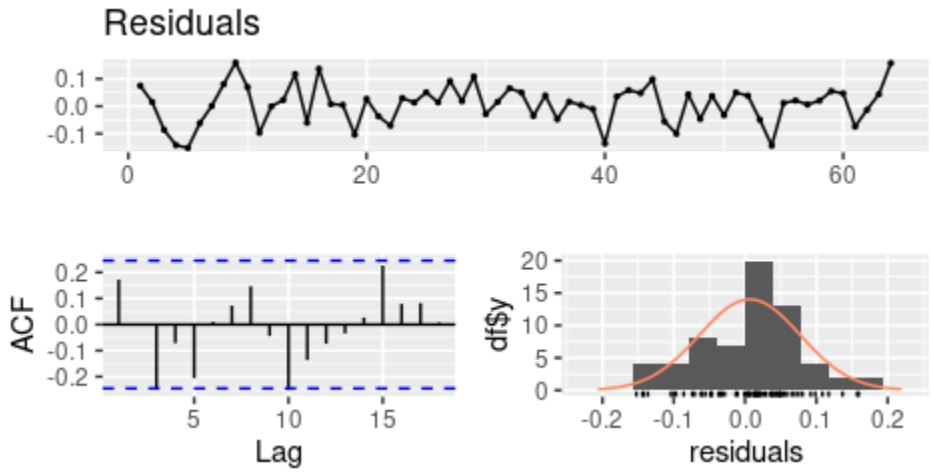
Gráfica 6. Errores de pronóstico modelo GARCH



Fuente: Elaboración propia

Se evidencia el comportamiento de los errores de pronóstico del modelo GARCH, encontrando que estos no presentan autocorrelación.

Gráfica 7. Errores de pronóstico modelo T-GARCH



Fuente: Elaboración propia

La gráfica muestra el comportamiento de los errores de pronóstico del modelo T-GARCH, encontrando que estos no presentan autocorrelación.